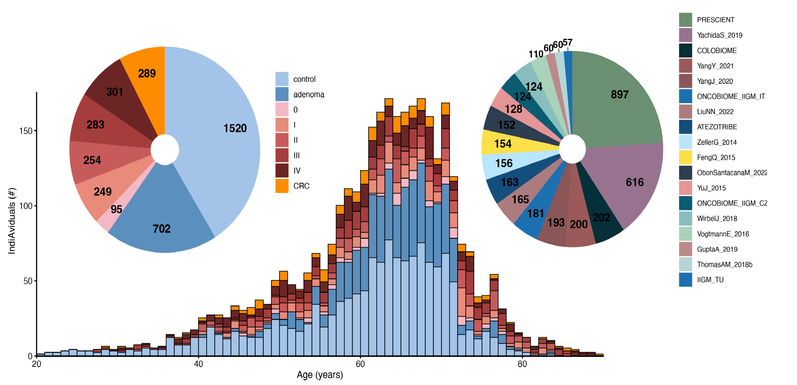
Integrated meta-analysis of the gut virome in colorectal cancer

Chahat Upreti[1,2], Jack T. Sumner[1,2], Jordan Jensen[1,3], Gianmarco Piccinno[4], Ana Nogal[5], Long H. Nguyen[1,6] , Nicola Segata[4], Eric Franzosa[1,2,3], Kelsey N. Thompson[1,2,3], Curtis Huttenhower[1,2,3,7]

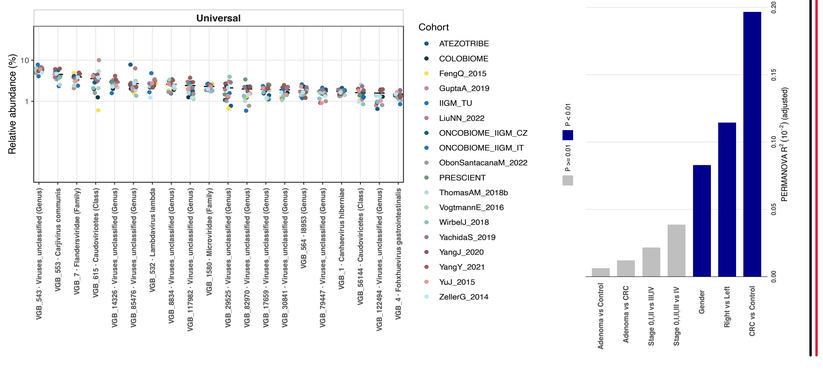
[1]Dept. of Biostatistics, Harvard T.H. Chan School of Public Health, [2]Broad Institute of MIT and Harvard, [3]Harvard Chan Microbiome in Public Health Center, [4]University of Trento, [5]Department of Epidemiology, Harvard T.H. Chan School of Public Health, [6]Massachusetts General Hospital and Harvard Medical School, [7]Dept. of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health

Colorectal cancer (CRC) is the third most common cancer worldwide and has been strongly linked to alterations in the gut microbiome. While bacterial components of the CRC microbiome have been extensively characterized, the human gut virome — comprising diverse viruses including bacteriophages that shape microbial community dynamics and influence host health — remains comparatively underexplored, particularly with respect to interactions with gut bacteria and progression along the adenoma–carcinoma sequence. Here, we systematically profiled the gut virome using BAQLaVA across eighteen publicly available gut metagenomic studies from around the world (n = 3,740), including healthy, adenoma, and CRC samples. We subsequently characterized global virome compositional changes across disease states, identified CRC-associated viral biomarkers, examined virus–host associations, and evaluated the predictive capacity of the gut virome for CRC classification.

Overview of the dataset

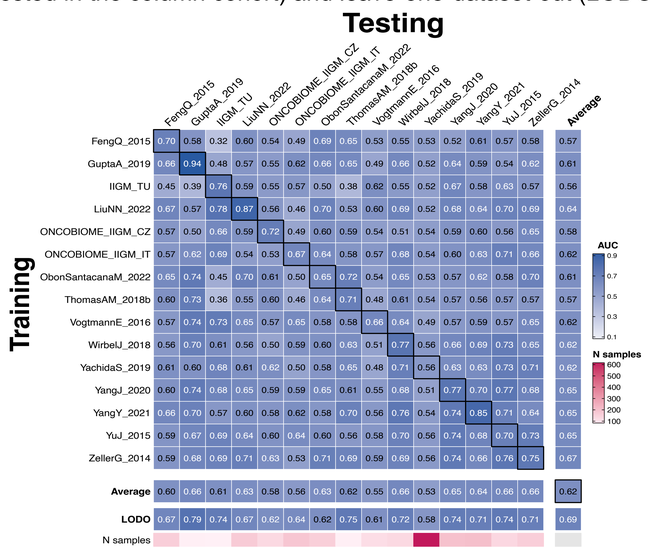


Several VGBs, including known viruses like *Lambdavirus lambda* and the classic crAssphage *Carjivirus communis* were found in all cohorts of our database. PERMANOVA analysis reveals that the distinction between CRC and Control samples exerts significant influence on the dataset variance.



