Viruses of the Microbiome Outperform Bacteria in Cancer Classification Models

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

Contents

Introduction

Cancer has remained a devastating and persistent plague on humanity despite our profound efforts to diminish its impact. Although we still have a long way to go with treating cancer, we have certainly made some progress in recent years. Perhaps one of the most impactful advances we have made has not been in treating cancer, but rather detecting cancer at an early enough stage so as to make our current treatments as effective as possible. This has been evident in a variety prominant cancers, one of the most notable being colorectal cancer.

Colorectal cancer is the second leading cause of cancer-related deaths in the United States. 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]. Not only is this a leading cancer in the United States, but it is a major source of morbidity and mortality throughout the world. Although it is a major health problem, the impact of colorectal cancer has been greatly diminished by improved screening and prevention efforts.

Developement of colorectal cancer is a stepwise process that begins when healthy tissue developes into a pre-cancerous polyp (i.e. ademona) in the large intestine. If left untreated, the adenoma will devlope into a cancerous lession that can invade and metastisize, leading to grave illness and death. Developement to cancer can be prevented when adenomas are detected and removed during routine screening, often as a colonoscopy. In fact, survival for colorectal cancer patients exceeds 90% when the lessions are detected early and removed. These methods are effective, but their invasiveness has caused a strong lack of compliance, speaking to a need to non-invasive screening methods. One such method is through screening of associated microbial communities.

Although the cause of colorectal cancer remains unclear, its occurance has been strongly associated with gut bacterial communities. This association is so striking that bacterial community signatures have been leveraged as biomarkers to greatly improve colorectal cancer detection[2,3]. While an understanding of colorectal cancer bacterial communities have proven fruitful both for prediction and understanding the underlying etiology, bacteria are only a subset of the microbiome and a more complete picture has yet to be reached. To this end, we evaluated the role the virus component of the microbiome (the virome) might be playing in colorectal cancer developement and its utility as a prognostic marker.

Due to their mutagentic abilities and their propensity for functional manipulation, viruses are strongly associated with, and in many cases cause, cancer. Additionally, because bacteria are thought to play a role in colorectal cancer developement, we hypothesize that bacteriophages could likewise have an indirect impact by acting through bacteria.

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. The implications of these findings are threefold. First, this suggests a biological role for the virome in colorectal cancer developement and that more than bacteria are involved in the process. Second, we present an avenue for even higher performance classification modeling of colorectal cancer using stool samples, an accurate but non-invasive alternative to colorectal cancer screening. Third, this provides early evidence for the importance of studying the virome as a component of the microbiome, especially in cancer.

Results

The Colorectal Cancer Virome Cohort

Our cohort consisted of 90 human subjects, 30 of which served as healthy controls, 30 of which had ademona lesions consistent with the 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[2]. The other half of the stool samples were purified for virus like particles (VLPs) and virome genomic DNA extraction, followed by shotgun metagenomic sequencing. The virus purificaiton allows us to observe the *active virome* because we are only sequencing those viruses that were encapsulated (we do not detect viruses integrated in bacterial genomes).

