# Viruses of the Microbiome Outperform Bacteria in Colorectal Cancer Classification

Geoffrey D Hannigan Melissa B Duhaime Mack T Ruffin IV
Charlie C Koumpouras Patrick D Schloss

**Abstract**

Colorectal cancer is the second leading cause of cancer-related deaths 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 the context of colorectal cancer, despite their oncogenic potential. We evaluated the gut virome role in colorectal cancer using a cohort of 90 human subjects with either healthy, pre-cancerous, or cancerous large intestines. We utilized 16S rRNA gene, whole shotgun metagenomic, and purified virus metagenomic methods to compare the role of the virome to that of the bacterial community. We found that alpha and beta diversity metrics are insufficient for detecting an association between the virome and colorectal cancer, but virome-based classification models are both highly associated with colorectal cancer and outperform those based on the bacterial community signature. The association between bacterial communities and colorectal cancer is driven by a small subset of taxa, with the most significant player belonging to the genus *Fusobacterium*. The virome link to colorectal cancer is driven more equally by all members of the community. Bacteriophages, not eukaryotic viruses, make up the majority of the CRC-associated virome, suggesting the community is indirectly linked to colorectal cancer, modulating bacterial community structure and functionality. Although the community as a whole is associated with colorectal cancer, virome members with broader host ranges have a more important role in the cancer state. These results suggest that phage communities are strongly associated with colorectal cancer, and that broadly tropic phages play a more significant role. Because of its importance, the virome should be considered in future colorectal, and other cancer studies.

## Contents

# 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 the most impactful advances we have made has not been in cancer treatment, but rather in detecting cancer at early stages so as to improve treatment efficacy. This has been evident in a variety prominent cancers, 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 will be 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[2,3].

Development of colorectal cancer is a stepwise process that begins when healthy tissue develops into a pre-cancerous polyp (i.e. adenoma) in the large intestine[4]. 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 adenomas are detected and removed during routine screening[2,5]. Survival for colorectal cancer patients may exceed 90% when the lesions are detected early and removed. 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[6–9]. This association has allowed bacterial community signatures to been leveraged as biomarkers to greatly improve colorectal cancer detection[8,10]. While an understanding of colorectal cancer bacterial communities has proven fruitful both for disease prediction 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, viruses are strongly associated with, and in many cases cause, cancer[11–14]. Additionally, because bacteriophages are crucial for bacterial community stability and composition[15–17], and because bacteria have been implicated as oncogenic agents[8,9,18], bacteriophages are potentially indirectly linked to cancer. The gut virome (the virus community of the gut) has the potential to impact disease, and has been associated with a diseases including periodontal disease[19], HIV[20], antibiotic exposure[21,22], urinary tract infections[23], and irritable bowel disease[24]. We aim to take this line of research further by beginning to assess the role of the virome in cancer.

Here we present a study of the colorectal cancer virome and its utility for prognosis and diagnosis. By creating effective classification models using virus community signatures, we are able to accurately classify stool samples as cancerous, pre-cancerous, or healthy, while outperforming bacterial models. We find that unlike the bacterial community association with colorectal cancer, which is strongly linked to *Fusobacterium*, the virus community as a whole is relatively equally important. We also report that the impact of bacteriophages on colorectal cancer is linked to phage host range. The implications of these findings are threefold. **First**, this suggests a biological role for the virome in colorectal cancer development and that more than bacteria are involved in the process. **Second**, we present an avenue for higher performance classification modeling of colorectal cancer using stool samples. **Third**, this provides early evidence for the importance of studying the virome as a component of the microbiome, especially in cancer. We expect this study to open avenues for continued research into the role of the virome in human cancer development.

# Results

## The Colorectal Cancer Virome Cohort

Our cohort consisted of 90 human subjects, 30 of which served as healthy controls, 30 of which had adenoma lesions consistent with a pre-cancerous state, and 30 which had carcinoma lesions consistent with colorectal

cancer **(Figure 1)**. Half of the stool was aliquoted and used to sequence the bacterial communities using both 16S rRNA and shotgun sequencing techniques. The 16S rRNA sequences were reported in a previous publication[8]. 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 virus purification allowed us to observe the *active virome* because we only sequence those viruses that were encapsulated.

