RENDERING THE UNSEEN VIROME VISIBLE

by

Michael Joseph Tisza

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

July 2020

Abstract

Thousands of distinct viruses have been discovered and have had their genomes sequenced. Despite, or perhaps because of this, the concept of “viruses” remains fluid. No gene or gene family is conserved between viruses, making them “polyphyletic.” What makes a genetic element a virus? How many kinds of viruses exist? When a new genetic element is discovered, how can one determine if it is a virus? These questions are complicated, and further experimentation to explore the "virus sequence space" will be required. Both genetic interpretation and modeling of important virus genes as well as isolation and analysis of virus particles can lead us through the vast diversity of the virus sequence space, and maybe allow us to "touch the walls" at the extremes of this space. Ultimately, these considerations will aid us in answering more practical questions. For example, how do the multitudes of viruses living in and on us affect our well-being? This dissertation presents original research that pushes the field of virology forward by striking out into the unexplored reaches of the virus sequence space, expanding our knowledge of virus genome sequences, i.e. the virome. Orthogonal techniques are developed and implemented to latch on to and explore distinctive virus-like signals, including protection of virus genomes from nucleases, circular DNA molecules, and three-dimensional structure conservation of capsids and other virion proteins. Additionally, the development and public release of a bioinformatic tool, Cenote-Taker2, addresses the persistent problems of finding familiar and divergent virus sequences of "known types" in complex datasets and accurately annotating these sequences for distribution to the scientific public. This should accelerate research across the field of virology. Finally, in the last chapter, sequencing data from thousands of human metagenomes is interrogated to pull out high-quality sequences from over 80,000 virus taxa, and strong associations are defined between over a thousand of these virus taxa and a variety of human chronic disease states.

Primary Reader and Advisor: Christopher B. Buck

Secondary Reader: Richard Roden

Acknowledgements

I would like to make clear that any success I have had has been more successful and every failure less painful because of the people who support me. My parents, Janet Tisza and John E. Tisza (may he rest in peace), managed to turn an intransigent boy with no respect for authority into a somewhat productive member of society. My brothers John S. Tisza and Daniel Tisza have always kept me on my toes.

My wife, Monica Garcia, and our daughter, Ana Tisza, provide a constant fountain of love and joy that I look forward to returning to each day after working in the lab. Monica, having received her PhD several years ago, has also provided guidance during difficult times of my scientific training. I smile each time I think about how lucky I am to have Monica and Ana in my life.

My interest in pursuing science was sparked in large part by my professor at Augustana College, Stephanie Fuhr. Jeffrey Chang, professor at University of Texas Health Science Center in Houston, was crazy enough to give me a job as a research assistant after college, and he was patient enough to teach me the ropes of scientific experimentation and thinking. Michael Lichten and Orna Cohen-Fix, former directors of the National Institutes of Health - Johns Hopkins University Graduate Partnership Program, were models of creative and thorough scientists as I began my PhD studies.

I had the incredible fortune to join Chris Buck's lab (Tumor Virus Molecular Biology Section) in August of 2015 to conduct research in pursuance of my PhD. Chris and I are both contrarian dreamers, and it still seems implausible that we were able to, by and large, hold our combative personalities at bay to have a happy, respectful, and productive mentor-mentee

relationship. Diana Pastrana is the yin to Chris's yang, and while she had no formal obligation to mentor me, she has taught me an incredible amount about science and what it means to be a mentor. I will never be able to repay Chris, Diana, or any of the other people who have guided me on my journey, but I hope to take what I've learned so far and pay it forward to the next generation of scientists.

I also want to acknowledge current and past members of Chris's lab who I've had the pleasure of working with, including Eileen Geoghegan, Alberto Peretti, Nicole Welch, Brittany Stewart, Gabe Starrett, Anna Belford, Mary Piaskowski, and Tristan Neal.

I am grateful to all of my committee members, Alison McBride, Richard Roden, and Jocelyne DiRuggiero. Former committee member James Taylor, who passed away before my defense, was also a big scientific influence on me.

Finally, I have many colleagues in graduate school that I consider close friends, including Brendan Miller, Diego Rivera Gelsinger, Allison Dennis, Jeff Maltas, and Dennis Burke. While we probably killed the mood at multiple parties by discussing our personal struggles with highly technical experiments that others couldn't relate to, these friends provided a scientific and emotional well of support that I am very grateful for.

