bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The Diagnostic Potential & Interactive Dynamics of the Colorectal Cancer Virome

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

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

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

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

[*]To whom correspondence may be addressed.

Corresponding Author Information

Patrick D Schloss, PhD

1150 W Medical Center Dr. 1526 MSRB I

Ann Arbor, Michigan 48109

Phone: (734) 647-5801

Email: pschloss@umich.edu

Article Class: Research Article

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

18 # Abstract

19 Human viruses (those that infect human cells) have been associated with many cancers, largely due to
20 their mutagenic and functionally manipulative abilities. Despite this, cancer microbiome studies have almost
21 exclusively focused on bacteria instead of viruses. We began evaluating the cancer virome by focusing
22 on colorectal cancer, a primary cause of morbidity and mortality throughout the world, and a cancer linked
23 to altered colonic bacterial community compositions but with an unknown association with the gut virome.
24 We used 16S rRNA gene, whole shotgun metagenomic, and purified virus metagenomic sequencing of
25 stool to evaluate the differences in human colorectal cancer virus and bacterial community composition.
26 Through random forest modeling we identified differences in the healthy and colorectal cancer virome. The
27 cancer-associated virome consisted primarily of temperate bacteriophages that were also predicted to be
28 bacteria-virus community network hubs. These results provide foundational evidence that bacteriophage
29 communities are associated with colorectal cancer and potentially impact cancer progression by altering the
30 bacterial host communities.

31 # Importance

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

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

40 # Introduction

41 The human gut virome is the community of all viruses found in the gut, including bacteriophages (viruses
42 that only infect bacteria), eukaryotic viruses (viruses that only infect eukaryotic cells), and human-specific
43 viruses (viruses that only infect human cells). Due to their mutagenic abilities and propensity for functional
44 manipulation, human viruses are strongly associated with, and in many cases cause, cancer (1–4). Because
45 bacteriophages are crucial for bacterial community stability and composition (5–7), and members of those
46 bacterial communities have been implicated as oncogenic agents (8–11), bacteriophages have the potential
47 to indirectly impact cancer as well. The gut virome therefore has a potential to be associated with, and
48 potentially impact, human cancer. Altered human virome composition and diversity have already been
49 identified in diseases including periodontal disease (12), HIV (13), cystic fibrosis (14), antibiotic exposure (15,
50 16), urinary tract infections (17), and inflammatory bowel disease (18). The strong association of bacterial
51 communities with colorectal cancer, the previous identification of human-specific viruses that cause cancer,
52 and the precedent for the virome to impact other human diseases suggest that colorectal cancer may be
53 associated with altered virus communities.

54 Colorectal cancer is the second leading cause of cancer-related deaths in the United States (19). The
55 US National Cancer Institute estimates over 1.5 million Americans were diagnosed with colorectal cancer
56 in 2016 and over 500,000 Americans died from the disease (19). Growing evidence suggests that an
57 important component of colorectal cancer etiology may be perturbations in the colonic bacterial community
58 (8, 10, 11, 20, 21). Work in this area has led to a proposed disease model in which bacteria colonize the
59 colon, develop biofilms, promote inflammation, and enter an oncogenic synergy with the cancerous human
60 cells (22). This association also has allowed researchers to leverage bacterial community signatures as
61 biomarkers to provide accurate, noninvasive colorectal cancer detection from stool (8, 23, 24). While an
62 understanding of colorectal cancer bacterial communities has proven fruitful both for disease classification
63 and for identifying the underlying disease etiology, bacteria are only a subset of the colon microbiome.
64 Viruses are another important component of the colon microbial community that have yet to be studied in
65 the context of colorectal cancer. We evaluated disruptions in virus and bacterial community composition

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

66 in a human cohort whose stool was sampled at the three relevant stages of cancer development: healthy,
67 adenomatous, and cancerous.

68 Colorectal cancer progresses in a stepwise process that begins when healthy tissue develops into a
69 precancerous polyp (i.e., adenoma) in the large intestine (25). If not removed, the adenoma may develop
70 into a cancerous lesion that can invade and metastasize, leading to severe illness and death. Progression to
71 cancer can be prevented when precancerous adenomas are detected and removed during routine screening
72 (26, 27). Survival for colorectal cancer patients may exceed 90% when the lesions are detected early and
73 removed (26). Thus, work that aims to facilitate early detection and prevention of progression beyond early
74 cancer stages has great potential to inform therapeutic development.

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

83 Results

84 Sample Collection and Processing

85 Our study cohort consisted of stool samples collected from 90 human subjects, 30 of whom had healthy
86 colons, 30 of whom had adenomas, and 30 of whom had carcinomas **(Figure 1)**. Half of each stool sample
87 was used to sequence the bacterial communities using both 16S rRNA gene and shotgun sequencing
88 techniques. The 16S rRNA gene sequencing was performed for a previous study, and the sequences were

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

89 re-analyzed using contemporary methods (8). The other half of each stool sample was purified for virus
90 like particles (VLPs) before genomic DNA extraction and shotgun metagenomic sequencing. In the VLP
91 purification, cells were disrupted and extracellular DNA degraded **(Figure 1)** to allow the exclusive analysis
92 of viral DNA within virus capsids. In this manner, the *extracellular virome* of encapsulated viruses was
93 targeted.

94 Each extraction was performed with a blank buffer control to detect contaminants from reagents or other
95 unintentional sources. Only one of the nine controls contained detectable DNA at a minimal concentration of
96 0.011 ng/μl, thus providing evidence of the enrichment and purification of VLP genomic DNA over potential
97 contaminants **(Figure S1 A)**. As expected, these controls yielded few sequences and were almost entirely
98 removed while rarefying the datasets to a common number of sequences **(Figure S1 B)**. The high quality
99 phage and bacterial sequences were assembled into highly covered contigs longer than 1 kb **(Figure S2)**.
100 Because contigs represent genome fragments, we further clustered related bacterial contigs into operational
101 genomic units (OGUs) and viral contigs into operational viral units (OVUs) **(Figure S2 - S3)** to approximate
102 organismal units.

103 Unaltered Diversity in Colorectal Cancer

104 Microbiome and disease associations are often described as being of an altered diversity (i.e. “dysbiotic”).
105 Therefore, we first evaluated the influence of colorectal cancer on virome OVU diversity. We evaluated
106 differences in communities between disease states using the Shannon diversity, richness, and Bray-Curtis
107 metrics. We observed no significant alterations in either Shannon diversity or richness in the diseased states
108 as compared to the healthy state **(Figure S4 C-D)**. There was no statistically significant clustering of the
109 disease groups (ANOSIM p-value = 0.6, **Figure S4**). Notably, there was a significant difference between
110 the few blank controls that remained after rarefying the data and the other study groups (ANOSIM p-value
111 < 0.001, **Figure S5)**, further supporting the quality of the sample set. In summary, standard alpha and beta
112 diversity metrics were insufficient for capturing virus community differences between disease states **(Figure**
113 **S4)**. This is consistent with what has been observed when the same metrics were applied to 16S rRNA gene

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

114 sequences and metagenomic samples (8, 23, 24) and points to the need for alternate approaches to detect
115 the impact of colorectal cancer disease state on these community structures.

116 Virome Composition in Colorectal Cancer

117 As opposed to the diversity metrics discussed above, OTU-based relative abundance profiles generated
118 from 16S rRNA gene sequences are effective for classifying stool samples as originating from individuals
119 with healthy, adenomatous, or cancerous colons (8, 23). By using classification models instead of attempting
120 to identify single differentially abundant OTUs, these and other studies have been successful in capturing
121 complex community relationships in which differences in taxonomic relative abundance are considered in
122 the context of other taxa. The exceptional performance of bacteria in these classification models supports a
123 role for bacterial functionality in colorectal cancer. We built off of these findings by evaluating the ability of
124 virus community signatures to classify stool samples and compared their performance to models built using
125 bacterial community signatures.

126 To identify the altered virus communities associated with colorectal cancer, we built and tested random forest
127 models for classifying stool samples as belonging to individuals with either cancerous or healthy colons. We
128 confirmed that our bacterial 16S rRNA gene model replicated the performance of the original report which
129 used logit models instead of random forest models **(Figure 2 A)** (8). We then compared the bacterial OTU
130 model to a model built using OVU relative abundances. The viral model performed as well as the bacterial
131 model (corrected p-value = 0.6), with the viral and bacterial models achieving mean area under the curve
132 (AUC) values of 0.768 and 0.775, respectively **(Figure 2 A - B)**. To evaluate the ability of both bacterial and
133 viral biomarkers to classify samples, we built a combined model that used both bacterial and viral community
134 data. The combined model did not yield a statistically significant performance improvement beyond the viral
135 (corrected p-value = 0.08) and bacterial (corrected p-value = 0.1) models, yielding an AUC of 0.807 **(Figure**
136 **2 A - B)**.

137 We compared viral metagenomic methods to bacterial metagenomic methods by building a viral model and
138 a model built using OGU relative abundance profiles from bacterial metagenomic shotgun sequencing data.

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

139 This bacterial model performed worse than the other models (mean AUC = 0.458) **(Figure 2 A - B)**. To
140 determine the cause of the discrepancy between the two bacterial sequencing methods, we attempted to
141 compare the approaches at a common sequencing depth. This revealed that the bacterial 16S rRNA gene
142 model was strongly driven by sparse and low abundance OTUs **(Figure S6)**. Removal of OTUs with a
143 median abundance of zero resulted in the removal of six OTUs, and a loss of model performance down
144 to what was observed in the metagenome-based model **(Figure S6 A)**. The majority of these OTUs had a
145 relative abundance lower than 1% across the samples **(Figure S6 B)**. Although the features in the viral model
146 also were of low abundance **(Figure S8 F)**, the coverage was sufficient for high model performance, likely
147 because viral genomes are orders of magnitude smaller than bacterial genomes.