Gut virome can predict CRC status

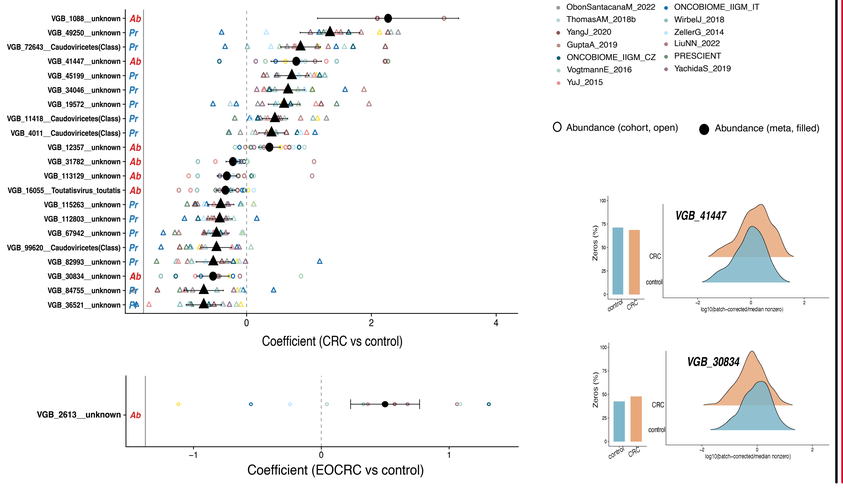
We checked viral-feature random forest performance for CRC versus control using pairwise cross-cohort prediction (each cell: model trained in the row cohort, tested in the column cohort) and leave-one-dataset-out (LODO) testing.



We see heterogeneous transferability: many cohort pairs achieve moderate-to-high AUC when training and testing contexts align, but a subset of off-diagonal cells fall toward chance-level performance, indicating that viral signal for CRC is not uniformly portable across studies and that cohort-specific biology/batch effects likely contribute to the weakest cells in the matrix.

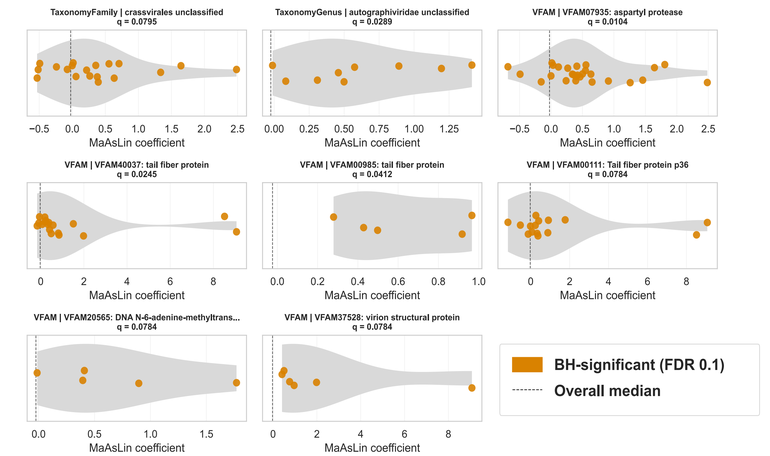
Meta-analytic viral biomarkers of CRC & EOCRC

Differentially abundant viral biomarkers associated with CRC and EOCRC were identified using the MMUPHin meta-analysis framework integrated with MaAsLin3, enabling study-specific modeling followed by pooled meta-analytic effect estimation across cohorts. We found several biomarkers for CRC and EOCRC.



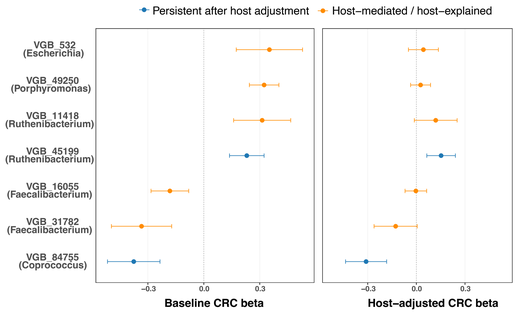
Viral host explains much of viral CRC association

We tested whether viral traits co-vary with CRC associations. Strong hits include tail fiber proteins and a DNA adenine methyltransferase, which are associated with host-sensing and defense, suggesting that CRC-associated changes in the gut virome are enriched for genes involved in recognizing or persisting with bacterial hosts.



Adjusting for host effect explains CRC association for most, but not all VGBs

We linked VGBs to their predicted bacterial hosts and fit mixed models with and without host abundance to test how much CRC–virus signal tracks the host. Host adjustment explained CRC association significance for a majority of VGBs (48/67, including CRC biomarker VGBs), yet several VGBs also showed CRC associations that persisted after adjustment.



Acknowledgments

This research is supported by the PROSPECT grant, funded by the Cancer Grand Challenges program (UK/US). The authors also gratefully acknowledge the use of the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

Discover bioBakery software and tutorials via:

http://huttenhower.sph.harvard.edu/biobakery

@hutlab