Table of Contents

List of Figures

2 Discovery of several thousand highly diverse circular DNA viruses

3 Bibiviruses are a New, Unusual Virus Family Common in the Human Gut

4 Cenote-Taker2 Democratizes Virus Discovery and Sequence Annotation

5 The Human Virome: Over Eighty Thousand Distinct Viruses and Specific Associations with Chronic Diseases

1 Introduction

The concept of “viruses” represents a hodgepodge of biological entities that were mysterious and ill-defined when first discovered in the 1890s and, despite many important advances, remain mysterious and ill-defined today. Members of the non-scientific public might tell you that viruses are tiny germs that can make you sick and can't be killed with antibiotics, and many scientists would struggle to tell you much more than that. One thing that scientists are discovering is that humans have a complex relationship with viruses. On the one hand, a small number of pathogenic viruses have caused incalculable death and destruction throughout human history, and some continue to do so today. On the other hand, each of us is covered in viruses from head to toe, often without any ill effects. In this dissertation, I aim to explore and expand the concept of viruses by studying previously unrecognizable virus genome sequences, i.e. viral dark matter, and investigate the role of viruses in human health outside the traditional scope of acute infection.

Viruses have likely been around since the beginning of life on Earth, and every species of cellular life probably has at least one species of virus that infects it[1]. Humans, for example, are infected by hundreds of distinct virus species[2]. It has been estimated there are more virus particles on Earth than there are stars in the universe[3] (~10^31 particles), and the genetic diversity of viruses dwarfs that of the diversity of all cellular organisms (bacteria, archaea, and eukaryotes) combined[4]. This comparison may not go far enough, however, because all cellular life shares a common ancestor and all cellular genomes share a set of conserved genes. Viruses,

meanwhile, are “polyphyletic.” New viral entities have arisen multiple times over the course of life on this planet and many groups of viruses share no common genes with other groups[1].

The term “virus” is a historically anchored umbrella term for a variety of disparate elements. In the 1890s, bacteria had been discovered and the germ theory of disease was widely accepted. Two scientists, Dmitri Ivanovsky and Martinus Willem Bejierinck, worked independently to demonstrate that an infectious replicating agent responsible for a disease in tobacco plants was distinct from bacteria, which other scientists had previously described as plant disease agents. This new agent's main peculiarity was that it was small enough to pass through a "bacteria-proof" filter. In contrast to previously studied bacteria, it was unable to replicate independent of its plant host. This "filterable infectious agent" was ultimately found to be tobacco mosaic virus, the first virus ever discovered[5].

This functional method of defining viruses has continued to this day and likely is responsible for the grouping of disparate elements into the category “viruses.” A working definition of a virus might be: **A genetic element (i.e. DNA or RNA genome) that replicates within a host cell and is packaged into a metabolically inert, self-encoded proteinaceous capsid shell capable of infecting new hosts**. Therefore, there is no specific genetic sequence requirement or threshold to include or exclude elements as viruses. This is important to consider when thinking about the field of virus discovery.

Since the 1890s, thousands of viruses have been discovered and, starting in 1976[6], sequenced[7]. These sequences have expanded our understanding of what has come to be known as the “virome”, i.e. **the sum total of all virus sequences in a given environment**, affirming that viruses are staggeringly diverse and have no single common ancestor. In the pre-

metagenomics era (the first virus metagenomics paper was published in 2002[8]), the vast majority of sequenced viruses were discovered from one of two sources: 1) multicellular organisms suffering from disease or 2) virus "plaques" of lysed bacteria on bacterial growth plates[9]. New viruses discovered by these methods provide observable virus particles, nucleic acids to sequence, and an associated host phenotype caused by the virus. They have also provided a baseline of reference sequences for the metagenomic era of virus discovery. A disadvantage of these approaches is that they do not capture all types of viruses. Specifically excluded from these assays are: animal and plant viruses that don't cause acute disease, viruses that cause acute disease in unstudied plants and animals, bacterial viruses that only infect difficult/impossible to culture bacteria, and bacterial viruses that are not capable of generating plaques.