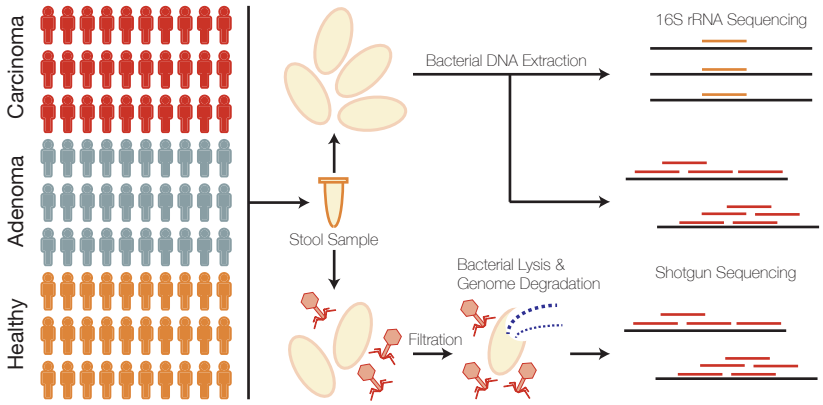


Figure 1: *Cohort and sample processing outline. Thirty subject stool samples were collected from healthy, adenoma (pre-cancer), and carcinoma (cancer) patients. Stool samples were split into two aliquots, the first of which was used for bacterial sequencing and the second which was used for virus sequencing.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

Virus DNA was purified by re-suspending the stool in saline magnesium buffer and removing contaminating cells (e.g. human, bacteria, etc) by filtering through a 0.22µm filter, followed by cell lysis with chloroform and degradation of the released genomic DNA with DNase **(Figure 1)**. The resulting genomic DNA was used to prepare a shotgun metagenomic sequencing library which sequenced on the Illumina HiSeq4000 platform. Each run was performed with a blank control to detect any contaminants from reagents. Only one of the controls detected DNA, which was of a minimal concentration, providing initial evidence of successful sequencing of VLP genomic DNA over potential contaminants **(Figure 6)**.

## Diversity is Insufficient for Virome-Based Cancer Classification

Microbiome disease-associations are often described as being under an altered diversity (i.e. dysbiosis). We evaluated the diversity of the virome and its association with colorectal cancer. We used beta-diversity to evaluate the differences in the communities between disease states. We utilized the Bray-Curtis dissimilarity metric to evaluate the differences between disease states. There was no observable clustering using NMDS ordination **(Figure 2 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 2 B)**. There were no significant differences between the disease groups, although there was a strongly significant difference between the negative controls and the rest of the study groups, further supporting the quality of our sample set **(Figure 9)**.

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 2 C-D)**. Overall, standard diversity metrics were insufficient for capturing the differences in the microbial 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 algorithms.

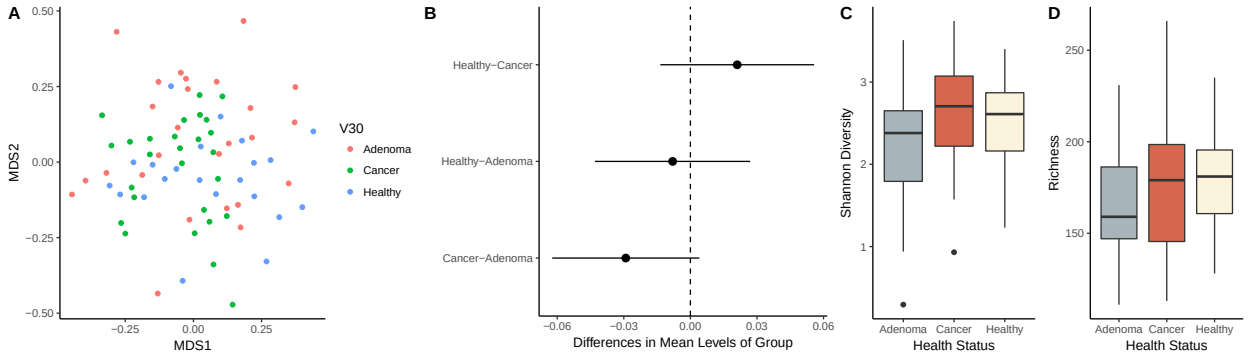


Figure 2: *Diversity calculations comparing cancer states of the colorectal virome, based on relative abundance of operational genomic units in each sample. A) NMDS ordination of community samples, colored for cancerous (green), pre-cancerous (red), and healthy (blue). B) Differences in means between 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.*

## Virome-Based Cancer Classification Outperforms Bacterial Models

Previous work has shown that 16S rRNA community signatures are effective for classifying stool samples as originating from healthy, pre-cancerous, or cancerous individuals[8,10]. This is valuable because it presents a potential alternative screening approach to an invasive colonoscopy. The exceptional performance of bacterial signatures in these predictive models also suggests 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 and 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. To improve performance and make the models computationally feasible, we only used OTUs and OGUs that were present in more than half of the samples. 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 findings from the original report which used logit models instead of random forest models **(Figure 3 A)**.

We compared the existing bacterial 16S rRNA gene model to a model built using the virome signatures. The viral model significantly outperformed the bacterial model by increasing the median AUC (area under the curve) from 0.80 to 0.86 **(Figure 3 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 median AUC of 0.56, lending support to our observation that high classification performance is a trait unique to the virome **(Figure 3 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 failed to improve performance beyond the model built using virome signatures alone (p-value $> 0.01$) **(Figure 3 A - B)**. Not only do bacterial community signatures fail to classify stool samples as well as the virome, but the bacteria do not have a synergistic impact on the virome classification model.

