Perturbed Gut Viral Ecology in Inflammatory Bowel Disease: a Multi-cohort Study

Jiaxian Shen[1,2,3], Etienne Nzabarushimana[1,2,3], Hanseul Kim[1,2,3], Jordan Jensen[5,6], Will Nickols[1,4,5], Daniel R. Sikavi, Evan Sang, Lathrop Chung, Philips Okeagu[8], Nanako Shirai[7], Eric A. Franzosa, Curtis Huttenhower[1,4,5,6], Andrew T. Chan[2,3,5], Kelsey N. Thompson[1,4,5,6], Long H. Nguyen[1,2,3,5]

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[2]Clinical and Translational Epidemiology Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
[3]Division of Gastroenterology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
[4]Broad Institute of MIT and Harvard, Cambridge, MA, USA
[5]The Harvard Chan Microbiome in Public Health Center, T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
[6]Department of Immunology and Infectious Diseases, T.H. Chan School of Public Health, Harvard University, Boston, MA, USA
[7]Harvard Medical School, Boston, MA, USA
[8]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

Background: The burden of inflammatory bowel disease (IBD) is rising globally. While prior studies have uncovered the critical role of gut dysbiosis in IBD and its subtypes, Crohn's disease (CD) and ulcerative colitis (UC), most have focused on its bacterial determinants and few have explored gut viral ecology.

Methods: Leveraging the recent development of a next-generation method for viral profiling, we uniformly processed and harmonized 2,574 IBD shotgun metagenomes from 580 individuals (320 CD, 124 UC, 136 non-IBD controls; **Fig. 1A**) enrolled in eight independent international cohorts from the Human Microbiome Bioactives Resource.

Results: In total, we detected 5,391 unique viral genome bins (VGBs, akin to bacterial species-level clades). The majority of these VGBs represented uncharacterized viral "dark matter," of which 78% were completely unclassified across all taxonomic levels. This represents a relative overrepresentation of novel viruses in the IBD gut when compared to the 7% unclassified background of our reference database. The gut virome was 1.5x more associated with disease status (i.e., IBD vs. control) than gut bacteria (R^2=2.1%, p=0.001, **Fig. 1B**). Aligning viral and bacterial profiles, we found that they were significantly coupled with similar clustering patterns (correlation=0.89, p=0.001). Interestingly, this coupling may be disrupted in CD and UC vs. non-IBD (p_{FDR}≤0.01, **Fig. 1C**). Additionally, viral and bacterial richness and evenness decreased monotonically from non-IBD to UC to CD (**Fig. 1D**). Using generalized multivariable linear models, we identified 343 differentially abundant VGBs (15.6% of those detected) in IBD compared to non-IBD (69% greater than the count for differentially abundant bacteria, **Fig. 1E**). As expected, most of these significantly altered viruses were uncharacterized (75%), indicating that these novel viruses may be a critical yet underexplored factor in IBD. Among characterized VGBs, *Vimunumvirus ST147VIM1phi71, Peduovirus P2*, *Felixounavirus felixO1, Lambdavirus lambda*, and *Punavirus P1*—several of which are putative phage of the IBD-associated bacteria, *Escherichia coli*—were significantly enriched in IBD (**Fig. 2A-B**). *Felixounavirus felixO1* and *Fohxhuevirus gastrointestinalis* were differentially abundant in CD vs. UC (**Fig. 2C**). Finally, using a machine learner, we found comparable accuracy in classifying IBD vs. non-IBD when using viral features compared to bacteria (AUC>0.95).

Conclusion: Our study represents the largest virally-targeted investigation of the IBD gut to-date. Next, we will expand our study to include comparatively understudied RNA viruses and will interrogate the interplay between microbial domains.