148 The association between the bacterial and viral communities and colorectal cancer was driven by a few
149 important microbes. *Fusobacterium* was the primary driver of the bacterial association with colorectal cancer,
150 which is consistent with its previously described oncogenic potential **(Figure 2 C)**(22). The virome signature
151 also was driven by a few OVUs, suggesting a role for these viruses in tumorigenesis **(Figure 2 D)**. It is also
152 important to note that while these viruses were driving the signature, the magnitude of their importance and
153 the significance of those values was noticibly less than the bacterial 16S signature, suggesting that unlike
154 what is observed in the bacteria, there are many viruses that are associated with the cancerous state. The
155 identified viruses were bacteriophages, belonging to *Siphoviridae*, *Myoviridae*, and phage taxa that could
156 not be confidently identified beyond their broad phage identification (i.e. “unclassified”). Viruses, which
157 were confirmed to not have genomic similarity to known bacterial genomes, were unidentifiable (denoted
158 “unknown”). This is common in viromes across habitats; studies have reported as much as 95% of virus
159 sequences belonging to unknown genomic units (14, 28–30). When the bacterial and viral community
160 signatures were combined, both bacterial and viral organisms drove the community association with cancer
161 **(Figure 2 E)**.

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

162 ## Phage Influence Between CRC Stages

163 Because previous work has identified shifts in which bacteria were most important at different stages
164 of colorectal cancer (8, 20, 22), we explored whether shifts in the relative influence of phages could be
165 detected between healthy, adenomatous, and cancerous colons. We evaluated community shifts between
166 the disease stage transitions (healthy to adenomatous and adenomatous to cancerous) by building random
167 forest models to compare only the diagnosis groups around the transitions. While bacterial OTU models
168 performed equally well for all disease class comparisons, the virome model performances differed **(Figure**
169 **S7 A-B)**. Like bacteria **(Figure S7 F-H)**, different virome members were important between the healthy to
170 adenomatous and adenomatous to cancerous stages **(Figure S7 C-E)**.

171 After evaluating our ability to classify samples between two disease states, we performed a three-class
172 random forest model including all disease states. The 16S rRNA gene model yielded a mean AUC of 0.784
173 and outperformed the viral community model, which yielded a mean AUC of 0.654 (p-value < 0.001, **Figure**
174 **S8 A-C**). The microbes important for the healthy versus cancer and healthy versus adenoma models were
175 also important for the three-class model **(Figure S8 D-E)**. The most important bacterium in the two and three
176 class models was the same *Fusobacterium* (OTU 4) **(Figure 2 C, Figure S8 D)**. The viruses most important
177 to the three-class model were identified as bacteriophages **(Figure 2 D, Figure S8 E)**, but not all important
178 OVUs were of increased abundance in the diseased state **(Figure S8 F)**.

179 ## Phage Dominance in CRC Virome

180 Differences in the colorectal cancer virome could have been driven by eukaryotic (human) viruses or by
181 bacteriophages. To better understand the types of viruses that were important for colorectal cancer, we
182 identified the virome OVUs as being similar to either eukaryotic viruses or bacteriophages. The most
183 important viruses to the classification model were identified as bacteriophages (**Figure S8)**. Overall, we
184 were able to identify 78.8% of the OVUs as known viruses, and 93.8% of those viral OVUs aligned to
185 bacteriophage reference genomes. It is important to note that this could have been influenced by our

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

186 methodological biases against enveloped viruses (more common of eukaryotic viruses than bacteriophage),
187 due to chloroform and DNase treatment for purification.

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

197 Community Context of Influential Phages

198 Because the link between colorectal cancer and the virome was driven by bacteriophages (as opposed
199 to non-bacterial viruses), we tested a potential hypothesis that the virome signal was a mere reflection of
200 the bacterial signal, and thus highly correlated with the bacterial signal. If this hypothesis were true, we
201 would expect a correlation between the relative abundances of influential bacterial OTUs and virome OVUs.
202 Instead, we observed a strikingly low correlation between bacterial and viral relative abundances **(Figure 4**
203 **A,C)**. Overall, there was an absence of correlation between the most influential OVUs and bacterial OTUs
204 **(Figure 4 B)**. This evidence supported our null hypothesis that the influential viral OVUs were not primarily
205 reflections of influential bacteria.

206 Given these findings, we posited that the most influential phages were acting by infecting a wide range of
207 bacteria in the overall community, instead of just the influential bacteria. In other words, we hypothesized
208 that the influential bacteriophages were community hubs (i.e. central members) within the bacteria and
209 phage interactive network. We investigated the potential host ranges of all phage OVUs using a previously
210 developed random forest model that relies on sequence features to predict which phages infected which

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

211 bacteria in the community **(Figure 5 A)** (34). The predicted interactions were then used to identify phage
212 community hubs. We calculated the alpha centrality (i.e. measure of importance in the ecological network) of
213 each phage OVU's connection to the rest of the network. The phages with high centrality values were defined
214 as community hubs. Next, the centrality of each OVU was compared to its importance in the colorectal cancer
215 classification model. Phage OVU centrality was significantly and positively correlated with importance to the
216 disease model (p-value = 0.004, R = 0.176), suggesting that phages that were important in driving colorectal
217 cancer also were more likely to be community hubs **(Figure 5 B)**. Together these findings supported our
218 hypothesis that influential phages were hubs within their microbial communities and had broad host ranges.

219 Discussion

220 Because of their propensity for mutagenesis and capacity for modulating their host functionality, many
221 human viruses are oncogenic (1–4). Some bacteria also have oncogenic properties, suggesting that
222 bacteriophages, a component of the human virome in addition to human-specific viruses, may play an
223 indirect role in promoting carcinogenesis by influencing bacterial community composition and dynamics
224 (8–10). Despite their carcinogenic potential and the strong association between bacteria and colorectal
225 cancer, a link between virus colorectal communities and colorectal cancer has yet to be evaluated. Here
226 we show that, like colonic bacterial communities, the colon virome was altered in patients with colorectal
227 cancer relative to those with healthy colons. Our findings support a working hypothesis for oncogenesis by
228 phage-modulated bacterial community composition.

229 Based on our findings, we have developed a conceptual model to be tested in our future studies aimed at
230 elucidating the role the colonic virome plays in colorectal cancer **(Figure 6 A)**. We found that basic diversity
231 metrics of alpha diversity (richness and Shannon diversity) and beta diversity (Bray-Curtis dissimilarity)
232 were insufficient for identifying virome community differences between healthy and cancerous states. By
233 implementing a machine learning approach (random forest classification) to leverage inherent, complex
234 patterns not detected by diversity measures, we were able to detect strong associations between the colon

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

235 virus community composition and colorectal cancer. The dsDNA virome of colorectal cancer was composed
236 primarily of bacteriophages. These phage communities were not exclusively predators of the most influential
237 bacteria, as demonstrated by the lack of correlation between the abundances of the bacterial and phage
238 populations. Instead, we identified influential phages as being community hubs, suggesting phages influence
239 cancer by altering the greater bacterial community instead of directly modulating the influential bacteria. Our
240 previous work has shown that modifying colon bacterial communities alters colorectal cancer progression and
241 tumor burden in mice (10, 20). This provides a precedent for phages indirectly influencing colorectal cancer
242 progression by altering the bacterial community composition. Overall, our data support a model in which
243 the bacteriophage community modulates the bacterial community, and through those interactions indirectly
244 influences the bacteria driving colorectal cancer progression **(Figure 6 A)**. Although our evidence suggested
245 phages indirectly influenced colorectal cancer development, we were not able to rule out the role of phages
246 directly interacting with the human host (35, 36).

247 In addition to modeling the potential connections between virus communities, bacterial communities, and
248 colorectal cancer, we also used our data and existing knowledge of phage biology to develop a working
249 hypothesis for the mechanisms by which this may occur. This was done by incorporating our findings into a
250 current model for colorectal cancer development **(Figure 6 B)** (22), although it is important to note that there
251 are also many other alternative hypotheses by which the system could be operating. We hypothesize that
252 the process begins with broadly infectious phages in the colon lysing, and thereby disrupting, the existing
253 bacterial communities. This shift opens novel niche space that enabled opportunistic bacteria (such as
254 *Fusobacterium nucleatum*) to colonize. Once the initial influential founder bacteria establish themselves in
255 the epithelium, secondary opportunistic bacteria are able to adhere to the founders, colonize, and establish
256 a biofilm. Phages may play a role in biofilm dispersal and growth by lysing bacteria within the biofilm, a
257 process important for effective biofilm growth (37). The oncogenic bacteria may then be able to transform the
258 epithelial cells and disrupt tight junctions to infiltrate the epithelium, thereby initiating an inflammatory immune
259 response. As the adenomatous polyps developed and progressed towards carcinogenesis, we observed a
260 shift in the phages and bacteria whose relative abundances were most influential. As the bacteria enter their
261 oncogenic synergy with the epithelium, we conjecture that the phages continue mediating biofilm dispersal.

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

262 This process would thereby support the colonized oncogenic bacteria by lysing competing cells and releasing
263 nutrients to other bacteria in the form of cellular lysates. In addition to highlighting the likely mechanisms by
264 which the colorectal cancer virome is interacting with the bacterial communities this model will guide future
265 research investigations of the role the virome plays colorectal cancer.

266 Our working hypothesis represents a conceptualization of areas for future work, which will be required
267 to characterize the colorectal cancer microbiome at the functional, mechanistic level. There are many
268 different ways in which this system may operating, and our working hypothesis is one. For example, it is
269 possible that the bacterial communities cause a change in the virome instead of the virome altering the
270 bacterial communities. To better understand this system, future studies will include larger cohort human
271 studies, further *in vitro* and *in vivo* mechanistic experimentation, and attempts at community studies using
272 absolute abundance values instead of relative abundance, which would allow for more accurate community
273 dynamic modeling. Overall, this study provides a conceptual foundation to direct future characterization of
274 the colorectal cancer microbiome at the functional, mechanistic level.

