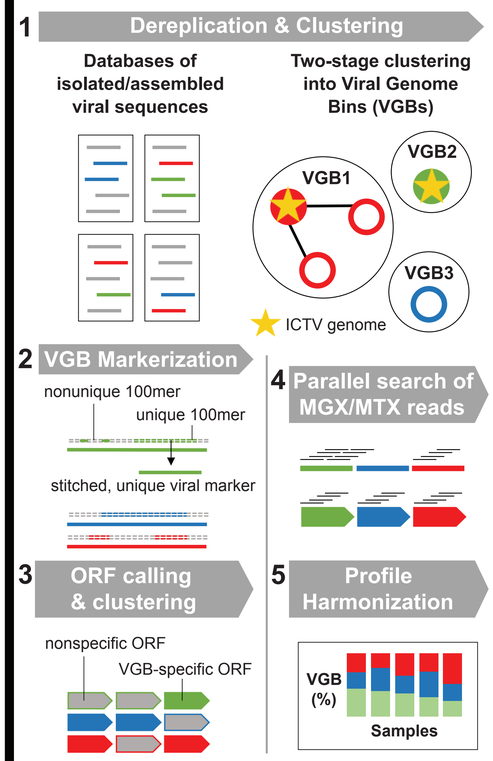
Dramatically improved viral profiling from metagenomes and metatranscriptomes using marker sequence identification

Jordan Jensen[1], Sagun Maharjan[1], Moreno Zolfo[2], Philipp C. Münch[3], Will Nickols[1], Bailey Bowcutt[1], Qianwen Li[1], Jennifer Lin[1], Nicola Segata[2], Eric A. Franzosa[1], Curtis Huttenhower[1]

[1]Harvard Chan Microbiome in Public Health Center, Harvard University, Boston, MA, USA; [2]Centre for Integrative Biology, University of Trento, Italy; [3]Department for Computational Biology of Infection Research, Helmholtz Center for Infection Research, Braunschweig, Germany.

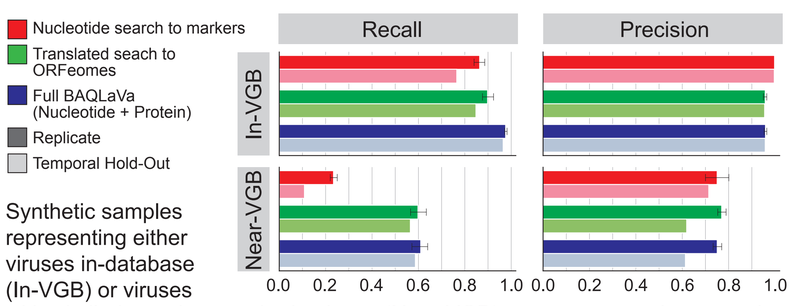
BAQLaVa viral profiling of metagenomes & metatranscriptomes

BAQLaVa (**B**ioinformatic **A**pplication for **Q**uantification and **La**beling of **V**iral t**a**xonomy), integrates reference-based profiling with nucleotide and protein databases to generate viral taxonomic profiles from shotgun DNA or RNA sequencing.



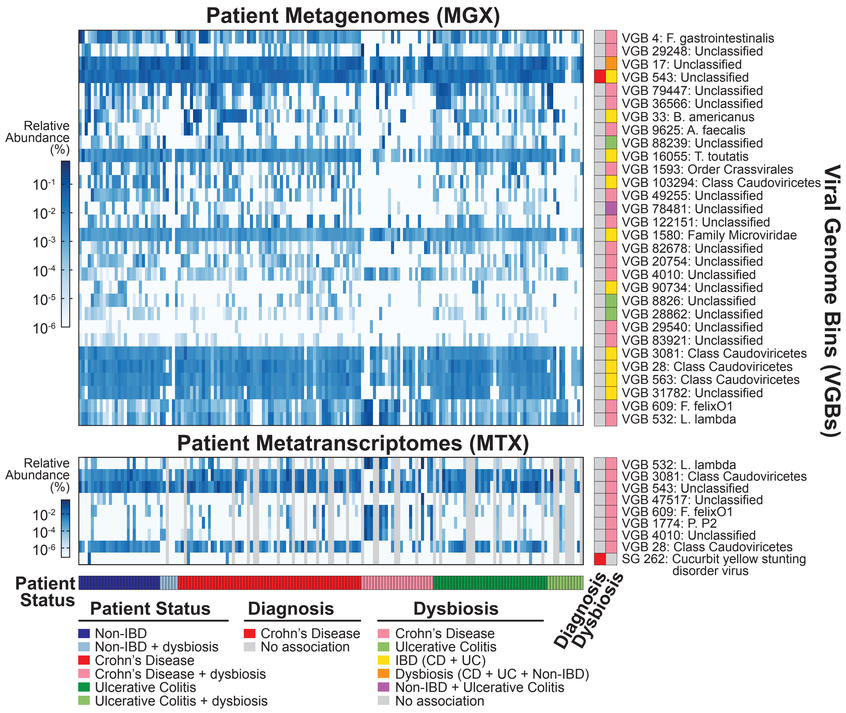
(1) BAQLaVa database generation: 122,099 Viral Genome Bins (VGBs) were generated by integrating well-characterized viral databases (ICTV and RefSeq) and novel viral metagenome assembled genome (vMAG) databases. Two-stage clustering included initial dereplication & clustering followed by a graph network clustering step to account for viral recombination. From each resulting VGB, we identified
(2) windows of unique marker nucleotide regions and
(3) panproteome contents to which translated amino acid matches can be made.
(4) In parallel, metagenomic (MGX) or metatranscriptomic (MTX) reads are searched against these databases to
(5) produce a harmonized VGB profile.