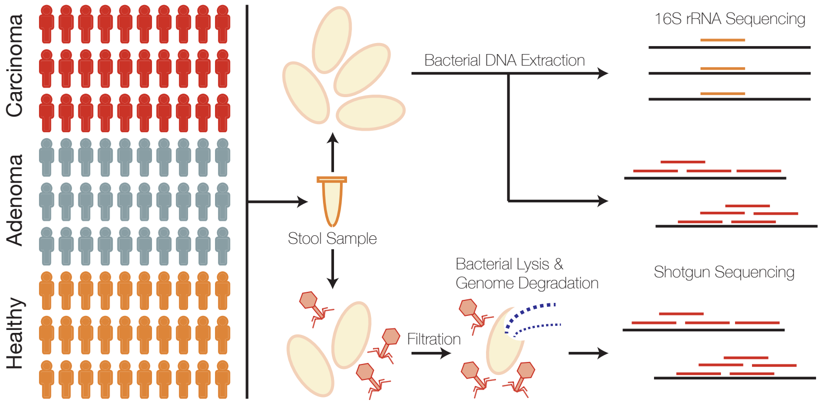

Figure 1: *Cohort and sample processing outline. Thirty subject stool samples were collected from healthy, adenoma (precancer), and carcinoma (cancer) patients. Stool samples were split into two aliquots, the first of which was used for bacterial sequencing and the second which was used for virus sequencing. 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 according to previous studies. Briefly, the stool was resuspended in saline magnesium buffer by rigorous vortexing. Bacteria, human, and other non-viral cells were removed by filtering through a 0.22µm filter, followed by cell lysis with chloroform and degredation of the released genomic DNA with DNase **(Figure 1)**. The resulting genomic DNA was used to prepare a sequencing library with the NexteraXT preparation kit, and was sequenced on the Illumina HiSeq2500 platform. To accomodate the low concentration of input DNA, we used the NexteraXT protocol with 18 PCR cycles instead of 12. 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, indicating successful sequencing of VLP genomic DNA over potential contaminants **(Figure 5)**.

Diversity Alone is Insuffient for Virome-Based Cancer Classification

We began our community analysis by evaluating the diversity of the system and its association with disease. We used beta-diversity to evaluate the differences in the communities between disease states. This allowed us to see how similar the communities were to each other. We utilized the Bray-Curtis dissimilarity metric to evaluate the differences between community 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 centroids of the sample clusters **(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, which further supports the quality of our sample set **(Figure 6 B)**.

Simple diversity metrics were insufficient for capturing the differences in the microbial communities between disease states. This suggests that a more sophisticated approach for understanding the microbial community may be required, beyond dissimilarity metrics.

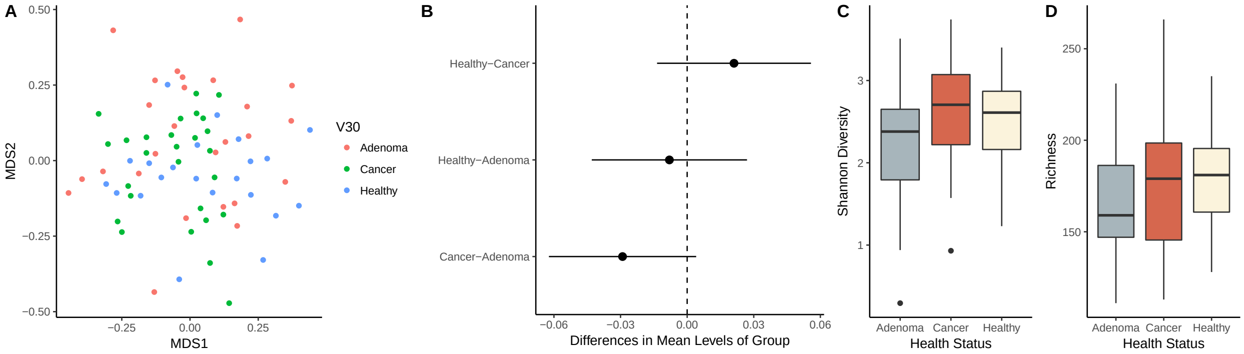

Figure 2: *Beta-diversity comparing disease states of the colorectal virome from stool samples. A) NMDS ordination of community samples, colored by disease state. B) Differences in means between disease group centroids with 95% confidence intervals based on an anosim test with a post hoc multivariate Tukey test. Comparisons in which the intervals cross the zero mean difference line (dashed line) were not significantly different.*

Virome-based cancer classification outperforms bacterial models

Previous work has shown that 16S rRNA community signatures are effective for classifying stool samples as coming from healthy, pre-cancerous, or cancerous individuals[2,3]. This is valuable because it presents a potential alternative screening approach to the invasive colonoscopy. This approach supplements other screening tests such as FIT. The exceptional performance of bacterial signatures in these predictive models suggests a role for bacteria in colorectal cancer. We built off of these findings by evaluating the ability of virus community signatures to classify stool samples and compared performance to models built using bacterial data.

We built and tested random forest models based on virus metagenomic communtiy signatures, whole metagenomic community signatures, and bacterial 16S rRNA gene signatures. 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 both datasets, and the only difference was the data used to learn the model. We confirmed that the model built using bacterial 16S data replicated the findings from the previous 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 greatly outperformed the bacterial model **(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 was the result of shotgun sequencing all of the microbial DNA within the system. This model performed extremely poorly, lending support to our observation that high classificaiton 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. **(Figure ??)**. 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 either.