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

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

283 # Materials and Methods

284 ## Analysis Source Code & Data Availability

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

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

288 https://github.com/SchlossLab/Hannigan_CRCVirome_mBio_2018

289 ## Study Design and Patient Sampling

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

300 ## 16S rRNA Gene Sequence Data Acquisition & Processing

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

303 http://www.mothur.org/MicrobiomeBiomarkerCRC

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

309 Whole Metagenomic Library Preparation & Sequencing

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

316 Virus Metagenomic Library Preparation & Sequencing

317 Genomic DNA was extracted from purified virus-like particles (VLPs) from stool samples, using a modified
318 version of a previously published protocol (29, 31, 43, 44). Briefly, an aliquot of stool (~0.1 g) was
319 resuspended in SM buffer (Crystalgen; Catalog #: 221-179) and vortexed to facilitate resuspension. The
320 resuspended stool was centrifuged to remove major particulate debris then filtered through a 0.22-μm filter
321 to remove smaller contaminants. The filtered supernatant was treated with chloroform for ten minutes
322 with gentle shaking, so as to lyse contaminating cells including bacteria, human, fungi, etc. The exposed
323 genomic DNA from the lysed cells was degraded by treating the samples with 5U of DNase for one hour
324 at 37C. DNase was deactivated by incubating the sample at 75C for ten minutes. The DNA was extracted
325 from the purified virus-like particles (VLPs) using the Wizard PCR Purification Preparation Kit (Promega).
326 Disease classes were staggered across purification runs to prevent run variation as a confounding factor.

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

327 As for whole community metagenomes, purified DNA was used to prepare a shotgun sequencing library
328 using the Illumina Nextera XT preparation kit according to the standard kit protocol. The tagmentation time
329 was increased from five minutes to ten minutes to improve DNA fragment length distribution. The PCR
330 cycle number was increased from twelve to eighteen cycles to address the low biomass of the samples, as
331 has been described previously (29). The library was sequenced using one lane of the Illumina HiSeq4000
332 platform and yielded 125 bp paired end reads.

333 Metagenome Quality Control

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

340 Contig Assembly & Abundance

341 Contigs were assembled using paired end read files that were purged of sequences without a corresponding
342 pair (e.g. one read removed due to low quality). The Megahit program (v1.0.6) was used to assemble contigs
343 for each sample using a minimum contig length of 1000 bp and iterating assemblies from 21-mers to 101-mers
344 by 20 (48). Contigs from the virus and whole metagenomic sample sets were concatenated within their
345 respective groups. Abundance of the contigs within each sample was calculated by aligning sequences
346 back to the concatenated contig files using the bowtie2 global aligner (v2.2.1), with a 25 bp seed length and
347 an allowance of one mismatch (49). Abundance was corrected for contig reference length and the number of
348 contigs included in each operational genomic unit. Abundance was also corrected for uneven sampling depth
349 by randomly sub-sampling virome and whole metagenomes to 1,000,000 and 500,000 reads, respectively,

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

350 and by removing samples with fewer total reads than the threshold. Thresholds were set for maximizing
351 sequence information while minimizing numbers of lost samples.

352 Operational Genomic Unit Classification

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

359 Diversity

360 Alpha and beta diversity were calculated using the operational viral unit abundance profiles for each sample.
361 Sequences were rarefied to 100,000 sequences. Samples with less than the cutoff were removed from the
362 analysis. Alpha diversity was calculated using the Shannon diversity and richness metrics. Beta diversity
363 was calculated using the Bray-Curtis metric (mean of 25 random sub-sampling iterations), and the statistical
364 significance between the disease state clusters was assessed using an analysis of similarity (ANOSIM) with
365 a post-hoc multivariate Tukey test. All diversity calculations were performed in R using the Vegan package
366 (51).

367 Classification Modeling

368 Classification modeling was performed in R using the Caret package (52). OTU, OVU, and OGU abundance
369 data was preprocessed by removing features (OTUs, OVUs, and OGUs) that were present in less than thirty
370 of the samples. This served both as an effective feature reduction technique and made the calculations
371 computationally feasible. The binary random forest model was trained using the Area Under the receiver

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

372 operating characteristic Curve (AUC) and the three-class random forest model was trained using the
373 mean AUC. Both were validated using five-fold nested cross validation to prevent over-fitting on the tuning
374 paramters. Each training set was repeated five times, and the model was tuned for mtry values. For
375 consistency and accurate comparison between feature groups (e.g., bacteria, viruses), the sample model
376 parameters were used for each group. The maximum AUC during training was recorded across twenty
377 iterations of each group model to test the significance of the differences between feature set performance.
378 Statistical significance was evaluated using a Wilcoxon test between two categories, or a pairwise Wilcoxon
379 test with Bonferroni corrected p-values when comparing more than two categories.

380 Taxonomic Identification of Operational Genomic Units

381 Operational viral units (OVUs) were taxonomically identified using a reference database consisting of
382 all bacteriophage and eukaryotic virus genomes present in the European Nucleotide Archives. The
383 longest contiguous sequence in each operational genomic unit was used as a representative sequence for
384 classification, as described previously (53). Each representative sequence was aligned to the reference
385 genome database using the tblastx alignment algorithm (v2.2.27) and a strict similarity threshold (e-value
386 < 1e-25) (54). Annotation was interpreted as phage, eukaryotic virus, or unknown. As an additional quality
387 control step, these OVUs were also aligned to the bacterial reference genome set from the European
388 Nucleotide Archives using the blastn algorithm (e-value < 1e-25) and OVUs with similarity to bacterial
389 genomes and not viral genomes were removed from analysis.

390 Ecological Network Analysis & Correlations

391 The ecological network of the bacterial and phage operational genomic units was constructed and analyzed
392 as previously described (34). Briefly, a random forest model was used to predict interactions between
393 bacterial and phage genomic units, and those interactions were recorded in a graph database using *neo4j*
394 graph databasing software (v2.3.1). The degree of phage centrality was quantified using the alpha centrality

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

395 metric in the igraph CRAN package. A Spearman correlation was performed between model importance and
396 phage centrality scores.

397 Phage Replication Style Identification

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

409 Funding Information

410 GD Hannigan was supported in part by the Molecular Mechanisms in Microbial Pathogenesis Training
411 Program (T32 AI007528). PD Schloss was supported by funding from the National Institutes of Health
412 (P30DK034933). MT Ruffin was supported by funding from the National Institutes of Health (5U01CA86400).
413 The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

414 # Acknowledgments

415 The authors thank the Schloss lab members for their underlying contributions, and the Great Lakes-New
416 England Early Detection Research Network for providing the fecal samples that were used in this study. The
417 authors also thank the participants in the study cohort.

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

418 # References

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

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

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

425 4. Chang Y, Cesarman E, Pessin MS, Lee F, Culpepper J, Knowles DM, Moore PS. 1994. Identification of
426 herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma. Science 266:1865–1869.

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

429 6. Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pašić L, Thingstad TF, Rohwer F, Mira A.
430 2009. Explaining microbial population genomics through phage predation. Nature Reviews Microbiology
431 7:828–836.

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

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

436 9. Garrett WS. 2015. Cancer and the microbiota. Science 348:80–86.

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

439 11. Arthur JC, Perez-Chanona E, Mühlbauer M, Tomkovich S, Uronis JM, Fan T-J, Campbell BJ, Abujamel
440 T, Dogan B, Rogers AB, Rhodes JM, Stintzi A, Simpson KW, Hansen JJ, Keku TO, Fodor AA, Jobin C. 2012.

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

441 Intestinal inflammation targets cancer-inducing activity of the microbiota. Science 338:120–123.

442 12. Ly M, Abeles SR, Boehm TK, Robles-Sikisaka R, Naidu M, Santiago-Rodriguez T, Pride DT. 2014. Altered
443 Oral Viral Ecology in Association with Periodontal Disease. mBio 5:e01133–14–e01133–14.

444 13. Monaco CL, Gootenberg DB, Zhao G, Handley SA, Ghebremichael MS, Lim ES, Lankowski A, Baldridge
445 MT, Wilen CB, Flagg M, Norman JM, Keller BC, Luévano JM, Wang D, Boum Y, Martin JN, Hunt PW,
446 Bangsberg DR, Siedner MJ, Kwon DS, Virgin HW. 2016. Altered Virome and Bacterial Microbiome in
447 Human Immunodeficiency Virus-Associated Acquired Immunodeficiency Syndrome. Cell Host and Microbe
448 19:311–322.

449 14. Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, Silva J, Tammadoni S, Nosrat B, Conrad D,
450 Rohwer F. 2009. Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and
451 non-cystic fibrosis individuals. PLOS ONE 4:e7370.

452 15. Abeles SR, Ly M, Santiago-Rodriguez TM, Pride DT. 2015. Effects of Long Term Antibiotic Therapy on
453 Human Oral and Fecal Viromes. PLOS ONE 10:e0134941.

454 16. Modi SR, Lee HH, Spina CS, Collins JJ. 2013. Antibiotic treatment expands the resistance reservoir and
455 ecological network of the phage metagenome. Nature 499:219–222.

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

458 18. Norman JM, Handley SA, Baldridge MT, Droit L, Liu CY, Keller BC, Kambal A, Monaco CL, Zhao G,
459 Fleshner P, Stappenbeck TS, McGovern DPB, Keshavarzian A, Mutlu EA, Sauk J, Gevers D, Xavier RJ,
460 Wang D, Parkes M, Virgin HW. 2015. Disease-specific alterations in the enteric virome in inflammatory
461 bowel disease. Cell 160:447–460.

462 19. Siegel R, Desantis C, Jemal A. 2014. Colorectal cancer statistics, 2014. CA: a cancer journal for
463 clinicians 64:104–117.

464 20. Zackular JP, Baxter NT, Chen GY, Schloss PD. 2016. Manipulation of the Gut Microbiota Reveals Role

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

465 in Colon Tumorigenesis. mSphere 1:e00001–15.