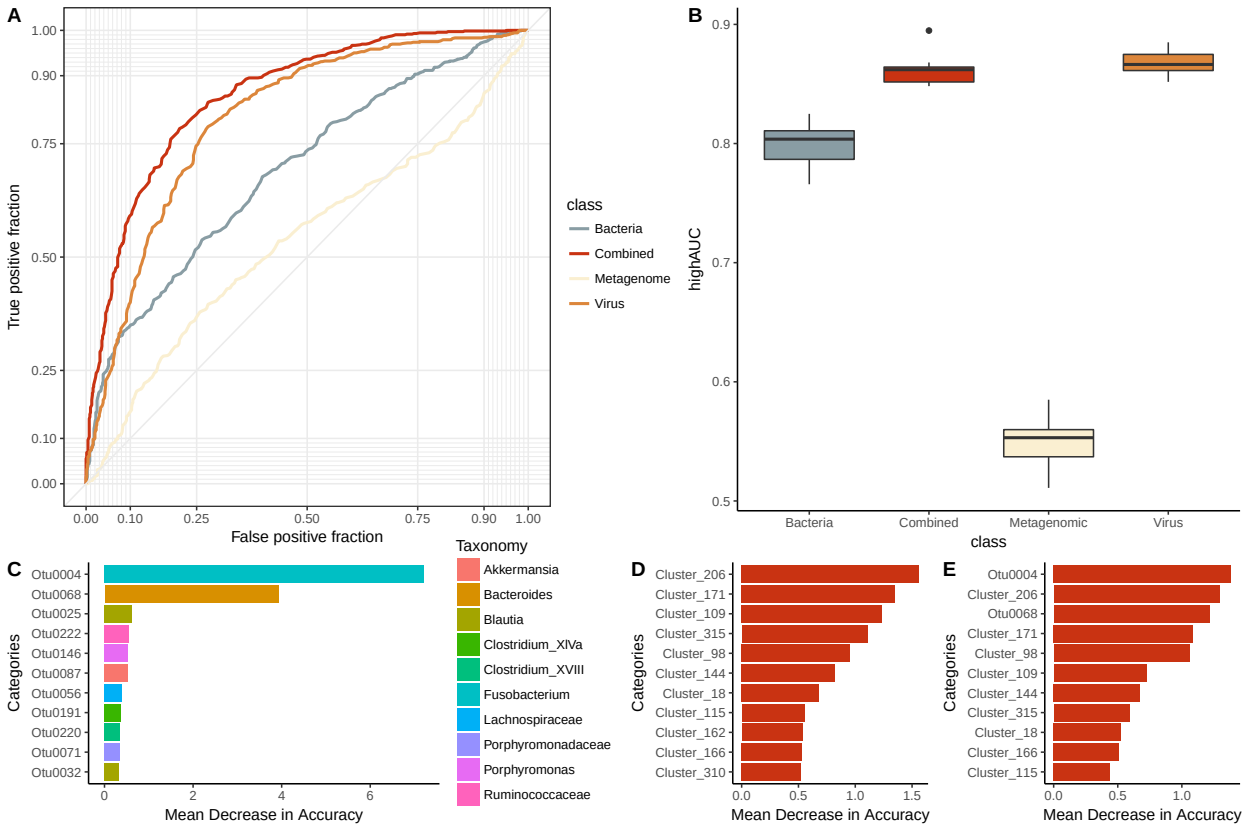


Figure 3: *Results from healthy vs cancer classification models built using virome signatures, bacterial 16S signatures, whole metagenomic signatures, and a combination of virome and 16S signatures. A) ROC curve for visualizing the performance of each of the models for classifying stool as coming from 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. A pairwise Wilcoxin test with a Bonferroni multiple hypothesis correction demonstrated that all models are significantly different except for the difference between the virome + 16S model and the virome alone (alpha = 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. 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.*

## Virome Performance is Consistent in Three-Class Classification Model

After evaluating our ability to classify samples as cancerous or healthy, we incorporated the pre-cancerous adenoma samples into the model and evaluated our ability to classify the groups out of the total dataset **(Figure 4)**. We used a set of three-class random forest models for the bacterial, whole metagenome, and viral sample sets. Again the virome significantly outperformed the bacterial community signatures **(Figure 4 A-C)**. The virus signature allowed for consistently high resolution for each of the disease classes, while the resolution varied with the bacterial signatures. The bacterial community allowed for high accuracy in classifying cancerous stool from pre-cancerous or healthy, but struggled to classify adenomas or healthy samples from the remainder of the dataset. On average the bacterial model performed significantly poorer than the virome model **(Figure 4 A-C)**.