Nucleotide & translated search each contribute unique functionality to BAQLaVa



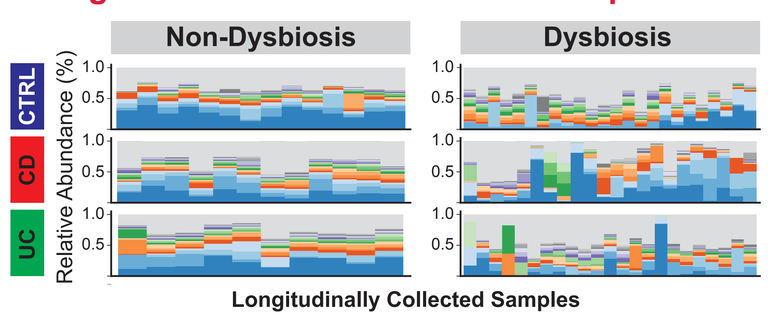
Synthetic samples representing either viruses in-database (In-VGB) or viruses with similarity to but not in-database (Near-VGB) were generated to examine whether BAQLaVa can predict a nearest neighbor when encountering novel genomes. Nucleotide search alone exhibits the highest precision, useful for targeting profiling to In-VGB viruses only. Nucleotide and translated search combined exhibit high recall for Near-VGB viruses, indicating its use in capturing a wider viral profile in the absence of a fully characterized viral landscape.

Loss of viral diversity in IBD-associated dysbiosis



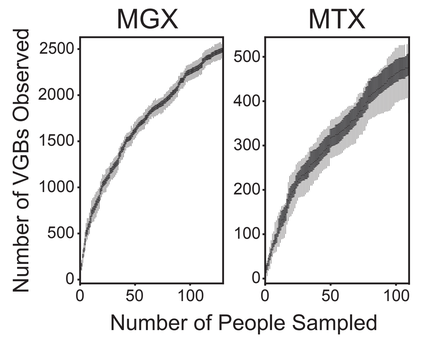
BAQLaVa was used to profile 1,626 MGX and 816 paired MTX from 130 individuals with inflammatory bowel disease (IBD) or non-IBD controls. Using MaAsLin 3, we identified significant changes to the virome that occur in both active and inactive IBD, including a strong loss of diversity in dysbiosis. We also observe that IBD is associated with an overrepresentation of novel VGB clusters containing no previously characterized viruses.

Individuals who exhibit microbial dysbiosis show higher baseline variation in virome profiles



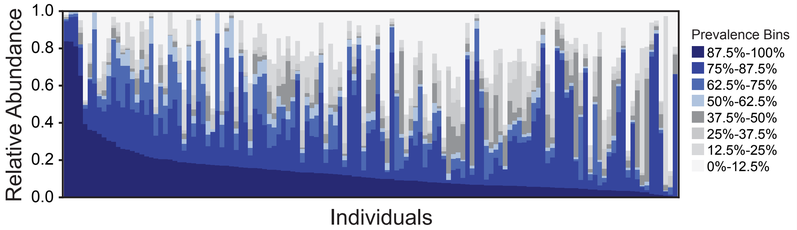
Example longitudinal virome profiles for individuals within subsets of disease and dysbiosis reveals that individuals who exhibit any days of dysbiosis tend to display more variation in weekly virome profiles than individuals who do not experience any dysbiotic days, regardless of disease status.

Viral properties observed across 130 longitudinally sampled individuals



Saturation curves of the 130 individuals across VGBs observed have not yet hit saturation points. Consistent with reports of viral saturation not having been reached at the species level, this underscores the importance of viral profiling methods that can identify closest neighbors when encountering novel viruses, as with BAQLaVa translated search to ORFeomes.

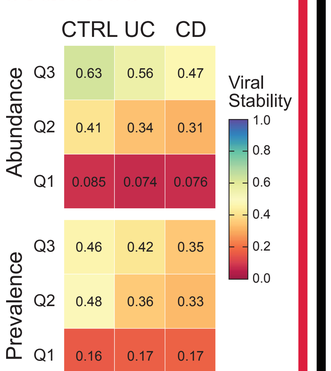
Viromes exhibit bimodal VGB representation by prevalence



Relative abundance plots of viromes shaded by VGB prevalence reveals that viruses fall largely into one of two groups: highly prevalent viruses (those found in ~75% of individuals or more), or individual-specific viruses (found in ~37.5% of individuals or less). Viromes are rarely dominated by VGBs of moderate population prevalence.

VGB stability is associated with both abundance and prevalence

Analysis of consecutive longitudinal samples revealed greater within-subject stability of viruses that were locally abundant or globally prevalent in the population. Additionally, we show that IBD reduces stability when compared to healthy controls.



Acknowledgements

We thank research support by Astellas Pharma Inc. and the Gates Foundation. We especially appreciate the participants in the HMP2 cohort who made this study possible.

https://huttenhower.sph.harvard.edu/
https://huttenhower.sph.harvard.edu/baqlava/