466 21. Dejea CM, Wick EC, Hechenbleikner EM, White JR, Mark Welch JL, Rossetti BJ, Peterson SN, Snesrud
467 EC, Borisy GG, Lazarev M, Stein E, Vadivelu J, Roslani AC, Malik AA, Wanyiri JW, Goh KL, Thevambiga
468 I, Fu K, Wan F, Llosa N, Housseau F, Romans K, Wu X, McAllister FM, Wu S, Vogelstein B, Kinzler KW,
469 Pardoll DM, Sears CL. 2014. Microbiota organization is a distinct feature of proximal colorectal cancers.
470 Proceedings of the National Academy of Sciences of the United States of America 111:18321–18326.

471 22. Flynn KJ, Baxter NT, Schloss PD. 2016. Metabolic and Community Synergy of Oral Bacteria in Colorectal
472 Cancer. mSphere 1:e00102–16.

473 23. Baxter NT, Ruffin MT, Rogers MAM, Schloss PD. 2016. Microbiota-based model improves the sensitivity
474 of fecal immunochemical test for detecting colonic lesions. Genome medicine 8:37.

475 24. Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, Amiot A, Böhm J, Brunetti F, Habermann
476 N, Hercog R, Koch M, Luciani A, Mende DR, Schneider MA, Schrotz-King P, Tournigand C, Tran Van Nhieu
477 J, Yamada T, Zimmermann J, Benes V, Kloor M, Ulrich CM, Knebel Doeberitz M von, Sobhani I, Bork P.
478 2014. Potential of fecal microbiota for early-stage detection of colorectal cancer. Molecular systems biology
479 10:766–766.

480 25. Fearon ER. 2011. Molecular genetics of colorectal cancer. Annual review of pathology 6:479–507.

481 26. Levin B, Lieberman DA, McFarland B, Smith RA, Brooks D, Andrews KS, Dash C, Giardiello FM, Glick
482 S, Levin TR, Pickhardt P, Rex DK, Thorson A, Winawer SJ, American Cancer Society Colorectal Cancer
483 Advisory Group, US Multi-Society Task Force, American College of Radiology Colon Cancer Committee.
484 2008. Screening and surveillance for the early detection of colorectal cancer and adenomatous polyps,
485 2008: a joint guideline from the American Cancer Society, the US Multi-Society Task Force on Colorectal
486 Cancer, and the American College of Radiology., pp. 130–160. *In* CA: A cancer journal for clinicians. The
487 University of Texas MD Anderson Cancer Center, Houston, TX, USA. John Wiley & Sons, Ltd.

488 27. Zauber AG. 2015. The impact of screening on colorectal cancer mortality and incidence: has it really

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

489 made a difference? Digestive diseases and sciences 60:681–691.

490 28. Pedulla ML, Ford ME, Houtz JM, Karthikeyan T, Wadsworth C, Lewis JA, Jacobs-Sera D, Falbo J, Gross
491 J, Pannunzio NR, Brucker W, Kumar V, Kandasamy J, Keenan L, Bardarov S, Kriakov J, Lawrence JG,
492 Jacobs WR, Hendrix RW, Hatfull GF. 2003. Origins of highly mosaic mycobacteriophage genomes. Cell
493 113:171–182.

494 29. Hannigan GD, Meisel JS, Tyldsley AS, Zheng Q, Hodkinson BP, SanMiguel AJ, Minot S, Bushman FD,
495 Grice EA. 2015. The Human Skin Double-Stranded DNA Virome: Topographical and Temporal Diversity,
496 Genetic Enrichment, and Dynamic Associations with the Host Microbiome. mBio 6:e01578–15.

497 30. Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, Chaffron S, Cruaud C, Vargas
498 C de, Gasol JM, Gorsky G, Gregory AC, Guidi L, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Poulos
499 BT, Schwenck SM, Speich S, Dimier C, Kandels-Lewis S, Picheral M, Searson S, Tara Oceans Coordinators,
500 Bork P, Bowler C, Sunagawa S, Wincker P, Karsenti E, Sullivan MB. 2015. Ocean plankton. Patterns and
501 ecological drivers of ocean viral communities. Science 348:1261498–1261498.

502 31. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, Lewis JD, Bushman FD. 2011. The human gut
503 virome: Inter-individual variation and dynamic response to diet. Genome Research 21:1616–1625.

504 32. Hannigan GD, Zheng Q, Meisel JS, Minot SS, Bushman FD, Grice EA. 2017. Evolutionary and functional
505 implications of hypervariable loci within the skin virome. PeerJ 5:e2959.

506 33. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI. 2010. Viruses in the faecal
507 microbiota of monozygotic twins and their mothers. Nature 466:334–338.

508 34. Hannigan GD, Duhaime MB, Koutra D, Schloss PD. 2018. Biogeography and environmental conditions
509 shape bacteriophage-bacteria networks across the human microbiome. PLoS Computational Biology
510 14:e1006099.

511 35. Lengeling A, Mahajan A, Gally DL. 2013. Bacteriophages as Pathogens and Immune Modulators? mBio

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

512 4:e00868–13–e00868–13.

513 36. G rski A, Mi dzybrodzki R, Borysowski J, D browska K, Wierzbicki P, Ohams M, Korczak-Kowalska
514 G yna, Olszowska-Zaremba N, usiak-Szelachowska M, K ak M, Jo czyk E, Kaniuga E, Go a A, Purchla S,
515 Weber-D browska B, Letkiewicz S awomir, Fortuna W, Szufnarowski K, Pawe czyk Z aw, Rog P, K osowska D.
516 2012. Phage as a Modulator of Immune Responses, pp. 41–71. *In* Bacteriophages, part b. Bacteriophage
517 Laboratory, Ludwik Hirszfeld Institute of Immunology; Experimental Therapy, Polish Academy of Sciences,
518 Wrocław, Poland. agorski@ikp.pl; Elsevier.

519 37. Rossmann FS, Racek T, Wobser D, Puchalka J, Rabener EM, Reiger M, Hendrickx APA, Diederich
520 A-K, Jung K, Klein C, Huebner J. 2015. Phage-mediated Dispersal of Biofilm and Distribution of Bacterial
521 Virulence Genes Is Induced by Quorum Sensing. PLoS Pathogens 11:e1004653–17.

522 38. Brockhurst MA, Koskella B. 2013. Experimental coevolution of species interactions. Trends in ecology
523 & evolution 28:367–375.

524 39. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. 2013. Development of a dual-index
525 sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina
526 sequencing platform. Applied and Environmental Microbiology 79:5112–5120.

527 40. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB,
528 Parks DH, Robinson CJ, Sahl JW, Stres B, Thallinger GG, Van Horn DJ, Weber CF. 2009. Introducing
529 mothur: open-source, platform-independent, community-supported software for describing and comparing
530 microbial communities. Applied and Environmental Microbiology 75:7537–7541.

531 41. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO. 2007. SILVA: a
532 comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible
533 with ARB. Nucleic Acids Research 35:7188–7196.

534 42. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. 2011. UCHIME improves sensitivity and speed
535 of chimera detection. Bioinformatics 27:2194–2200.

536 43. Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F. 2009. Laboratory procedures to generate viral

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

537 metagenomes. Nature protocols 4:470–483.

538 44. Kleiner M, Hooper LV, Duerkop BA. 2015. Evaluation of methods to purify virus-like particles for
539 metagenomic sequencing of intestinal viromes. BMC Genomics 16:7.

540 45. Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads.
541 EMBnetjournal 17:10.

542 46. Hannon GJ. 2010. FASTX-Toolkit GNU Affero General Public License.

543 47. Schmieder R, Edwards R. 2011. Fast identification and removal of sequence contamination from genomic
544 and metagenomic datasets. PLOS ONE 6:e17288.

545 48. Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, Yamashita H, Lam T-W. 2016. MEGAHIT v1.0:
546 A fast and scalable metagenome assembler driven by advanced methodologies and community practices.
547 METHODS 102:3–11.

548 49. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods
549 9:357–359.

550 50. Alneberg J, Bjarnason BS aacute ri, Bruijn I de, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ,
551 Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. Nature
552 Methods 1–7.

553 51. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, OHara RB, Simpson
554 GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. vegan: Community Ecology Package.

555 52. Kuhn M. 2016. caret: Classification and Regression Training. CRAN.

556 53. Guidi L, Chaffron S, Bittner L, Eveillard D, Larhlimi A, Roux S, Darzi Y, Audic S, Berline L, Brum JR, Coelho
557 LP, Espinoza JCI, Malviya S, Sunagawa S, Dimier C, Kandels-Lewis S, Picheral M, Poulain J, Searson S,
558 Tara Oceans Consortium Coordinators, Stemmann L, Not F, Hingamp P, Speich S, Follows M, Karp-Boss
559 L, Boss E, Ogata H, Pesant S, Weissenbach J, Wincker P, Acinas SG, Bork P, Vargas C de, Iudicone D,
560 Sullivan MB, Raes J, Karsenti E, Bowler C, Gorsky G. 2016. Plankton networks driving carbon export in the

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

561 oligotrophic ocean. Nature 532:465–470.

562 54. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+:

563 architecture and applications. BMC Bioinformatics 10:1.