Identifying important biological factors in colorectal cancer

Our model for classifying stool samples as cancerous or healthy performed significantly better when using virome signatures compared to bacterial 16S. Not only is this important for demonstrating an improved model, but it also suggests an underlying biological importance for viruses in colorectal cancer. We therefore

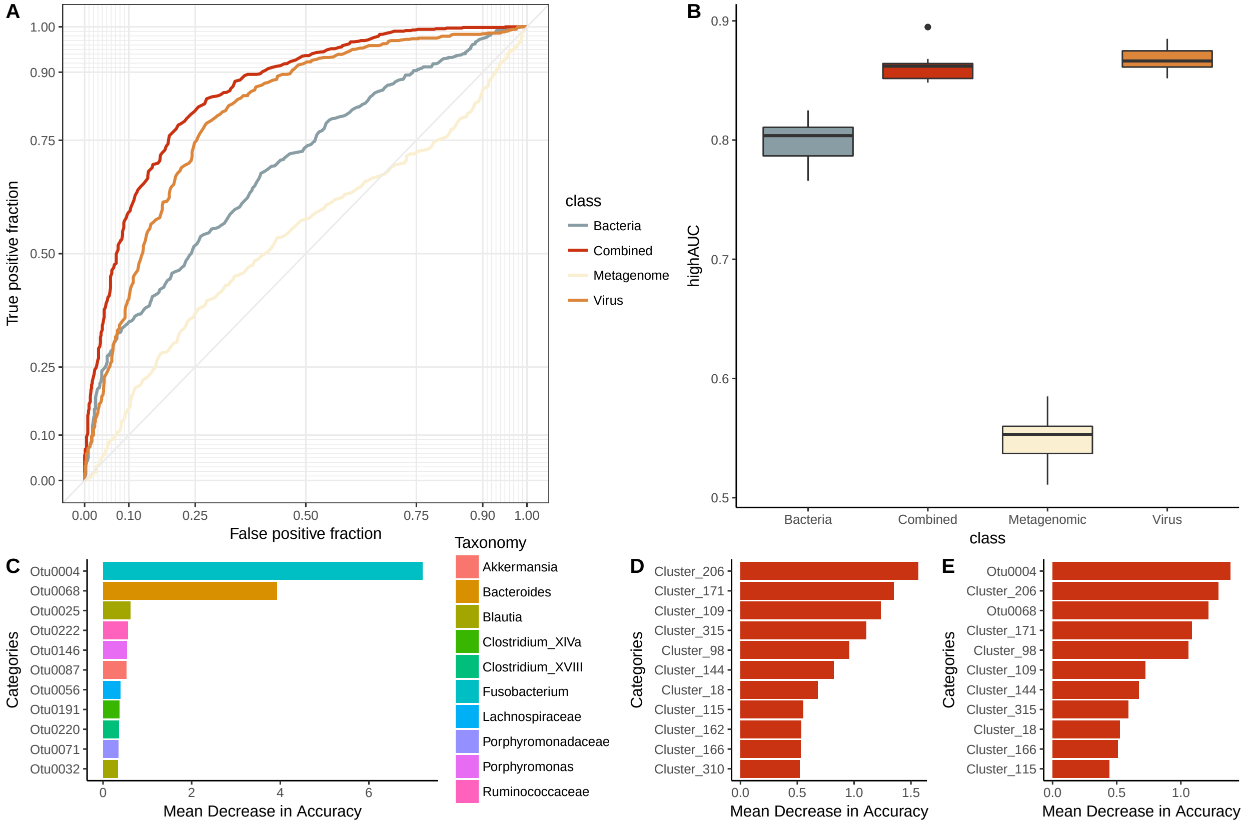

Figure 3: *Results from healthy vs cancer classification models built using virome signatures, bacterial 16S signatures, whole metagenomic signatures, and a combination of virome and 16S signatures. A) ROC curve for visualizing the performance of each of the models for classifying stool as coming from 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 bonferonni 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) Importance values of each operational taxonomic unit within the 16S classification model, colored by taxonomic identities. D) Importance of each operational genomic unit in the virome classification model. E) Importance of each operational genomic unit and operational taxonomic unit in the model using both 16S and virome features.*

used our predictive models to evaluate which viruses were most important for distinguishing between disease states, thus allowing us to identify the agents likely to play biological roles in disease.