The metagenomic era of virus discovery has been based on massively parallel sequencing of environmental and host-associated samples. Often, samples are enriched for virus particles[10], but enrichment is not strictly necessary[11]. From the sequences of these samples, viruses are identified based on their similarity to the sequences of known viruses. This has allowed the expansion of the virome at a remarkable rate. However, with a few exceptions[12], these methods have only expanded our knowledge of viruses of known types, while ignoring sequences with little or no resemblance to previously catalogued sequences, i.e. viral "dark matter"[12-14]. This phenomenon of ignoring unfamiliar sequences has contributed to a type of tunnel vision or streetlight effect in the field.

One of the reasons that viral dark matter is ignored is that it can be difficult or impossible, even from virus-enriched samples, to discriminate between a true virus sequence

and contamination from a bacterial sequence or a sequence from another type of mobile genetic element, such as a plasmid[15].

In this dissertation, Chapters 2 and 3 cover work overcoming the limitations of sequence-similarity-based methods of virus discovery. Specifically, unconventional strategies to detect highly divergent capsid genes or capsid genes of a previously unrecognized type are employed. Because, by definition, virus genomes must encode capsid gene(s)[1,16], the detection of a capsid gene is strong evidence that an unrecognized genetic element is a viral sequence.

Two homologous genes, such as capsid genes, can evolve to the point where nucleotide sequence similarity or even amino acid sequence similarity becomes undetectable. However, protein fold and three-dimensional protein structure are conserved to a greater degree than linear sequence in homologous genes[17]. Predicted folds can be inferred from a sequence and compared to a database of known protein structures in order to detect homology and structural conservation between highly divergent sequences where sequence alignment algorithms fail[17]. This strategy was employed to identify dozens of capsid genes from a set of dark matter sequences.

Capsid genes appear to have arisen *de novo* at least several times from various cellular sources[1,16,18]. It is unlikely that examples of all categories of viral capsids have been cataloged. One strategy to identify previously unrecognized capsid types uses artificial neural networks trained to identify important sequence features of capsid genes across virus families independent of alignment of long stretches of amino acid sequence[12]. In this dissertation, potential capsids identified with these neural networks were validated with wet bench

experiments showing formation of capsids from exogenously expressed potential capsid genes (Chapters 2 & 3), or *in vivo* virion formation in a natural host (Chapter 3).

An impressive body of literature has shown that communities of bacteria that live in and on us, i.e. the bacterial microbiome, play important roles in keeping us healthy[19,20], but can also cause or contribute to chronic diseases and health conditions[21,22]. It is likely that all of our resident bacteria become infected by viruses and that these bacterial viruses have indirect effects on human physiology. For example, bacterial viruses can regulate the abundance of their host bacteria[23], potentially ablating the effects of either "good" or "bad" bacteria. Bacteriophages can also render relatively harmless bacterial species pathogenic by transducing toxin genes[24]. Many of the genetic differences between benign and pathogenic bacterial strains are due to presence or absence of integrated or episomal mobile genetic elements, such as viruses[24,25].

Some studies have tried to unravel the role of the human virome in health and disease, occasionally with promising results[26-28]. However, almost without exception, reviews of the topic lament the fact that no comprehensive virus sequence database exists[13,29]. Experts conservatively estimate that hundreds of millions of virus species exist on Earth[4]. However, GenBank, the largest and most accessed public sequence database, has only 30,000 - 40,000 virus species represented, some without a complete genome sequence available.

To move human virome research forward, work for this dissertation has overcome several obstacles. As previously discussed, detecting highly divergent viruses and discriminating them from non-viral sequences is challenging. Furthermore, annotation of genes and other

features of new virus genomes is particularly challenging. The goal of the work was to overcome these challenges by developing tools to detect familiar as well as highly divergent virus sequences, annotate the genomes, and facilitate GenBank deposition. This tool, Cenote-Taker2, is documented in Chapter 4.

In Chapter 5, Cenote-Taker2 was used to mine thousands of publicly available sequence libraries from human metagenomes (gut, skin, oral, vaginal). The search detected 80,000 unique viruses, the vast majority of which are bacterial viruses. Cenote-Taker2 was applied to a dozen case-control studies that used massively parallel sequencing on stool and/or saliva from patients and control subjects. Significant associations between specific viruses and most of these disease states were found, and these virus associations were often stronger than the associations found for bacteria in these samples.

