The Colon Cancer Virome and Its Interactions with the Microbiome

Geoffrey D Hannigan, Melissa B Duhaime, Mack T Ruffin IV, Patrick D Schloss

Contents

Thoughts & Notes **2**

Working Hypothesis **2**

Introduction **2**

Results **2**

The Colorectal Cancer Virome Cohort 2
Changes in Genomic and Functional Diversity in Colorectal Cancer 3
Virome-based cancer classification outperforms bacterial models 3
Performance of cancer vs ademona vs healthy models between microbial signatures 5
Identifying Major Factors in Colorectal Cancer 6

Discussion **6**

Conclusions **6**

Methods **6**

Acknowledgements **6**

Conflicts of Interest **6**

Figures **6**

References **6**

Thoughts & Notes

I think I want this paper to be more focused. Instead of a general survey of the virome, I want to focus on the predictive modeling. How it performs, how it compares to the other sample sets, and what biological information we can get from it. Not only will this make the paper clearer with more of a purpose, but I think it will let me get it done sooner as well.

I think the *title* and general theme should reflect how well viruses (maybe metagenome functionality in general) predict cancer, and what elements are important.

General Landscape -> Predictive Modeling -> Networking to ID Virus OGUs

Working Hypothesis

Like bacteria, it is more of the minor players of the community that make that difference instead of major abundances.

What viruses are always present in cancer but not in healthy?

Introduction

Colorectal cancer is a common and severe disease that 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].

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, other microbial groups have yet to be evaluated and may provide additional insights into the link between disease and microbial communities.

Due to their mutagentic abilities and their general propensity for functional manipulation, viruses are strongly associated with, and in many cases cause cancer. Unfortunately there is not much research

Results

The Colorectal Cancer Virome Cohort

Stool samples were collected in accordance with our approaved IRB protocol. The cohort consisted of 90 human subjects, 30 of which served as healthy controls, 30 that had ademona lesions consistent with the pre-cancerous state, and 30 that 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).

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

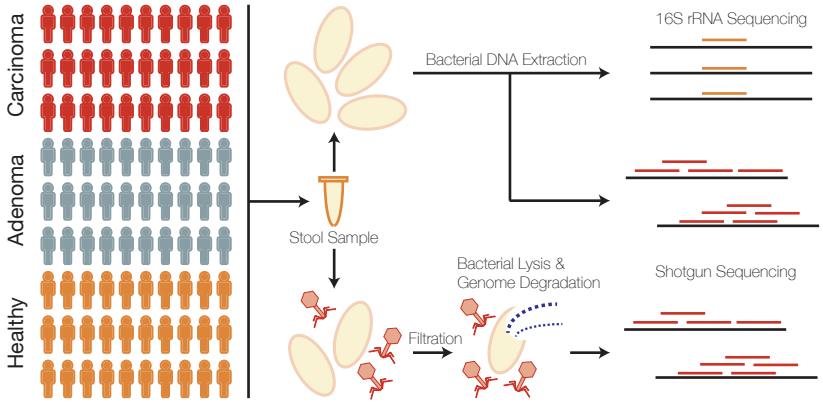

Figure 1: Cohort and sample processing outline. Thirty subject stool samples were collected from healthy, adenoma (precancer), and carcinoma (cancer) patients. Stool samples were split into two aliquots, the first of which was used for bacterial sequencing and the second which was used for virus sequencing.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

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 2)**.

Changes in Genomic and Functional Diversity in Colorectal Cancer

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 3 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 3 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. This further supports the quality of our sample set.

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 and bacteria communtiy 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. Abundance counts were subsampled to the same level and those samples