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 there was a discreency between the distribution of top variable importance between bacterial and viral models, with the most important bacterial taxa being responsible for 6.5% of the model while the top viral feature was only responsible for 1.5%. The top bacterial unit (as identified by 16S rRNA gene sequence) was fusobacterium, a bacterium long been associated with cancer and hypothesised to play a role in colorectal cancer.

Performance of model upon inclusion of adenoma samples

After evaluating our ability to classify samples as cancerous vs 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 both the bacterial and viral sample sets. Again the virome significantly outperformed the bacterial community signatures. The virus signature allowed for consistently high resolution for each of the disease classes, while the resolution varried with the bacterial signatures. The bacterial signature allowed for high accuracy in classifying cancerous stool apart from pre-cancerous or healthy, but struggled to classify adenomas or healthy samples away from the remainder of the dataset. On average the bacterial model performed poorly compared to the consistently higher viruses.

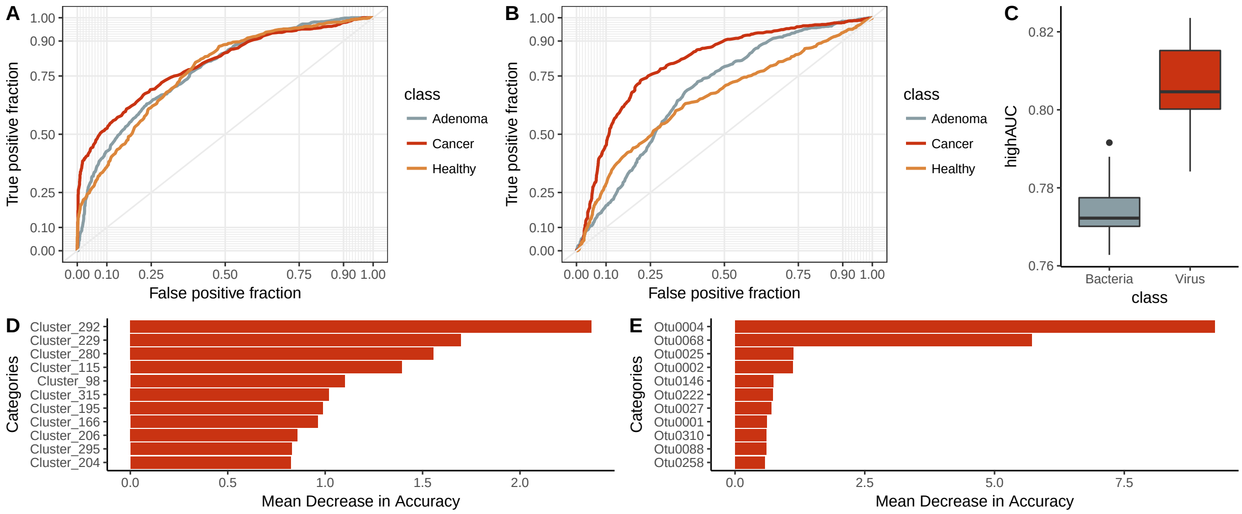

Figure 4: *ROC curves from three-class random forest tuned on mean AUC for A) virus and B) bacterial signatures. Each curve represents the ability of the specified class to be classified against the other two classes.*