2 Discovery of several thousand highly diverse circular DNA viruses

Adapted from: eLife publication, 10.7554/eLife.51971

Abstract

Although millions of distinct viral species likely exist, only approximately 9,000 are catalogued in GenBank's RefSeq database. We selectively enriched for the genomes of circular DNA viruses in over 70 animal samples, ranging from nematodes to human tissue specimens. A bioinformatics pipeline, Cenote-Taker, was developed to automatically annotate over 2,500 complete genomes in a GenBank-compliant format. The new genomes belong to dozens of established and emerging viral families. Some appear to be the result of previously undescribed recombination events between ssDNA and ssRNA viruses. In addition, hundreds of circular DNA elements that do not encode any discernable similarities to previously characterized sequences were identified. To characterize these "dark matter" sequences, we used an artificial neural network to identify candidate viral capsid proteins, several of which formed virus-like particles when expressed in culture. These data further the understanding of viral sequence diversity and allow for high throughput documentation of the virosphere.

Keywords

Viral metagenomics, virome, virus discovery, microbial genomics, evolution

Introduction

There has been a rush to utilize the massive parallel sequencing approaches to better understand the complex microbial communities associated with humans and other animals. Although the bacterial populations in these surveys have become increasingly recognizable [30], a substantial fraction of the reads and *de novo* assembled contigs in many metagenomics efforts are binned as genetic "dark matter," with no recognizable similarity to characterized sequences [31,32]. Some of this dark matter undoubtedly consists of viral sequences, which have remained poorly characterized due to their enormous diversity [7,33,34]. Recent efforts have shown that our understanding of viral diversity, even of viruses known to directly infect humans, has been incomplete [35-37]. To increase the power of future studies seeking to more comprehensively catalog the virome and find additional associations between viruses and disease, reference genomes for all clades of the virosphere need be identified, annotated, and made publicly accessible.

Virus discovery has typically proven to be more difficult than discovery of cellular organisms. Whereas all known cellular organisms encode conserved sequences (such as ribosomal RNA genes) that can readily be identified through sequence analysis, viruses, as a whole, do not have any universally conserved sequence components [38-41]. Nevertheless, some success has been achieved in RNA virus discovery by probing for the conserved sequences of their distinctive RNA-dependent RNA polymerase or reverse transcriptase genes in metatranscriptomic data [42]. Also, many bacteriophages of the order *Caudovirales*, such as the families *Siphoviridae*, *Podoviridae*, and *Myoviridae*, have been reported in high numbers due to

their and their hosts' culturability and their detectability using viral plaque assays [43-45]. The relatively abundant representation of these families in databases has allowed new variants to be recognized by high-throughput virus classification tools like VirSorter [46-48]. In contrast, many small DNA viruses are not easily cultured [49], use diverse genome replication strategies, and typically lack DNA polymerase genes such as those in large DNA viruses [50]. An additional challenge is that small DNA viruses with segmented genomes may have segments that do not encode recognizable homologs of known viral genes. Therefore, small DNA viruses are more sparsely represented in reference databases. However, some groups have been successful in discovery of small DNA genomes in a wide range of viromes [10,35,51-55].

Despite the apparent challenges in detecting small DNA viruses, many have physical properties that can be leveraged to facilitate their discovery. In contrast to the nuclear genomes of animals, many DNA virus genomes have circular topology, which allows selective enrichment through rolling circle amplification (RCA) methods [56]. Further, the unique ability of viral capsids to protect nucleic acids from nuclease digestion and to mediate the migration of the viral genome through ultracentrifugation gradients or size exclusion columns allows physical isolation of viral genomes.

The current study grew out of an effort to find papillomaviruses (small circular DNA viruses) in humans and economically important or evolutionarily informative animals [35,57]. The sampling included several types of animals that might serve as laboratory models (e.g., mice, fruit flies, soil nematodes). A number of papillomaviruses were detected among a vastly larger set of circular DNA sequences that were not easily identifiable in standard BLASTN searches. The goal of the present study is to catalog and annotate the circular DNA virome from these

animal tissues to understand the diversity and evolution of viral sequences. We developed a comprehensive bioinformatics pipeline, Cenote-Taker, to classify and annotate over 2,500 candidate viral genomes and generate GenBank-compliant output files. Cenote-Taker is available for free public use with a graphical user interface at http://www.cyverse.org/discovery-environment.