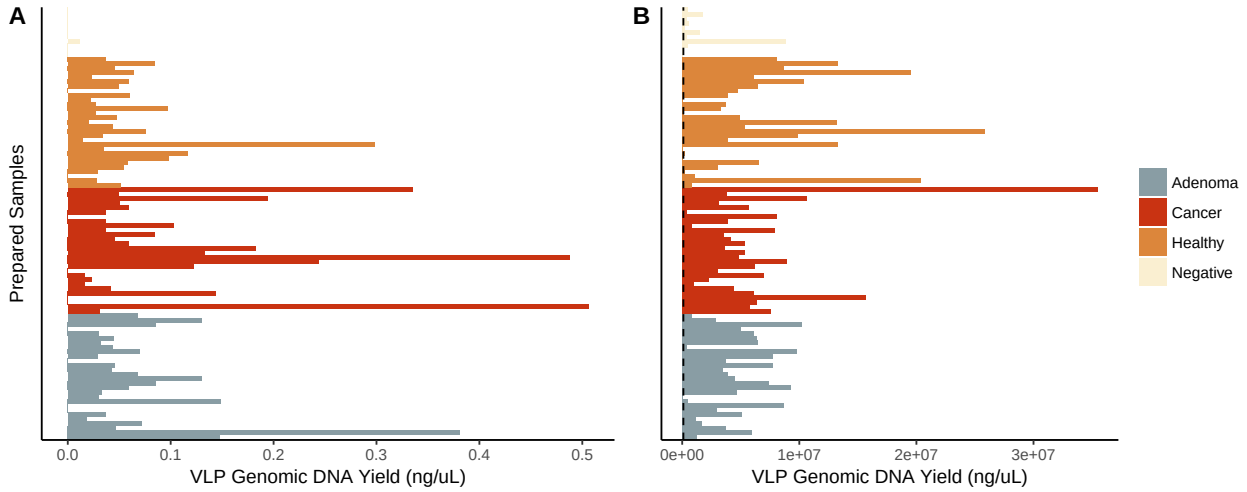

Figure 2: Basic Quality Control Metrics. A) VLP genomic DNA yield from all sequenced samples. Each bar represents a sample which is grouped and colored by its associated disease group. B) Sequence yield following quality control including quality score filtering and human decontamintion. Dashed line represents the subsampling depth used in the study.

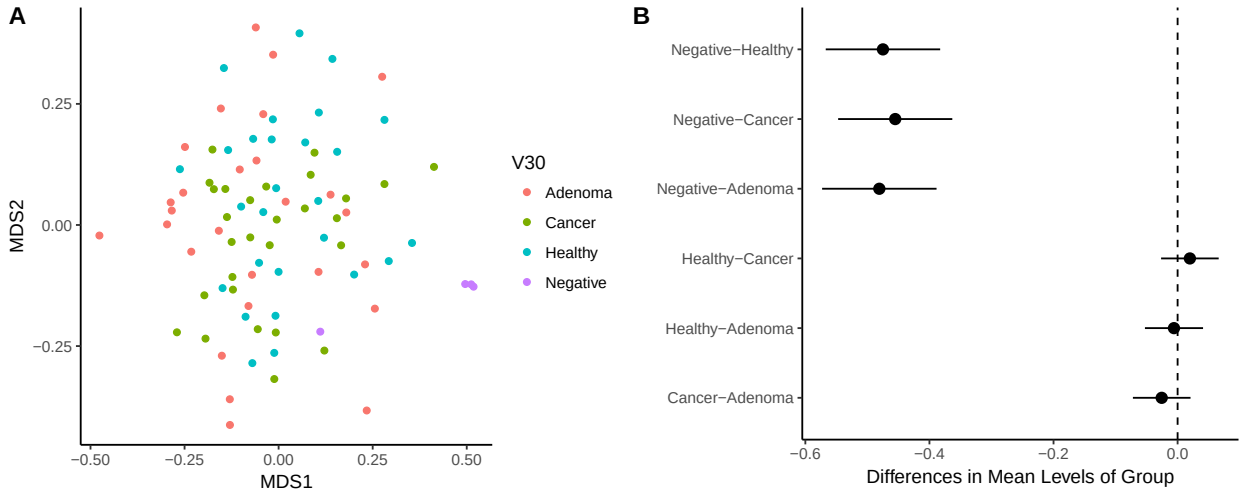

Figure 3: Beta-diversity comparing disease states of the colorectal virome from stool samples. A) NMDS ordination of community samples, colored by disease state. B) Differences in means between disease group centroids with 95% confidence intervals based on an anosim test with a post hoc multivariate Tukey test. Comparisons in which the intervals cross the zero mean difference line (dashed line) were not significantly different.

that failed to reach the defined sequencing threshold were excluded from analysis. The same model approach was used for both datasets, and the only difference was the data used to learn the model. We found that the model based on virus metagenomic signatures greatly outperformed the model built on bacterial 16S data **(Figure 1)**. This supports the utility of using virus communities as cancer biomarkers and suggests viruses play a role in colorectal cancer.