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

564 # Figure Legends

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.*

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 2: *Results from healthy vs cancer classification models built using virome signatures, bacterial 16S rRNA gene sequence signatures, whole metagenomic signatures, and a combination of virome and 16S rRNA gene sequence signatures. A) An example ROC curve for visualizing the performance of each of the models for classifying stool as coming from either an individual with a cancerous or healthy colon. B) Quantification of the AUC variation for each model, and how it compared to each of the other models based on 15 iterations. A pairwise Wilcoxon test with a false discovery rate multiple hypothesis correction demonstrated that all models are significantly different from each other (p-value < 0.01). C) Mean decrease in accuracy (measurement of importance) of each operational taxonomic unit within the 16S rRNA gene classification model when removed from the classification model. Mean is represented by a point, and bars represent standard error. D) Mean decrease in accuracy of each operational virus unit in the virome classification model. E) Mean decrease in accuracy of each operational genomic unit and operational taxonomic unit in the model using both 16S rRNA gene and virome features.*

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 3: *Lysogenic phage relative abundance in disease states. Phage OVUs were predicted to be either lytic or lysogenic, and the relative abundance of lysogenic phages was quantified and represented as a boxplot. No disease groups were statistically significant.*

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 4: *Relative abundance correlations between bacterial OTUs and virome OVUs. A) Pearson correlation coefficient values between all bacterial OTUs (x-axis) and viral OVUs (y-axis) with blue being positively correlated and red being negatively correlated. Bar plots indicate the viral (left) and bacterial (bottom) operational unit importance in their colorectal cancer classification models, such that the most important units are in the top left corner. B) Magnification of the boxed region in panel (A), highlighting the correlation between the most important bacterial OTUs and virome OVUs. The most important operational units are in the top left corner of the heatmap, and the correlation scale is the same as panel (A). C) Histogram quantifying the frequencies of Pearson correlation coefficients between all bacterial OTUs and virome OVUs.*

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 OVU. A linear regression line was fit to illustrate the correlation (blue) which was found to be statistically significantly and weakly correlated (p-value = 0.00409, R = 0.176).*

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 6: *Final working hypothesis from this study. These panels summarize our thoughts on our results and represent interesting future directions that we predict will build on the presented work. A) Basic model illustrating the connections between the virome, bacterial communities, and colorectal cancer. B) Working hypothesis of how the bacteriophage community is associated with colorectal cancer and the associated bacterial community.*

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

565 # Supplemental Figure Legends

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 or no-DNA negative control. B) Sequence yield following quality control including quality score filtering and human decontamination. Dashed line indicates rarefaction level* (10^6 *reads) in which all samples with lower sequence yields less than this level were excluded from downstream analysis. After rarefaction and removal of samples with less than* 10^6 *reads, 27 healthy, 28 cancerous, 27 adenomatous, and 3 negative control samples remained.*

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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 viral unit (OVU) length and sequence coverage for the virus metagenomic sample set. C) Heated scatter plot demonstrating the distribution of contig coverage and length for the whole metagenomic sample set. D) Scatter plot illustrating the distribution of operational genomic unit (OGU) length and sequence coverage for the whole metagenomic sample set.*

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.*

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.*

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure S5: *Beta-diversity analysis comparing Bray-Curtis dissimilarity between disease state and negative control community structures that were captured following sequence rarefaction. 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.*

bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure S6: *Comparison of bacterial 16S rRNA classification models with and without OTUs whose median relative abundance are greater than zero. A) Classification model performance (measured as area under the curve) for bacteria models using 16S rRNA data both with and without filtering of samples whose median was zero. Significance was calculated using a Wilcoxon rank sum test, and the resulting p-value is shown. The random area under the curve (0.5) is marked with a dashed line. B) Relative abundance of the six bacterial OTUs removed when filtered for OTUs with median relative abundance of zero. OTU relative abundance is separated by healthy (red) and cancerous (grey) samples. Relative abundance of 1% is marked by the dashed line.*

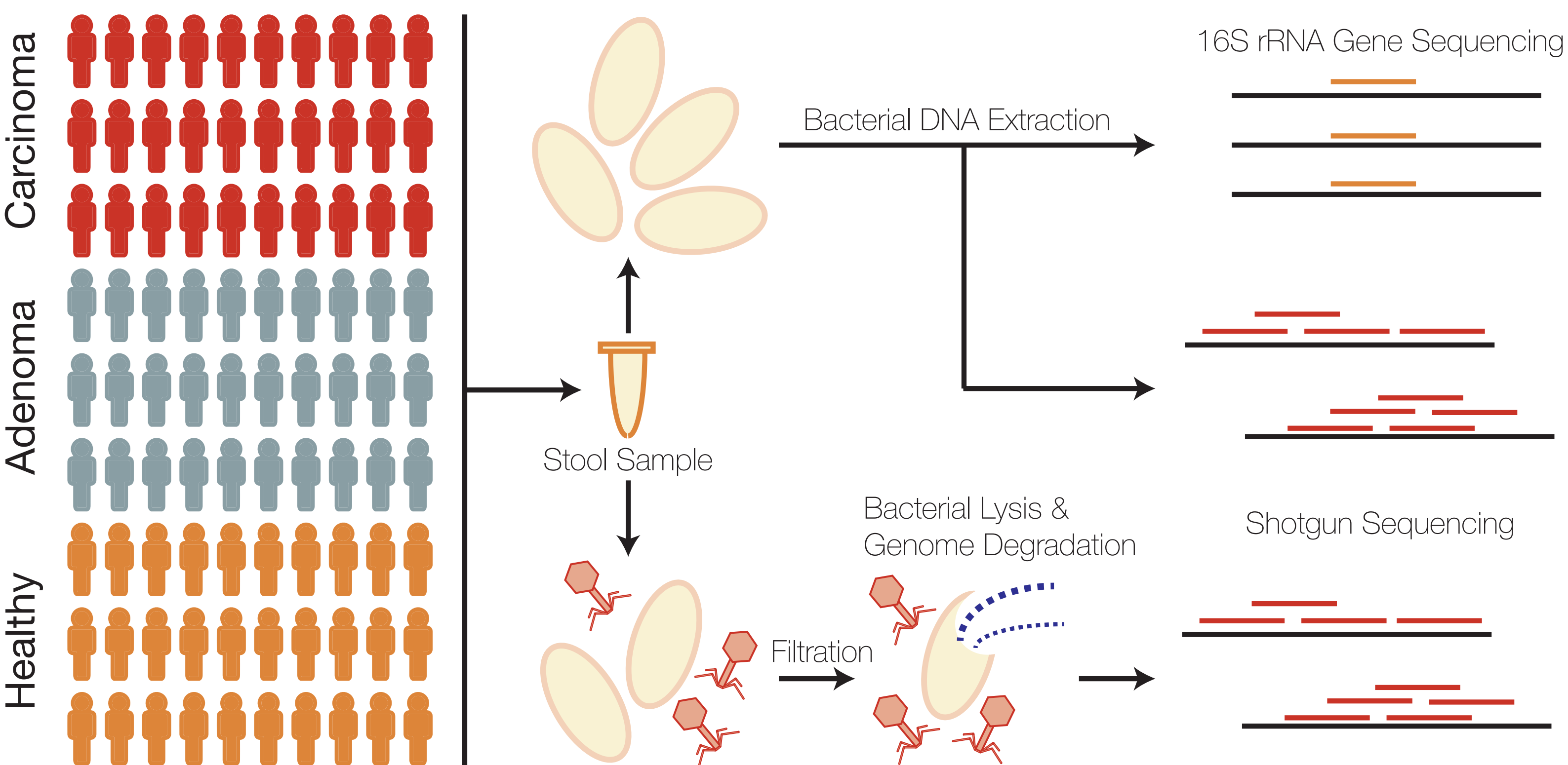
bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure S7: *Transition of colorectal cancer importance through disease progression. A) Virus and B) 16S rRNA gene model performance (AUC) when discriminating all binary combinations of disease types. Blue line represents mean performance from multiple random iterations. C-E) Top ten important phage OVUs when classifying each combination of disease state, as measured by the mean decrease in accuracy metric. Mean is represented by a point, and bars represent standard error. Disease comparison is specified in the top left corner of each panel. F-H) Top ten important bacterial 16S rRNA gene OTUs for classifying each disease state combination.*

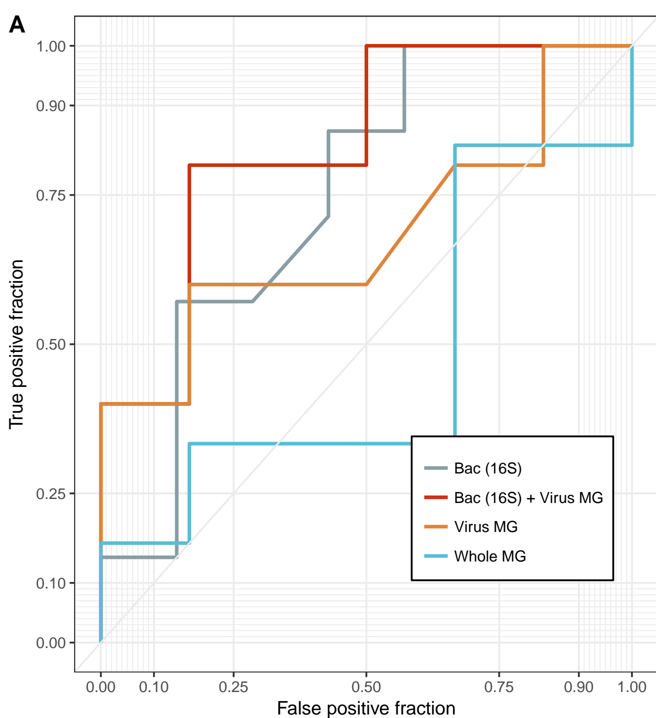
bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

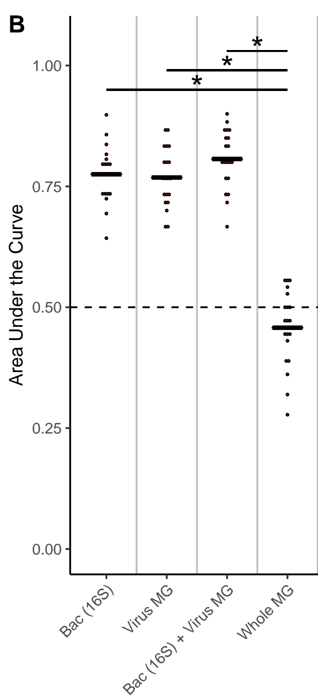
Figure S8: *ROC curves from A) virome and B) bacterial 16S three-class random forest models tuned on mean AUC. Each curve represents the ability of the specified class to be classified against the other two classes. C) Quantification of the mean AUC variation for each model based on 10 model iterations. A pairwise Wilcoxon test with a Bonferroni multiple hypothesis correction demonstrated that the models are significantly different (alpha = 0.01). D) Mean decrease in accuracy when virome operational genomic units and E) bacterial 16S OTUs are removed from the respective three-class classification models. Results based on 25 iterations. F) Relative abundance of the six most important virome OVUs in the model, with the most important on the right. Line indicates abundance mean.*