Results

Virion enrichment, genome sequencing, and annotation

We have previously developed methods for discovery of new polyomavirus and papillomavirus species in skin swabs and complex tissue specimens [57]. Nuclease-resistant DNA from purified virions was amplified by random-primed rolling circle amplification (RCA) and subjected to deep-sequencing. Reads were *de novo* assembled into contigs and analyzed with a bioinformatics pipeline, Cenote-Taker (a portmanteau of *cenote*, a naturally occurring circular water pool, and *note-taker*), to identify and annotate *de novo* assembled contigs with terminal direct repeats consistent with circular DNA molecules (Figure 2.1). In this pipeline, putative closed circular sequences of greater than 1000 nucleotides (nt) were queried against GenBank's nucleotide database using BLASTN to remove circles with extensive nucleotide identity (>90% across any 500 nt window) to known sequences. Sequences with >90% identity to previously reported viral sequences represented less than 1.5% of circular contigs and are not included in further analysis. Approximate taxonomy was determined by BLASTX to a protein database derived from RefSeq virus proteins and GenBank plasmid proteins (only hits better than 1×10^{-5} were considered). Open reading frames (ORFs) from remaining unidentified circular DNA

sequences >240 nucleotides (nt) in length were translated and used for RPS-BLAST queries of GenBank's Conserved Domain Database (CDD). ORFs that did not yield E values better than 1×10^{-4} in RPS-BLAST were subjected to BLASTP searches of viral sequences in GenBank's nr database [58-60]. For ORFs that were not confidently identified in BLAST searches, HHBlits [61] was used to search the CDD, Pfam [62], Uniprot [63], Scop [64], and PDB [65] databases. The results were used to annotate and name each sequence in a human-readable genome map as well as a format suitable for submission to GenBank. After checking the Cenote-Taker output of each genome, minor revisions were made, as needed, and files were submitted to GenBank (BioProject Accessions PRJNA393166 and PRJNA396064). All annotations meet or exceed recently proposed standards for uncultivated virus genomes [66]. Plasmid sequences were frequently detected and were discarded. Circular sequences were considered to be plasmid-like if they: 1) had a best BLASTX hit to a plasmid and 2) had no detectable virion structural genes.

Viral enrichment of the analyzed samples (based on ViromeQC [67], with alignment to prokaryotic single-copy housekeeping genes) was typically high (Supplementary File 1). However, even in the samples where enrichment was low, quality viral genomes could still be identified based on the bioinformatic analyses.