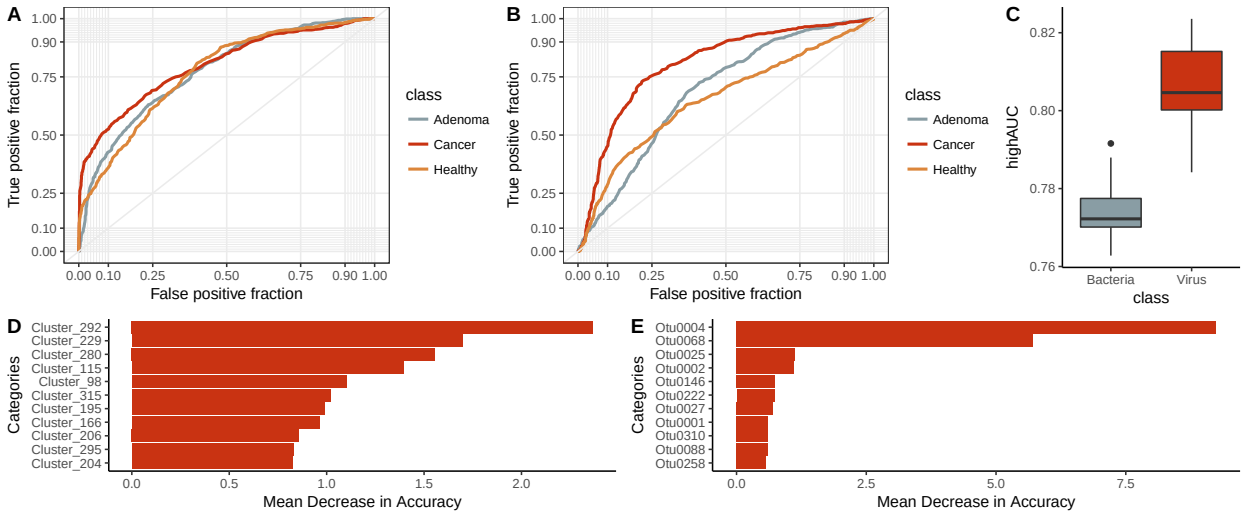


Figure 4: *ROC curves from A) virome and B) bacterial 16S three-class random forest models tuned on mean AUC. Each curve represents the ability of the specified class to be classified against the other two classes. C) Quantification of the mean AUC variation for each model. A pairwise Wilcoxin test with 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.*

## Bacteriophages Drive Link Between Virome and Colorectal Cancer

Our model for classifying stool samples as cancerous, pre-cancerous, or healthy performed significantly better when using virome signatures compared to bacterial 16S. Not only is this important for establishing an improved diagnostic model, but it also suggests an underlying biological importance for viruses in colorectal cancer. We used our predictive models to evaluate which virus OGUs were most highly linked to colorectal cancer.

We calculated the importance of each operational unit in each model by iteratively re-building the model without each unit and quantifying the resulting loss of accuracy. We found that, unlike bacterial community signatures which are driven by one or two primary taxa, the virome is driven more equally by all members of its community. The bacterial community cancer signature was driven by *Fusobacterium*, which was responsible for 6.5% of the model accuracy **(Figure 3 C, Figure 4 D)**. The top viral feature was only responsible for 1.5% **(Figure 3 D, Figure 4 E)**. There were no standout important viruses that allowed for cancer prediction, suggesting it is the community as a whole, and not a small subset of taxa, that is important for colorectal cancer progression.

The role of the virome in colorectal cancer could be driven directly by eukaryotic viruses or indirectly by bacteriophages which act through their bacterial hosts. To better understand the types of viruses that are important for colorectal cancer, we identified the longest sequences from the OGUs as the representative sequences to be used for taxonomic classification. These sequences were aligned (tblastx algorithm) to a set of all reference virus genomes, including bacteriophages and eukaryotic viruses. The tblastx algorithm is often used for virus/phage identification because the matches are based on amino acid similarity. A strict e-value threshold of 1e-25 was used to improve our confidence in the matches between the genome sequences. We were able to identify 79% (257 / 326) of the OGUs as viruses, and 95% of the viral OGUs aligned to bacteriophage reference genomes. Thus the majority of the OGUs are bacteriophages and not eukaryotic viruses, meaning the association between the virome and colorectal cancer is reliant almost entirely on bacteriophage communities.

## Broadly Infectious Phages Play Greater Role in Colorectal Cancer

Reference-based analyses of the virome perform poorly due to small reference databases and genomic modularity. Instead of relying on genome alignments to infer the hosts 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 5 A)**. We calculated the alpha centrality 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 = 1.5e-06, r = 0.28), indicating that phages with a broader host range play more important roles in colorectal cancer **(Figure 5 B)**.