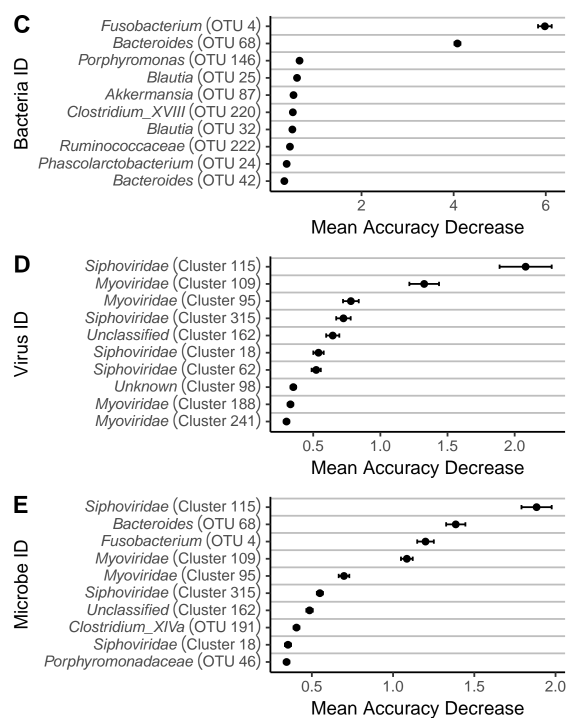
bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



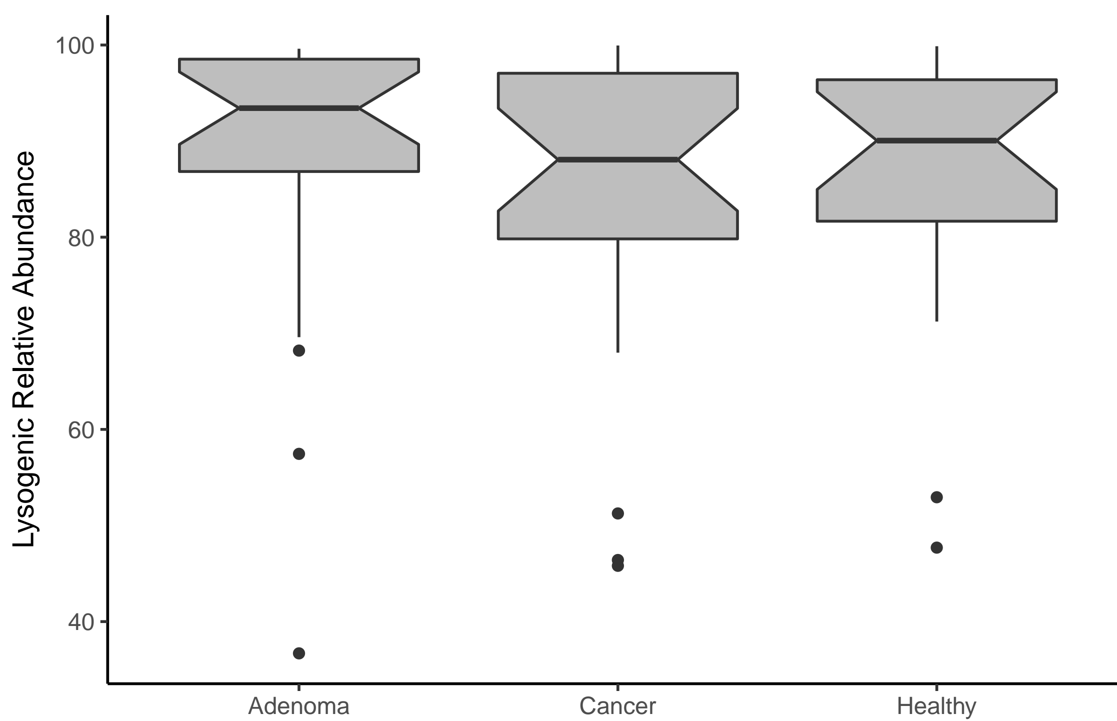
bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



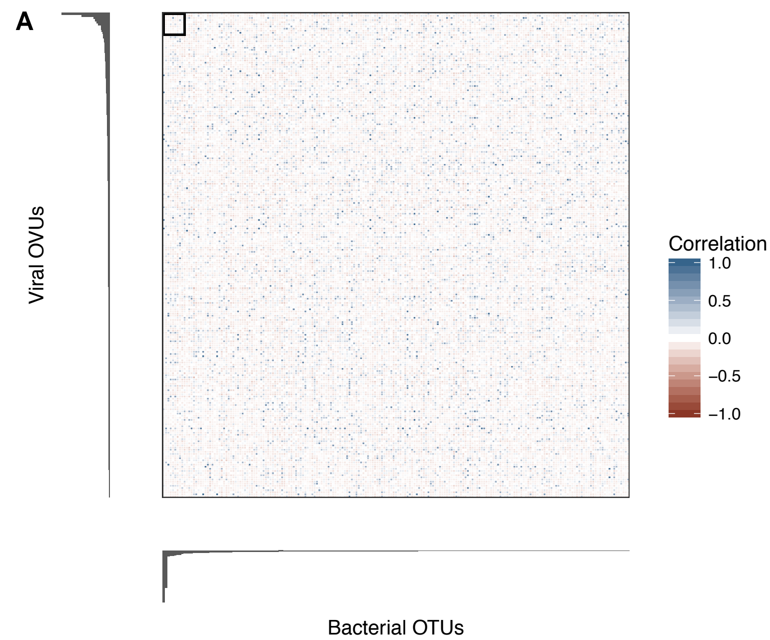


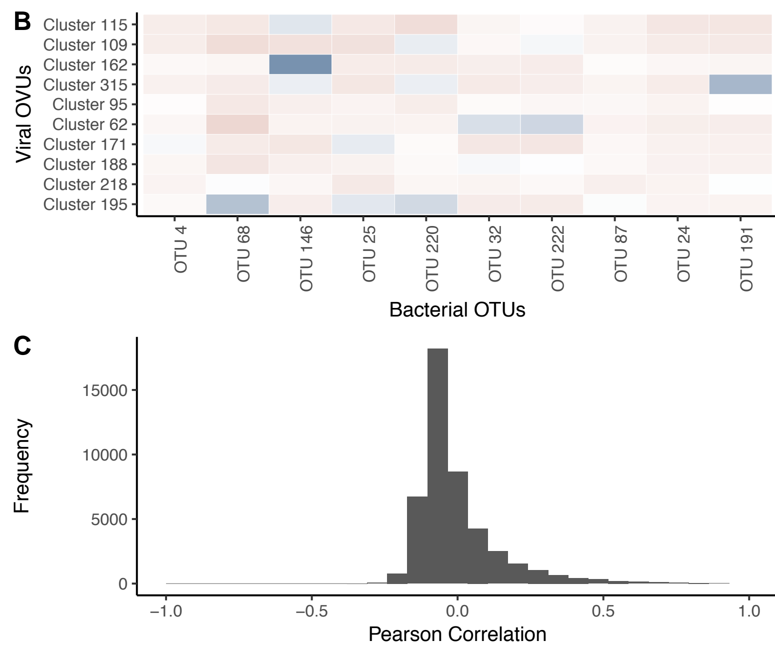


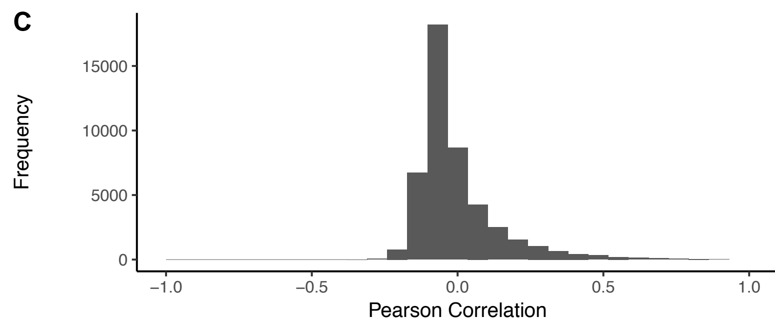
bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



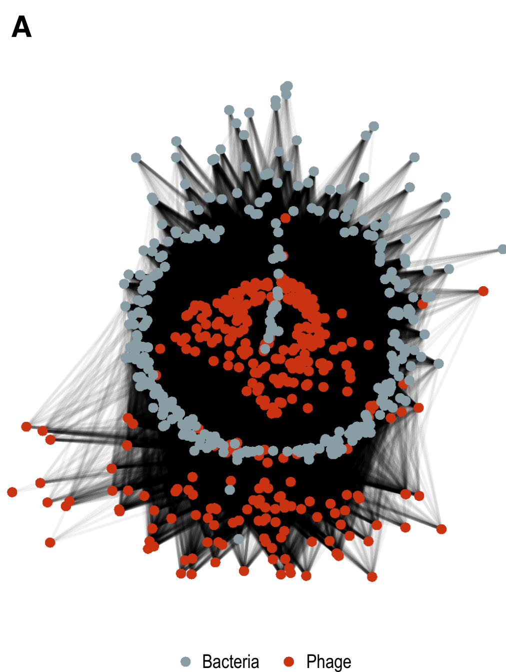
bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