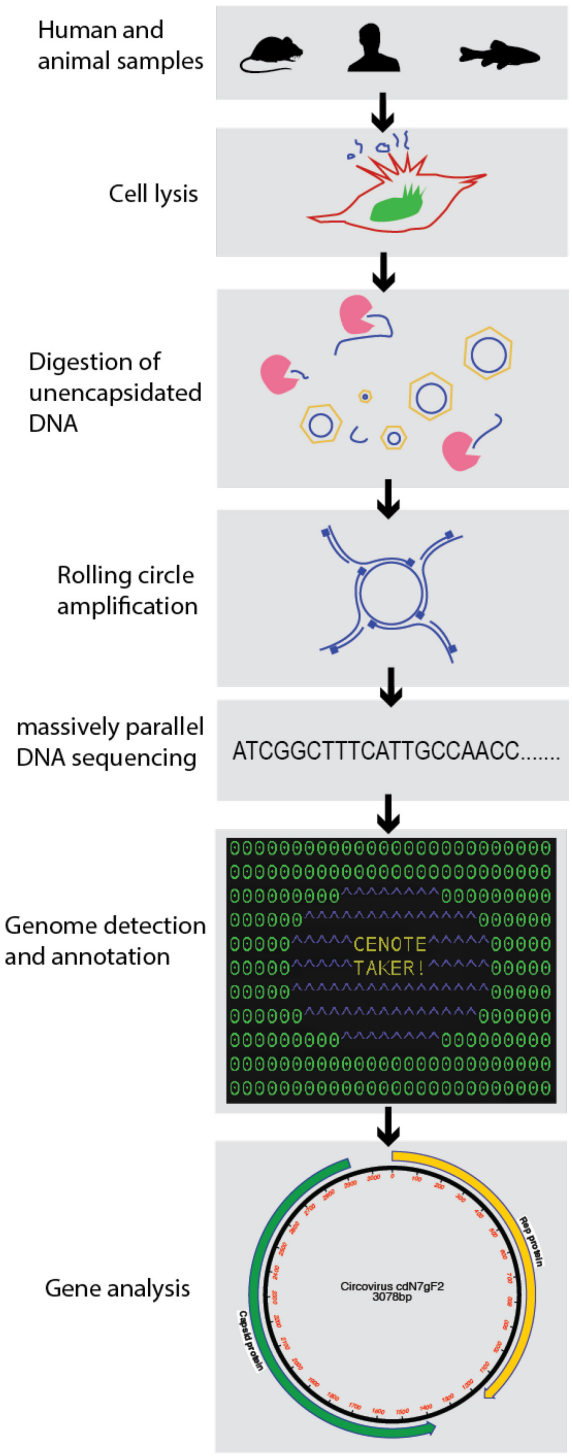


Figure 2.1: Virus Discovery Overview

Pictorial representation of virus discovery methods

Discovery of 2514 DNA viruses in animal metagenomes

Of the novel circular sequences detected in the survey, 1844 encode genes with similarity to proteins of ssDNA viruses and 55 encode genes with similarity to dsDNA viral proteins (Figure 2.2A). The large majority of genomes from this study are highly divergent from RefSeq entries (Figure 2.3). We discovered 868 genomes that had similarity to unclassified eukaryotic viruses known as circular replication associated protein (Rep)-encoding single-stranded DNA (CRESS) viruses. The group is defined by the presence of a characteristic rolling circle endonuclease/superfamily 3 helicase gene (Rep) [18,68], but has not been assigned to families by the ICTV or RefSeq. We estimate that 199 non-redundant unclassified CRESS virus genomes had been previously deposited in GenBank, and 85 are curated in RefSeq (Figure 2.2B). Also abundant was the viral family *Microviridae*, a class of small bacteriophages, with 670 complete genomes. This represents a substantial expansion beyond the 459 non-redundant microvirus genomes previously listed in GenBank (of which 44 were curated in the RefSeq database). Other genomes that were uncovered represent *Anelloviridae* (n=170), *Inoviridae* (n=70), *Genomoviridae* (n=58), *Siphoviridae* (n=18), unclassified phage (n=14), *Podoviridae* (n=10), *Myoviridae* (n=7) unclassified virus (n=6), *Papillomaviridae* (n=4), *Circoviridae* (n=3), unclassified *Caudovirales* (n=3), *Bacilladnaviridae* (n=2), *Smacoviridae* (n=2), and *CrAssphage-like* (n=2) (Figure 2.2B). A table of samples, metadata, and viruses can be accessed at (https://elifesciences.org/articles/51971/figures#supp2). Viral families were found in association with 23 different animal species (Figure 2.2C). It was not surprising to find bacterial viruses as all animals are presumed to have microbial communities, and our sampling often target tissues where these communities reside.

It is difficult to assign a host to most of the viruses from this study due to their divergence from known viral sequences. However, we searched the CRISPR database at (https://crispr.i2bc.paris-saclay.fr/crispr/BLAST/CRISPRsBlast.php), and three viruses had exact matches to CRISPR spacers in bacterial genomes (Siphoviridae sp. ctcj11:Shewanella sp. W3-18-1, Inoviridae sp. ctce6:Shewanella baltica OS195, Microviridae sp. ctbe523:Paludibacter propionicigenes WB4) and one virus had an exact match to the CRISPR spacer of an archaeon (Caudovirales sp. cthg227:Methanobrevibacter sp. AbM4), implying that these organisms are infected by these viruses. Further, the 142 anelloviruses found in human blood samples are almost certain to be bona fide human viruses based on their relatedness to known human anelloviruses.

In addition to circular genomes with recognizable similarity to known viruses, 609 circular contigs appeared to represent elements that lacked discernable similarity to known viruses (Figure 2.2A, C).

The vast majority of the *de novo* assembled circular genomes were <10 kb in length (Figure 2.3). This is largely due to the fact that large genomes are typically more difficult to *de novo* assemble from short reads. Despite these technical obstacles, our detection of a new tailed bacteriophage with a 419kb genome (Myoviridae sp. isolate ctbc_4, GenBank Accession: MH622943), along with 45 other >10 kb circular sequences (Figure 2.3), indicates that the methods used for the current work can detect large viral genomes.