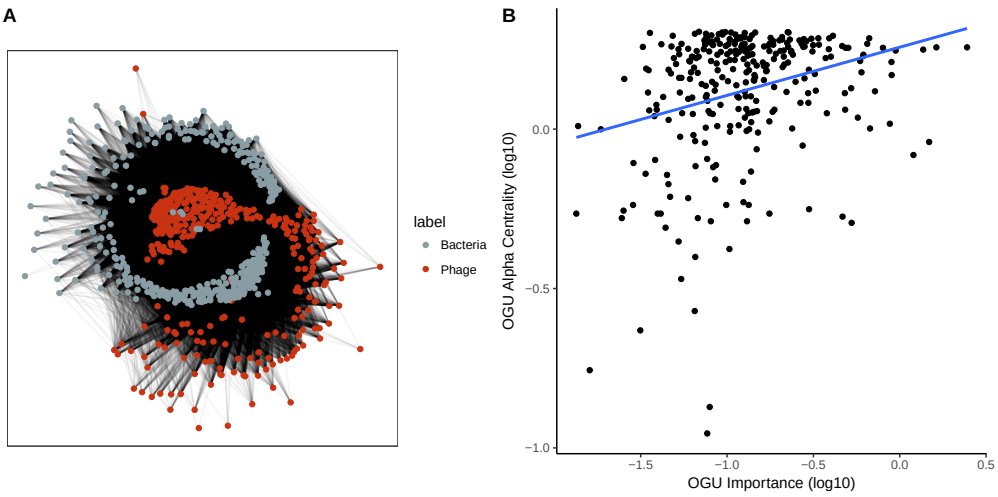


Figure 5: *Community network analysis utilizing predicted interactions between bacteria and phage operational genomic units. A) Visualization of the community network for our colorectal cancer cohort. B) Scatter plot illustrating the correlation between importance (mean decrease in accuracy) and the degree of centrality for each OGU. A linear regression line was fit to illustrate the correlation (blue) which was found to be statistically significantly weakly correlated (p-value = 1.5e-06, r = 0.28).*

# Discussion

Here we show that, like gut bacterial communities, the gut virome is associated with colorectal cancer. Alpha and beta diversity metrics are insufficient for detecting an association between the virome and colorectal cancer,

but virome-based classification models are both highly associated with colorectal cancer and outperform those based on the bacterial community signature. The association between bacterial communities and colorectal cancer is driven by a small subset of taxa, with the most significant player belonging to the genus *Fusobacterium*. The virome link to colorectal cancer is driven more equally by all members of the community. Bacteriophages, not eukaryotic viruses, make up the majority of the CRC-associated virome, suggesting the community is indirectly linked to colorectal cancer, modulating bacterial community structure and functionality. Although the community as a whole is associated with colorectal cancer, virome members with broader host ranges have a more important role in the cancer state.

We found that models based on whole metagenomic shotgun sequencing (primarily consisting of bacterial genomic DNA) performed very poorly compared to virome or 16S rRNA bacterial models. This was unexpected for us and could be the result of multiple factors. First, this could be the results of greater sharing of genomic signatures between bacterial taxa. Virus genomes are highly diverse, which is problematic for taxonomically identifying bacteriophages, but could be important when discriminating communities. Bacteria housekeeping genes and similar genes could be causing a loss of discriminatory information. Second, bacterial genomes are orders of magnitude larger than viral genomes, which would suggest that, given the same sequencing depth, we were better able to cover viral genomes than bacterial genomes. The whole metagenome model may perform better when a greater sequencing depth is achieved. These points warrant further investigation in future work, both in colorectal cancer and other cancers studied.

The finding that *Fusobacterium* is most highly associated with colorectal cancer confirms previous reports that support a similar role for the bacterium. Bacteria within the *Fusobacterium* genus are known to be oncogenic and have been suggested as playing a causative role in colorectal cancer. Although one bacterial taxa stood out as being highly associated with colorectal cancer, there were no such standouts from the virus communities. The entire community, which is predominantly bacteriophages, is important in colorectal cancer. That association is highly structured and dependent on other taxa within the community, as evidenced by the discrepancy between information gained by simple diversity techniques and machine learning modeling.

Although there was no standout phage taxa associated with colorectal cancer, our network analysis did reveal a structure to the phage OGUs that are playing the biggest role in the situation. Our community network analysis suggested that the importance of a phage to colorectal cancer is associated with its host specificity. Phages with broader host ranges are more highly associated with colorectal cancer. Together this suggests that not only is the virus community as a whole the important in colorectal cancer, but the more broadly infecting phages are linked to the cancer state.