Discussion

Conclusions

Methods

Acknowledgements

Conflicts of Interest

Supplemental Figures

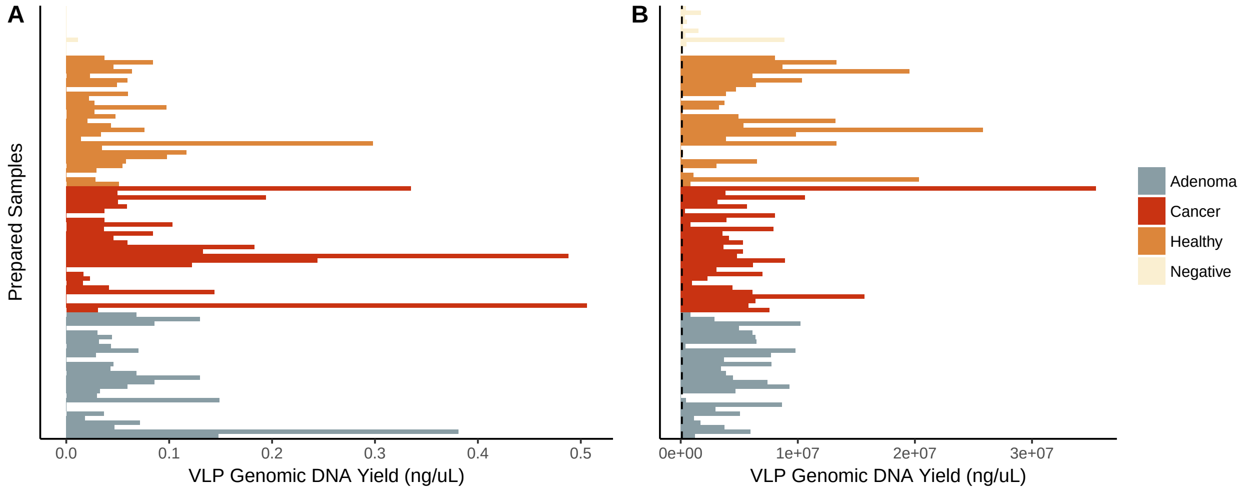

Figure 5: *Basic Quality Control Metrics. A) VLP genomic DNA yield from all sequenced samples. Each bar represents a sample which is grouped and colored by its associated disease group. B) Sequence yield following quality control including quality score filtering and human decontamintion. Dashed line represents the subsampling depth used in the study.*

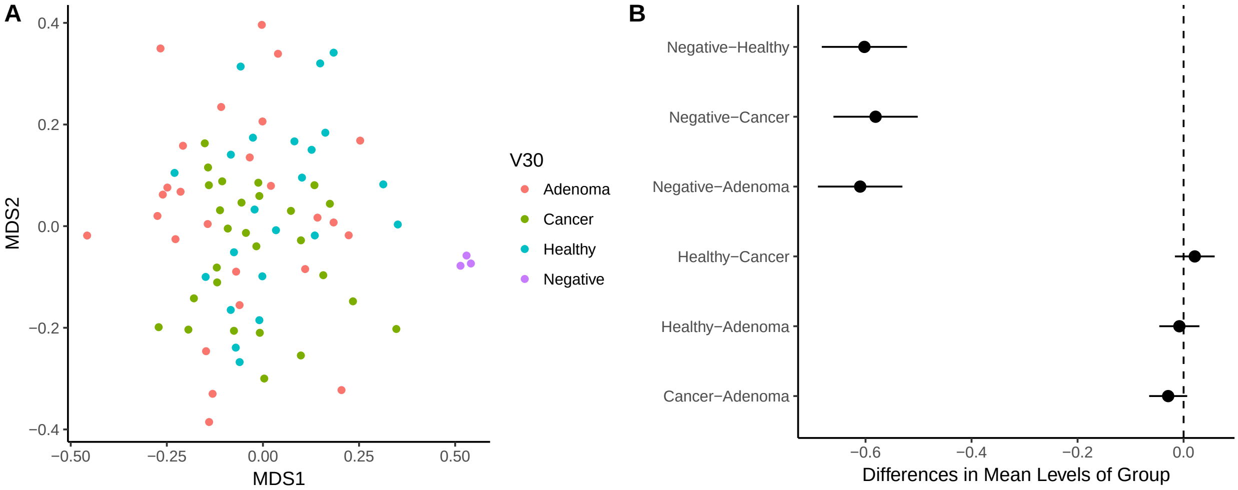

Figure 6: *Beta-diversity comparing disease states of the colorectal virome from stool samples. A) NMDS ordination of community samples, colored by disease state. B) Differences in means between disease group centroids with 95% confidence intervals based on an anosim test with a post hoc multivariate 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. 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).

3. 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).