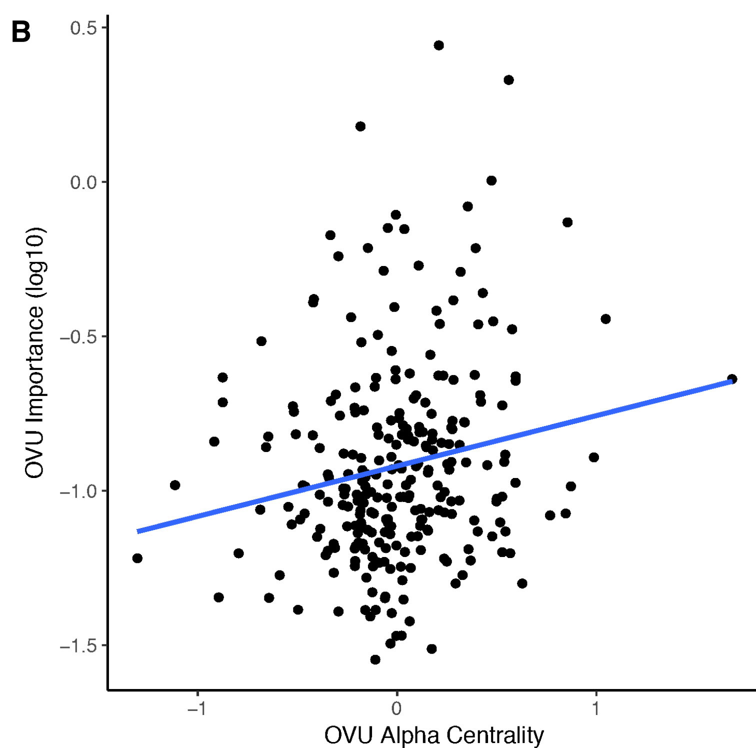




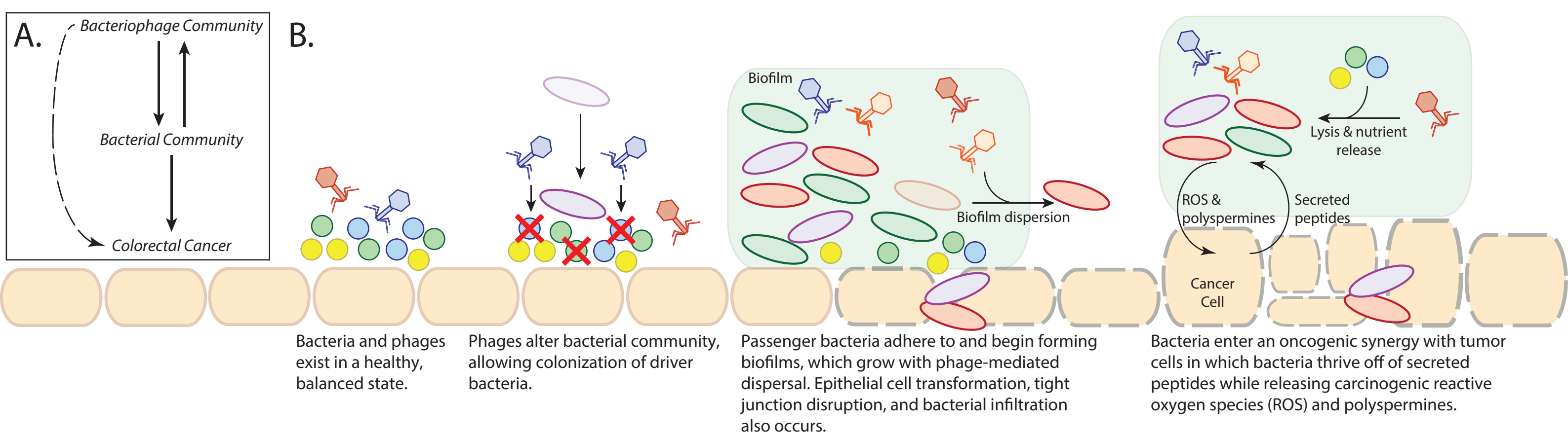


bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

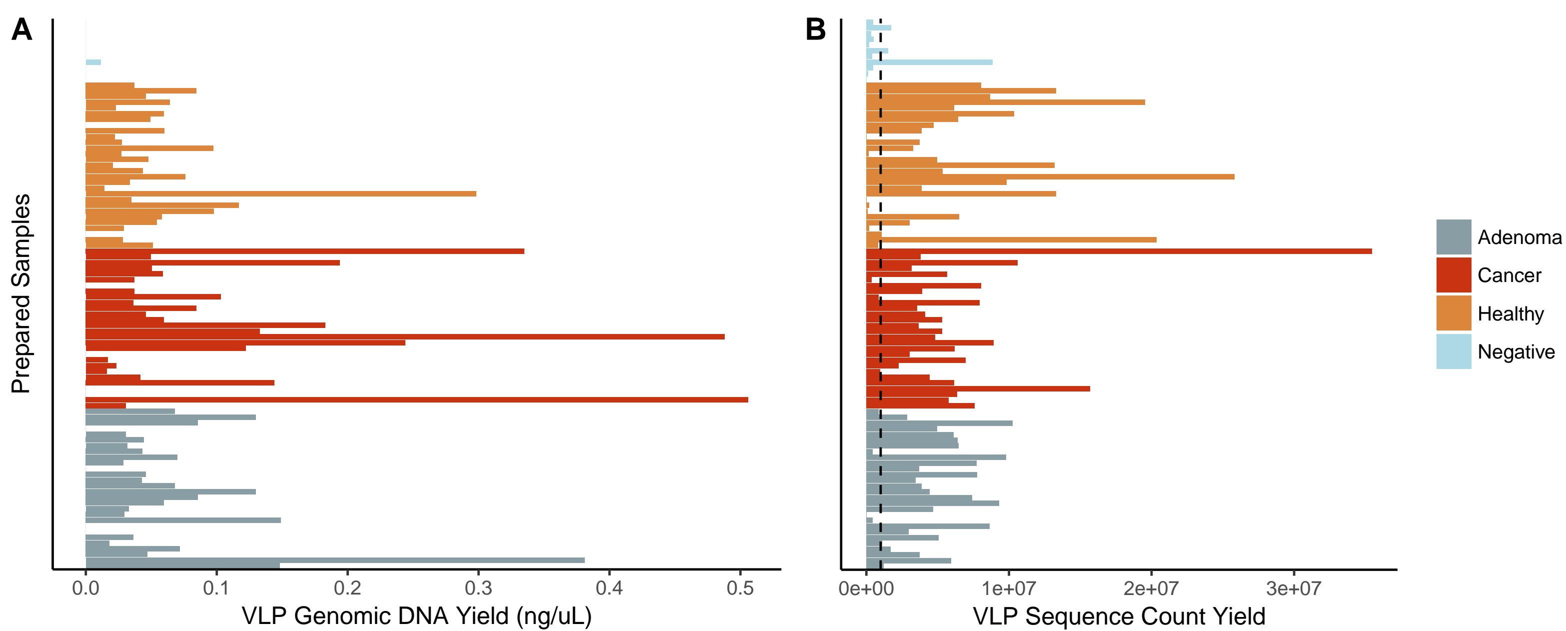




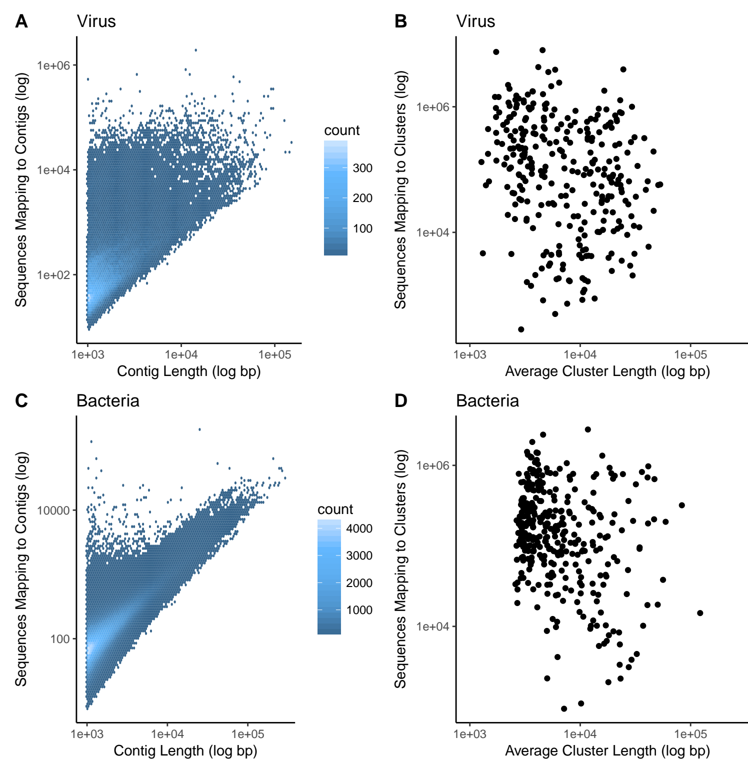
bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



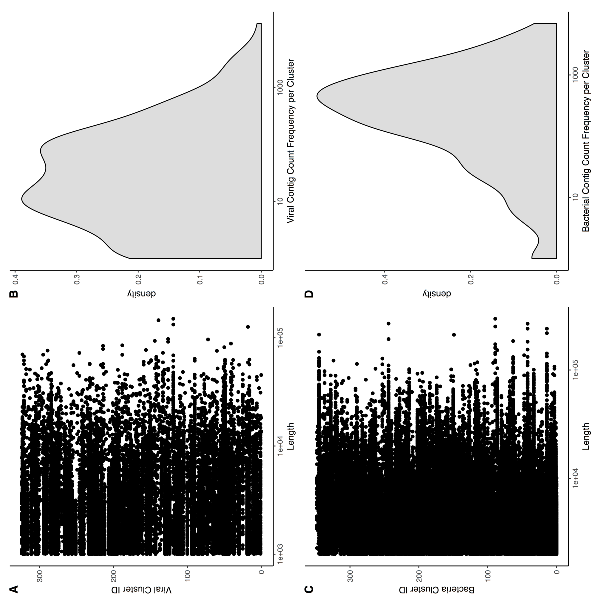
bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