When interpreting these results, it is worth noting that we are studying these communities as taxonomic units and with the methods presented, we are unable to make conclusions with a high taxonomic resolution. Bacterial operational taxonomic units can represent relatively large categories of bacterial taxa such as genera that 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 phage or bacterial strains or species.

In addition to the clear ramifications for understanding colorectal cancer, 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 more information than the bacterial communities. We firmly believe that the virome is a crucial component to the microbiome and that bacteriophages are important players. A bacteriophage and bacterial community cannot thrive without the other. 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 work flow 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[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, have not had surgery, have not had chemotherapy or radiation, and were free of known comorbidities including HIV, chronic viral hepatitis, HNPCC, FAP, and inflammatory bowel disease. Sample 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[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 analytical toolkit[25,26]. Briefly, the sequences were de-replicated, screened for chimeras using UCHIME[27] and the SILVA database[28], and binned into operational taxonomic units (OTUS) using a 97% similarity threshold.

## 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 125bp paired end reads.

## Virus Metagenomic Library Preparation & Sequencing

Genomic DNA was extracted from purified virus-like particles (VLPs) from stool samples, using a modified version of a previously published protocol[29]. 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) as previously described. 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[29]. The library was sequenced using one lane of the Illumina HiSeq4000 platform and yielded 125bp 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 CutAdapt with an error rate of 0.1 and an overlap of 10[30]. The FastX toolkit was used to quality trim the reads to a minimum length of 75bp and a minimum quality score of 30[31]. Reads mapping to the human genome were removed using the DeconSeq algorithm and default parameters[32].

## 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 was used to assemble contigs for each sample using a minimum contig length of 1000bp and iterating assemblies from 21-mers to 101-mers by 20[33]. 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, with a 25bp seed length and an allowance of one mismatch[34]. Abundance was corrected for contig reference length.

## 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 which bins related contigs by similar tetra-mer and co-abundance profiles within samples using a variational Bayesian approach[35]. CONCOCT was used with a length threshold of 1000bp for virus contigs and 2000bp 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 50,000 reads for beta diversity and 100,000 for alpha diversity. 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, 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.

## Classification Modeling

Classification modeling was performed in R using the Caret package. 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. Significance was evaluated under alpha = 0.01.

### 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 and a strict similarity threshold (e-value < 1e-25). 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. The degree of phage centrality was quantified using the alpha centrality metric in the igraph CRAN package. A Pearson correlation was performed between model importance and phage centrality scores. Because scores were not normally distributed, the values were log transformed.

## Conflicts of Interest

The authors declare no conflicts of interest.

## Supplemental Figures

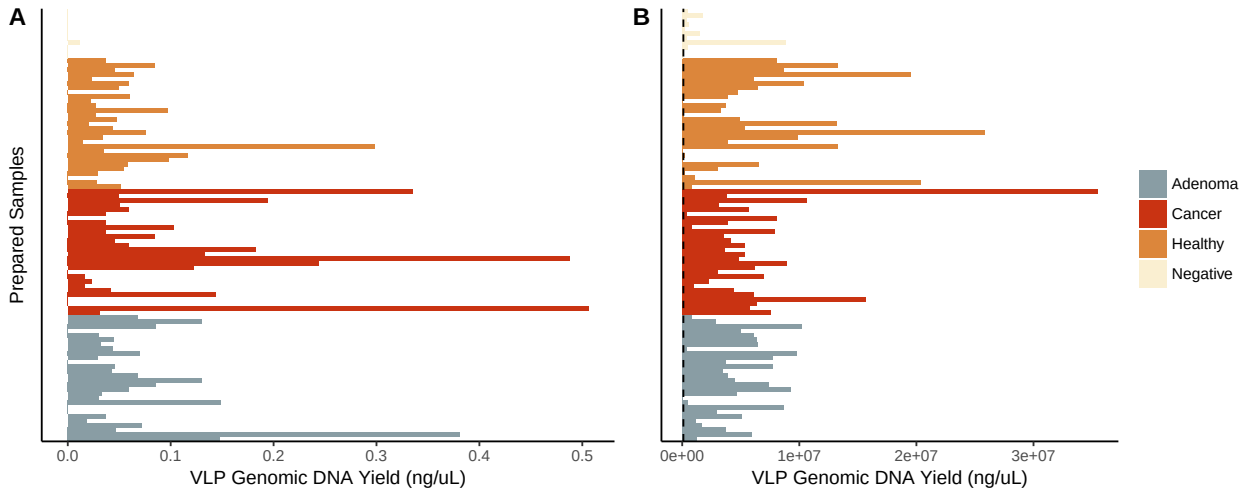


Figure 6: *Basic Quality Control Metrics. A) VLP genomic DNA yield from all sequenced samples. Each bar represents a sample which is grouped and colored by its associated disease group. B) Sequence yield following quality control including quality score filtering and human decontamination. Dashed line represents the sub-sampling depth used in the study.*