Keywords: inflammatory bowel disease, IBD, virome, gut microbiome, multi-cohort

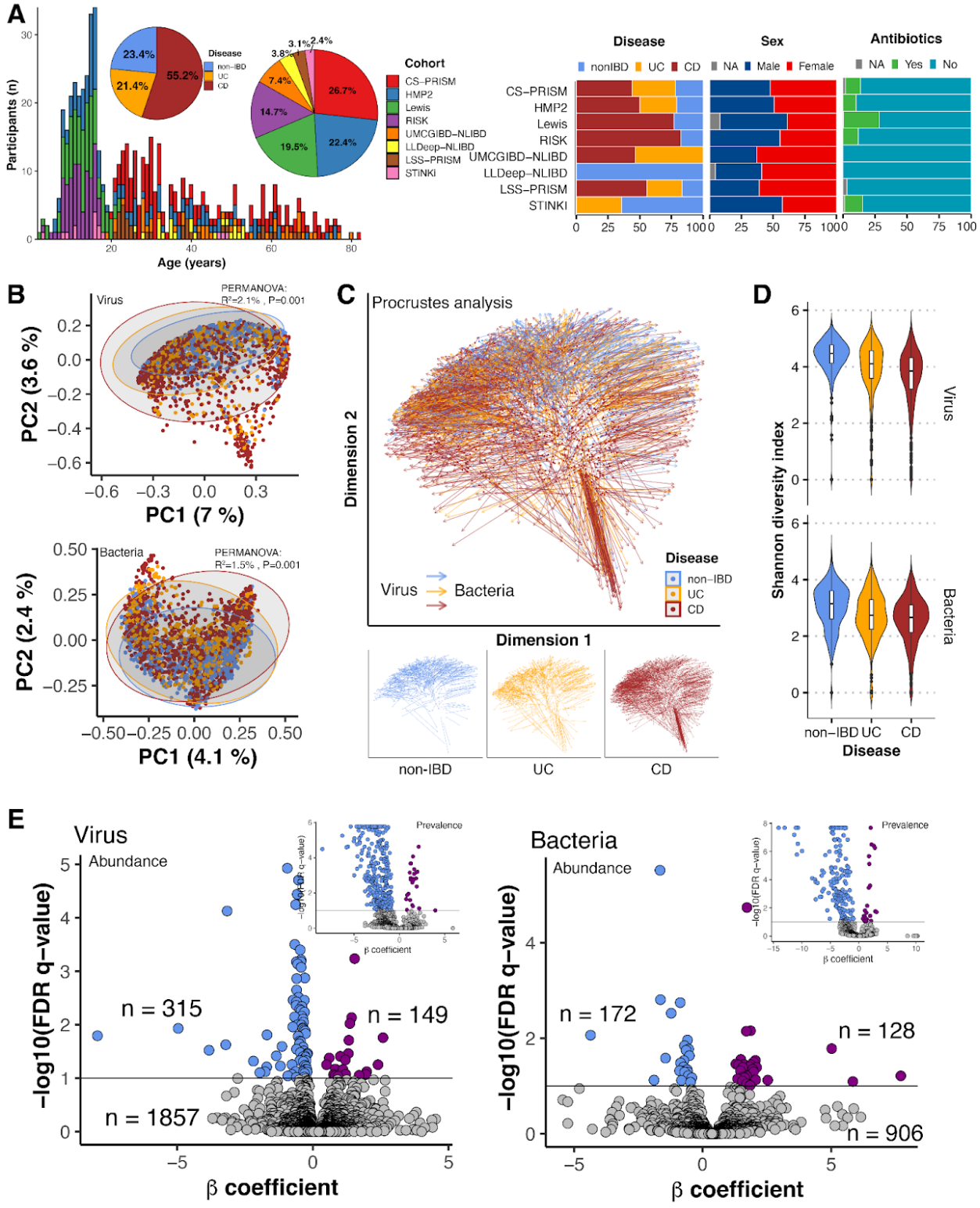

Fig. 1. Perturbed gut viral ecology in the IBD microbiome. (**A**) Baseline characteristics of 580 participants from eight cohorts by age, cohort, and disease phenotype as well as distribution of participant disease, sex, and prior antibiotic usage by each cohort. (**B**) Viral and bacterial beta diversity between non-IBD, UC, and CD. (**C**) Procrustes analysis showed that virome and bacteriome were significantly coupled and exhibited similar clustering patterns (Procrustes correlation = 0.89, p=0.001). While virome and bacteriome are globally correlated, compared with

non-IBD, the discrepancy was significantly larger in CD, followed by UC, as evidenced by the longer lines connecting multi-kingdom profiles (p_{FDR}≤0.01). (**D**) Viral and bacterial richness and evenness (indicated by Shannon Index) were highest in non-IBD controls, followed by UC, then CD. (**E**) Differentially abundant viruses and bacteria in IBD vs. non-IBD were identified using generalized multivariable linear models. Each point represents an individual viral genome bin (VGB) for viruses or species genome bin (SGB) for bacteria. Age, sex, antibiotic usage, and sequencing depth were adjusted for, and a per-participant random intercept was included in the models. Significance was determined as p_{FDR}<0.1 (indicated by a horizontal line). Multi-test *p*-values were adjusted using the Benjamini-Hochberg method.

[Short version]

Fig. 1. Perturbed gut viral ecology in IBD microbiome. (**A**) Baseline characteristics of participants by age, cohort, disease phenotype and distribution of participant disease, sex, and antibiotic usage per cohort. (**B**) Viral and bacterial beta diversity in non-IBD, UC, CD. (**C**) Procrustes analysis showed that viral and bacterial profiles were significantly coupled (correlation=0.89, *p*=0.001). Compared with non-IBD, the discrepancy was significantly larger in CD, followed by UC, as evidenced by the longer lines connecting multi-kingdom profiles (p_{FDR}≤0.01). (**D**) Viral and bacterial richness and evenness were highest in non-IBD, followed by UC, then CD. (**E**) Differentially abundant viruses and bacteria (p_{FDR}<0.1, multi-test adjusted by BH) in IBD vs. non-IBD were identified using generalized multivariable linear models. Each point was one VGB. Age, sex, antibiotic usage, and sequence depth were adjusted for, and a per-participant random intercept was included.