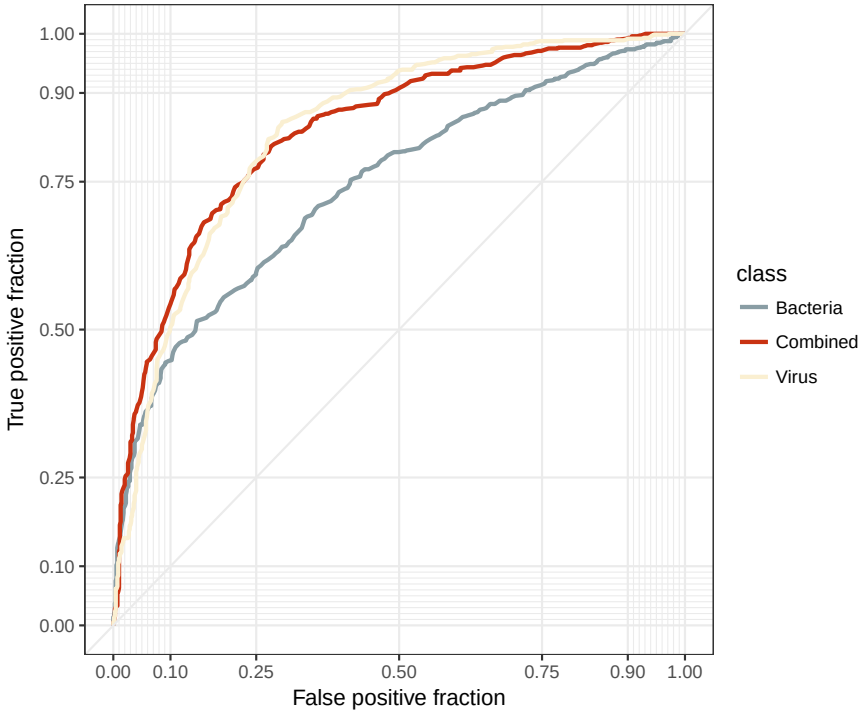

Figure 4: ROC curve from a random forest model built on virus (red) and bacteria (grey) community abundance data.

Try adding more information to the ROC curve:

- Diversity
- Metadata
- Whole Shotgun
- Functional Information
- Virome

Try running a three-way model that includes adenoma.

Show ability to distinguish between samples and negative controls?

Performance of cancer vs ademona vs healthy models between microbial signatures

After evaluating our ability to classify samples as cancerous vs healthy, we incorporated the pre-cancerous adenoma samples into the model and evaluated out ability to classify the groups out of the total dataset **(Figure 5)**.

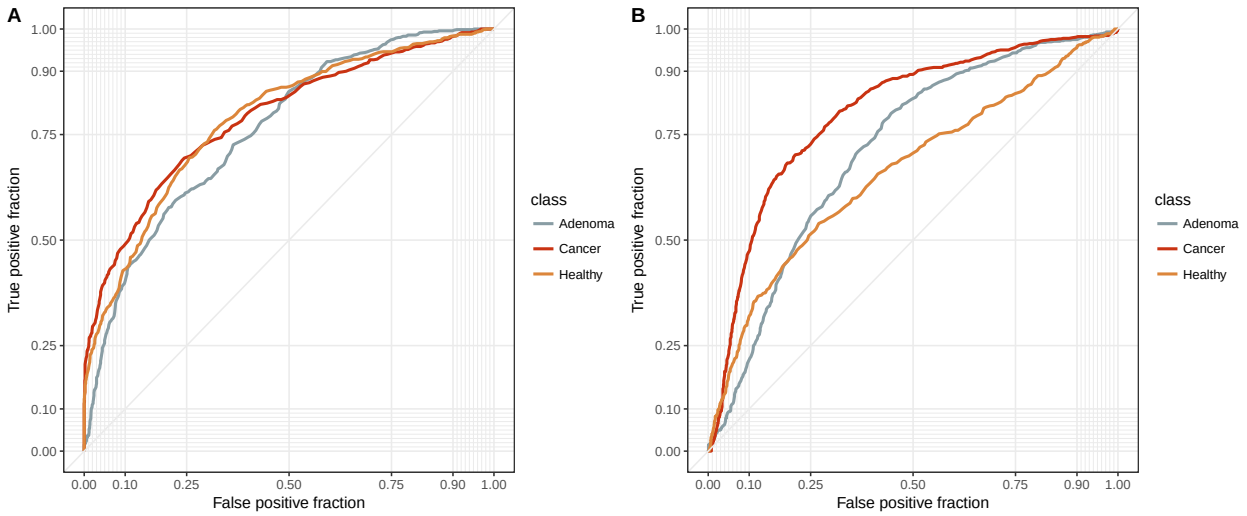

Figure 5: ROC curves from three-class random forest tuned on mean AUC for A) virus and B) bacterial signatures. Each curve represents the ability of the specified class to be classified against the other two classes.

Identifying Major Factors in Colorectal Cancer

Discussion

Conclusions

Methods

Acknowledgements

Conflicts of Interest

Figures

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).