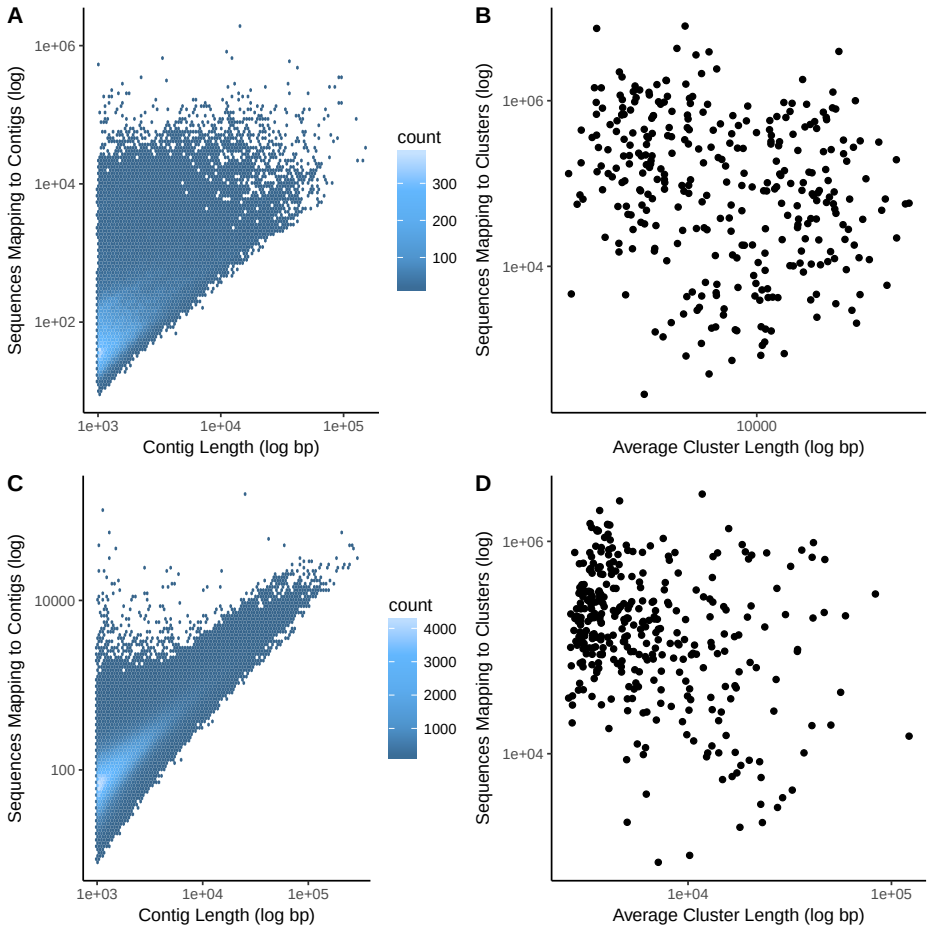


Figure 7: *Length and coverage statistics. A) Heated scatter plot demonstrating the distribution of contig coverage (number of sequences mapping to each contig) and contig length for the virus metagenomic sample set. B) Scatter plot illustrating the distribution of operational genomic unit (OGU) length 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.*