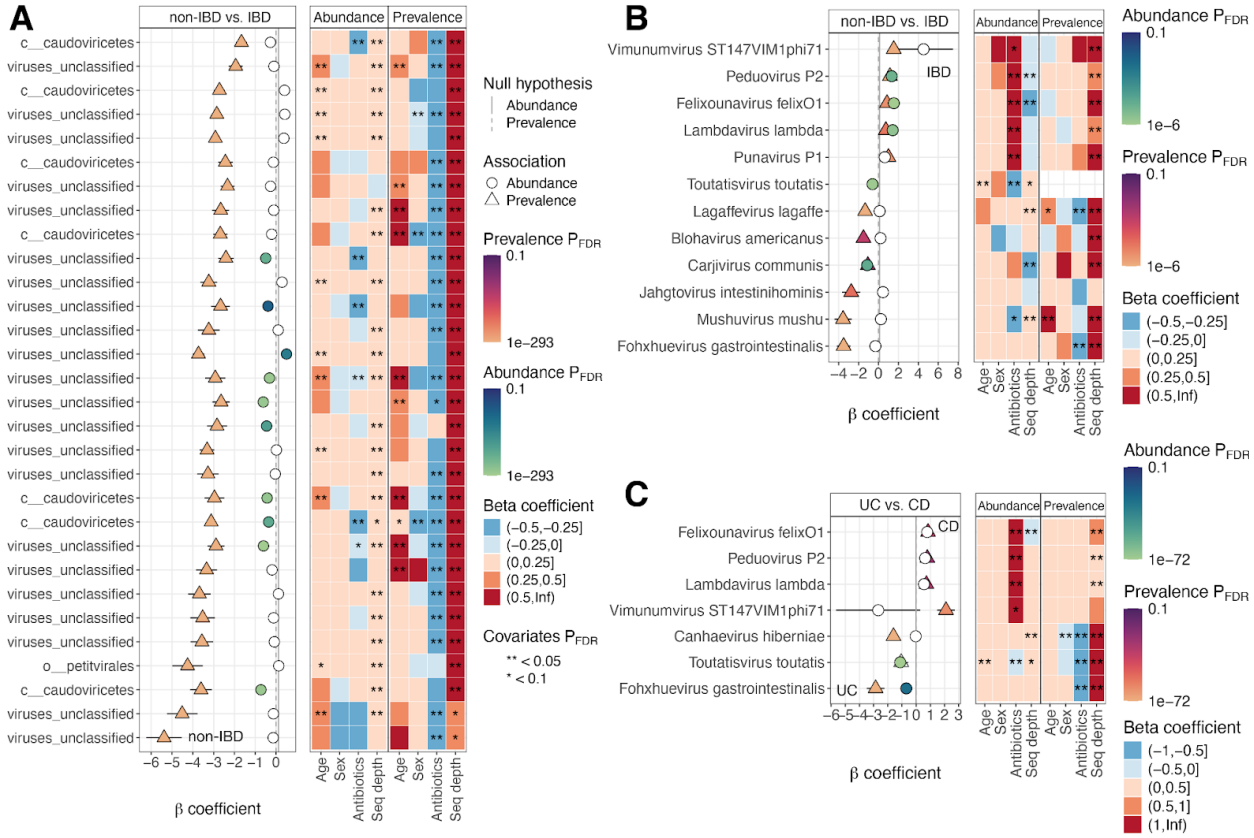

Fig. 2 Significantly altered viruses in IBD at the VGB level. (**A**) Top 30 significant viruses associated with IBD vs. non-IBD as ranked by p_{FDR}. VGBs that were completely uncharacterized across all taxonomic levels were labeled as "viruses_unclassified"; those with certain-level putative taxonomy were labeled by their highest-resolved taxonomic level according to the International Committee on Taxonomy of Viruses (ICTV; from realm, kingdom, phylum, class, order, family, genus, to species). ICTV-resolved viral species significantly associated with (**B**) IBD vs. non-IBD and (**C**) CD vs. UC. We adjusted for age, sex, antibiotics usage, and sequencing depth, and a per-participant random intercept was included in the models. Significance was determined as p_{FDR}<0.1. Multi-test *p*-values were adjusted

using the Benjamini-Hochberg method. Bars on the points represented the standard errors of the estimated beta coefficients.