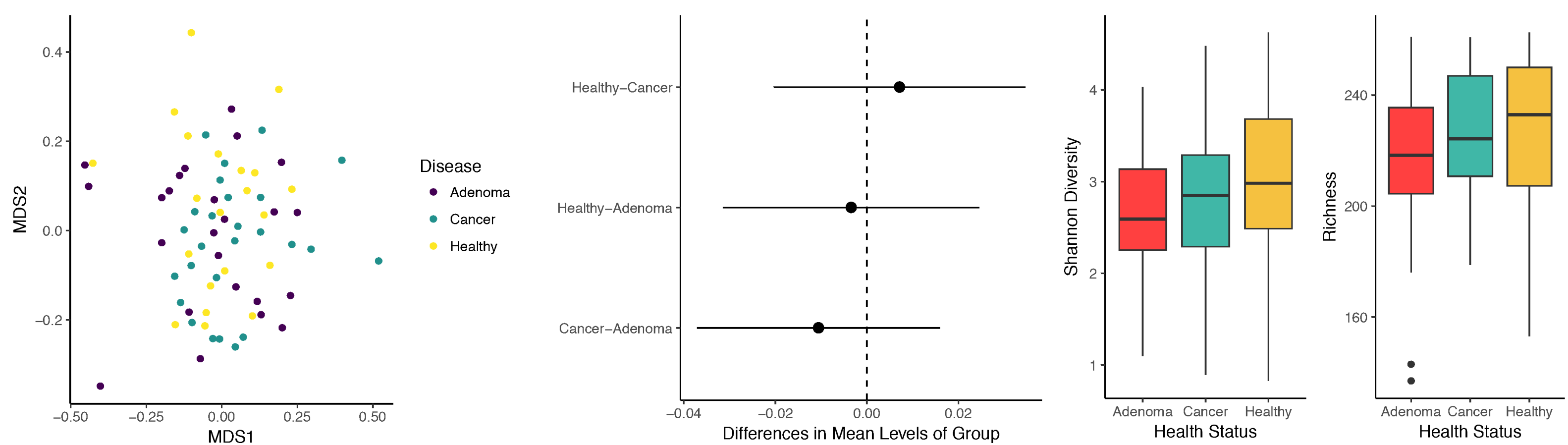
bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



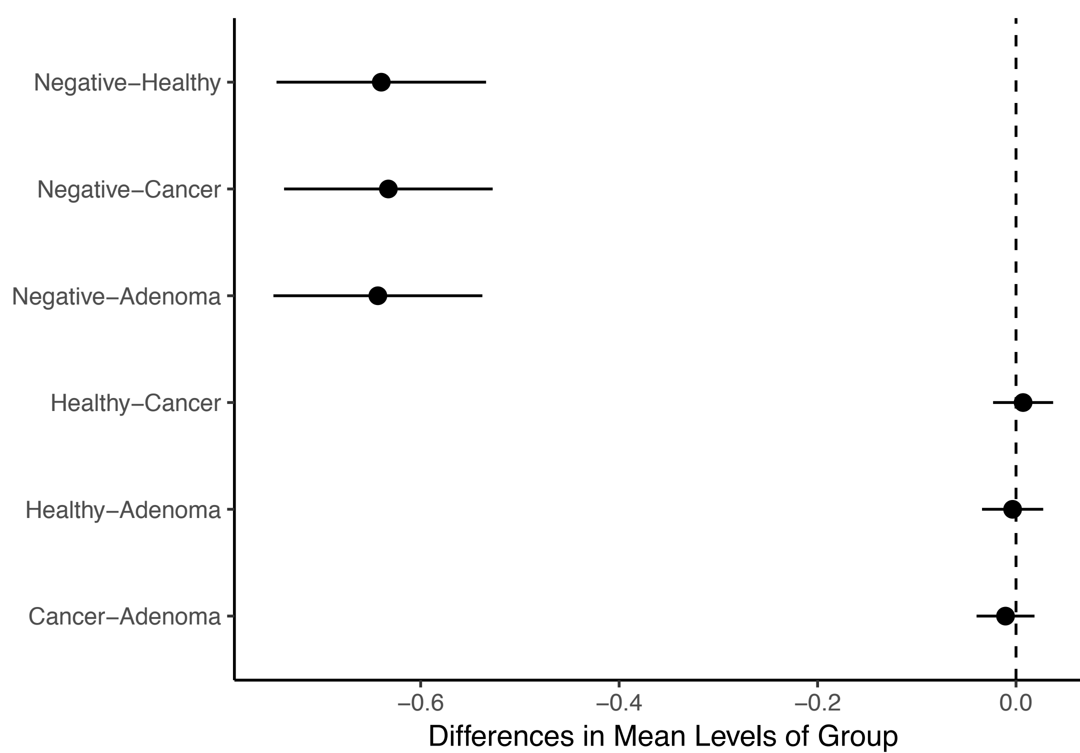
bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



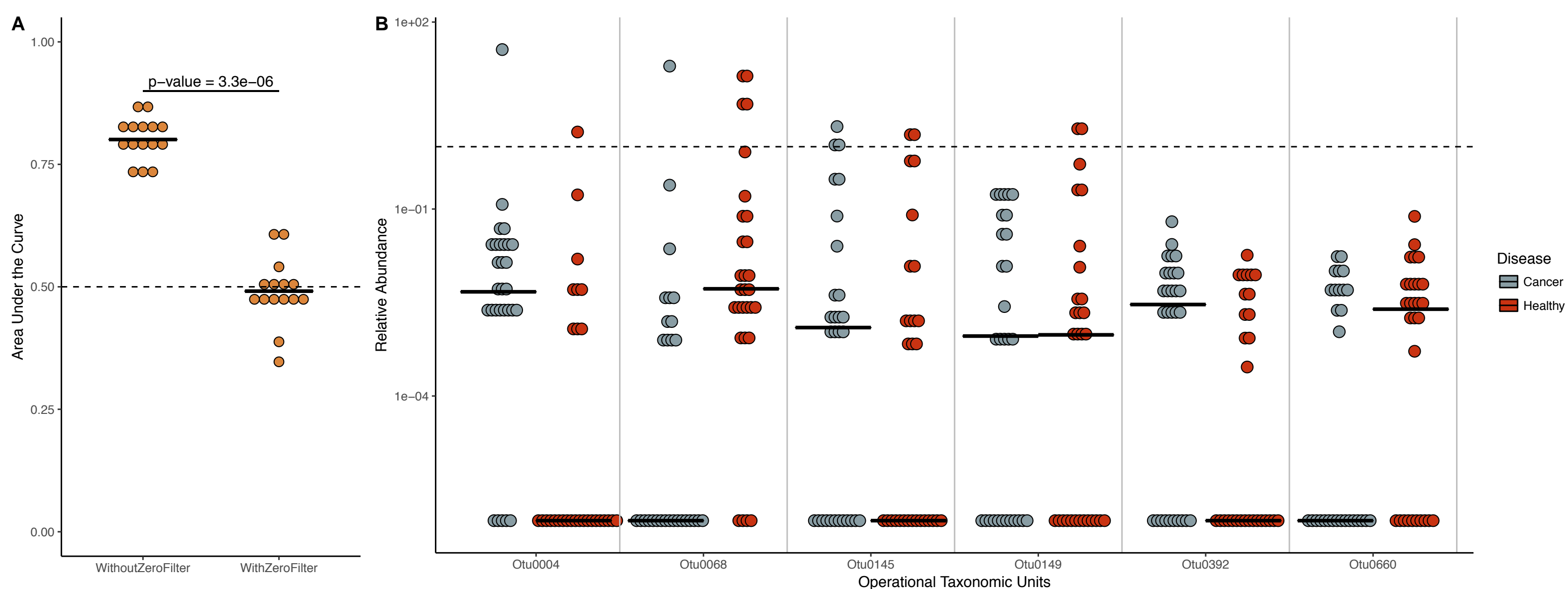
bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



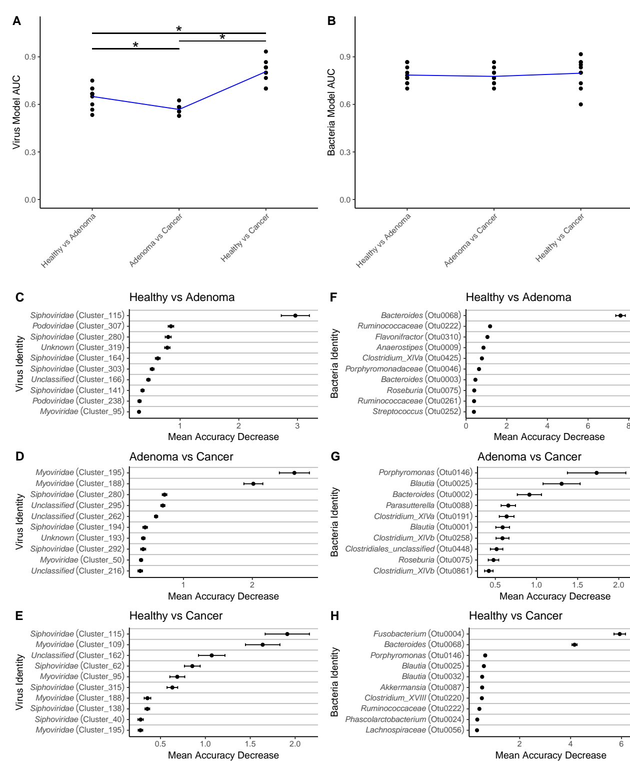
bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/152868; this version posted October 11, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.