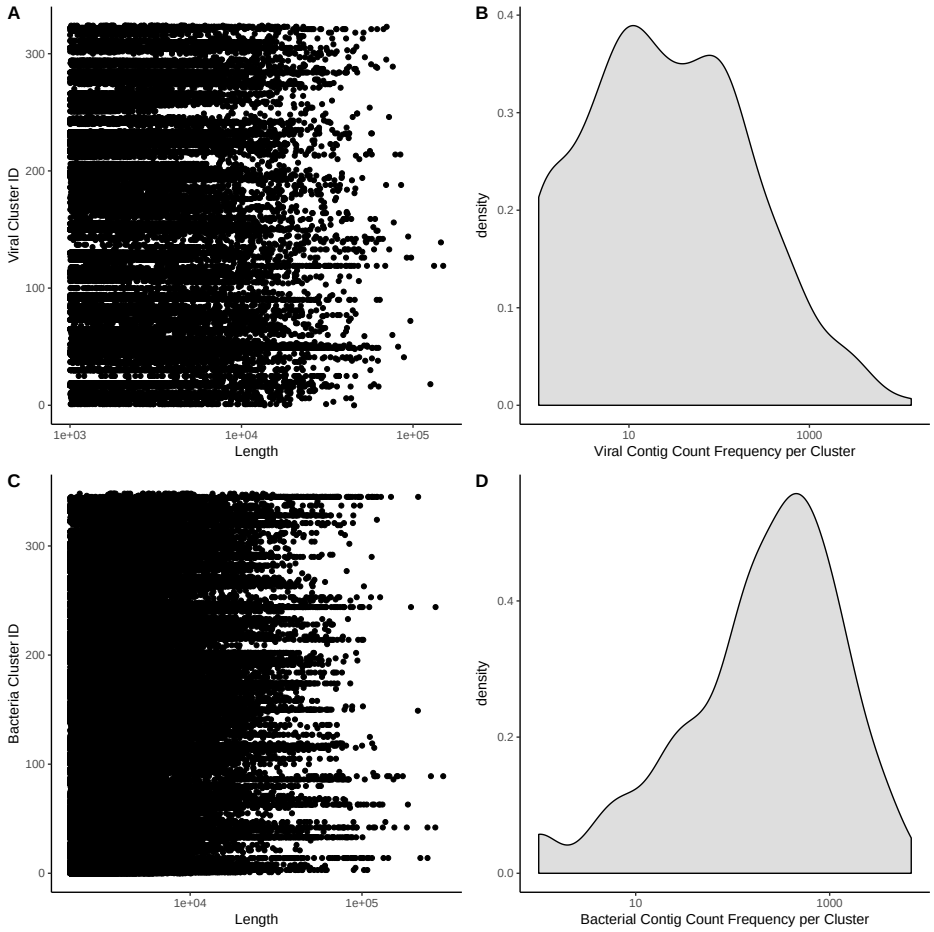


Figure 8: *Operational genomic unit composition stats. A) Strip chart demonstrating the length and frequency of contigs within each operational genomic unit of the virome sample set. The y-axis is the operational genomic unit identifier, and x-axis is the length of each contig, and each dot represents 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.*

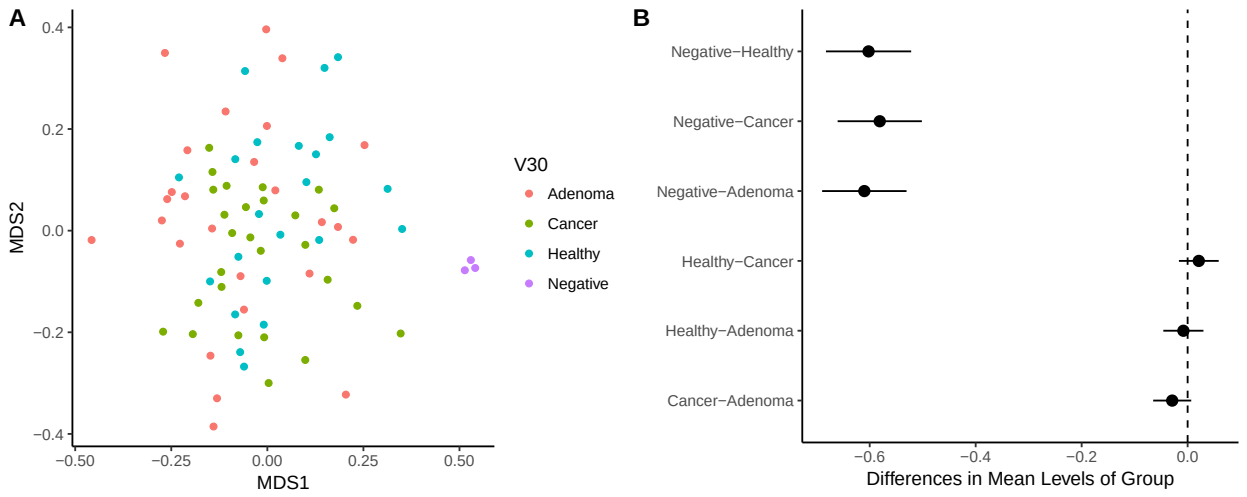


Figure 9: *Beta-diversity comparing disease states and the study negative controls. A) NMDS ordination of community samples, colored by disease state. B) Differences in means between disease group centroids with 95% confidence intervals based on an Anosim test with a post hoc multivariate Tukey test. Comparisons in which the intervals cross the zero mean difference line (dashed line) were not significantly different.*

# References


1. Howlader, N. *et al.* SEER Cancer Statistics Review, 1975-2013. *National Cancer Institute* (2016).

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

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

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

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

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

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

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

9. Baxter, N. T., Zackular, J. P., Chen, G. Y. & Schloss, P. D. Structure of the gut microbiome following colonization with human feces determines colonic tumor burden. *Microbiome* **2,** 20 (2014).

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

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

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

3623–3634 (2011).

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

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

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

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

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

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

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

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

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

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

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

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

25. Sze, M. A. & Schloss, P. D. Looking for a Signal in the Noise: Revisiting Obesity and the Microbiome. *mBio* **7,** e01018–16 (2016).

26. Schloss, P. D. *et al.* Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology* **75,** 7537–7541 (2009).

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

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

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

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

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

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

33. Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies

and community practices. *Methods* **102,** 3–11 (2016).

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

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