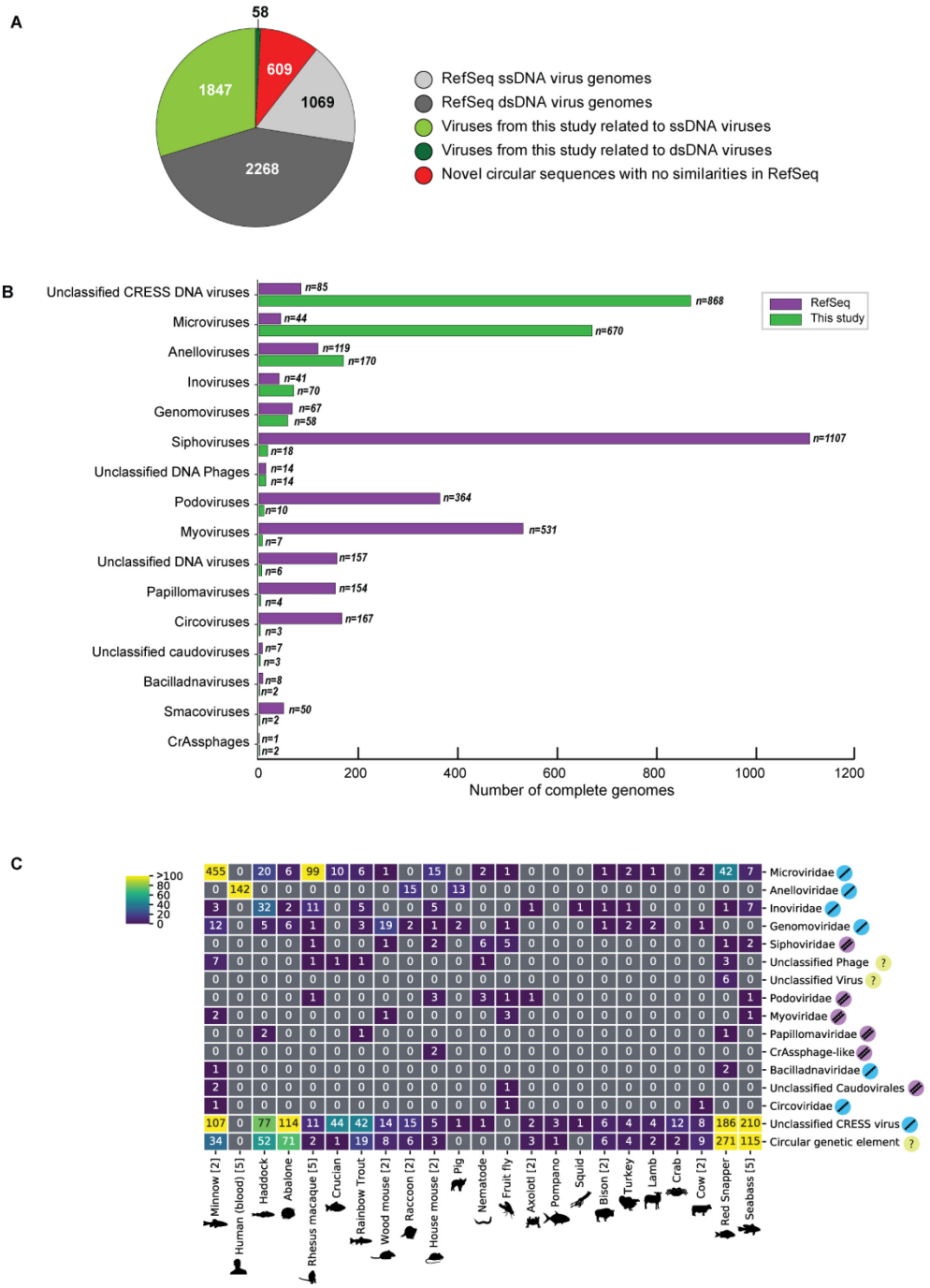


Figure 2.2: Novel viruses associated with animal samples

Gross characterization of viruses discovered in this project compared to NCBI RefSeq virus database entries. (A) Pie chart representing the number of viral genomes in broad categories. (B) Bar graph showing the number of new representatives of known viral families or unclassified groups. (C) Heatmap reporting number of genomes found associated with each animal species. Number of samples per species in brackets. Y axis = virus family/bin. Blue icon = ssDNA, purple icon = dsDNA, yellow icon = unknown strandedness.

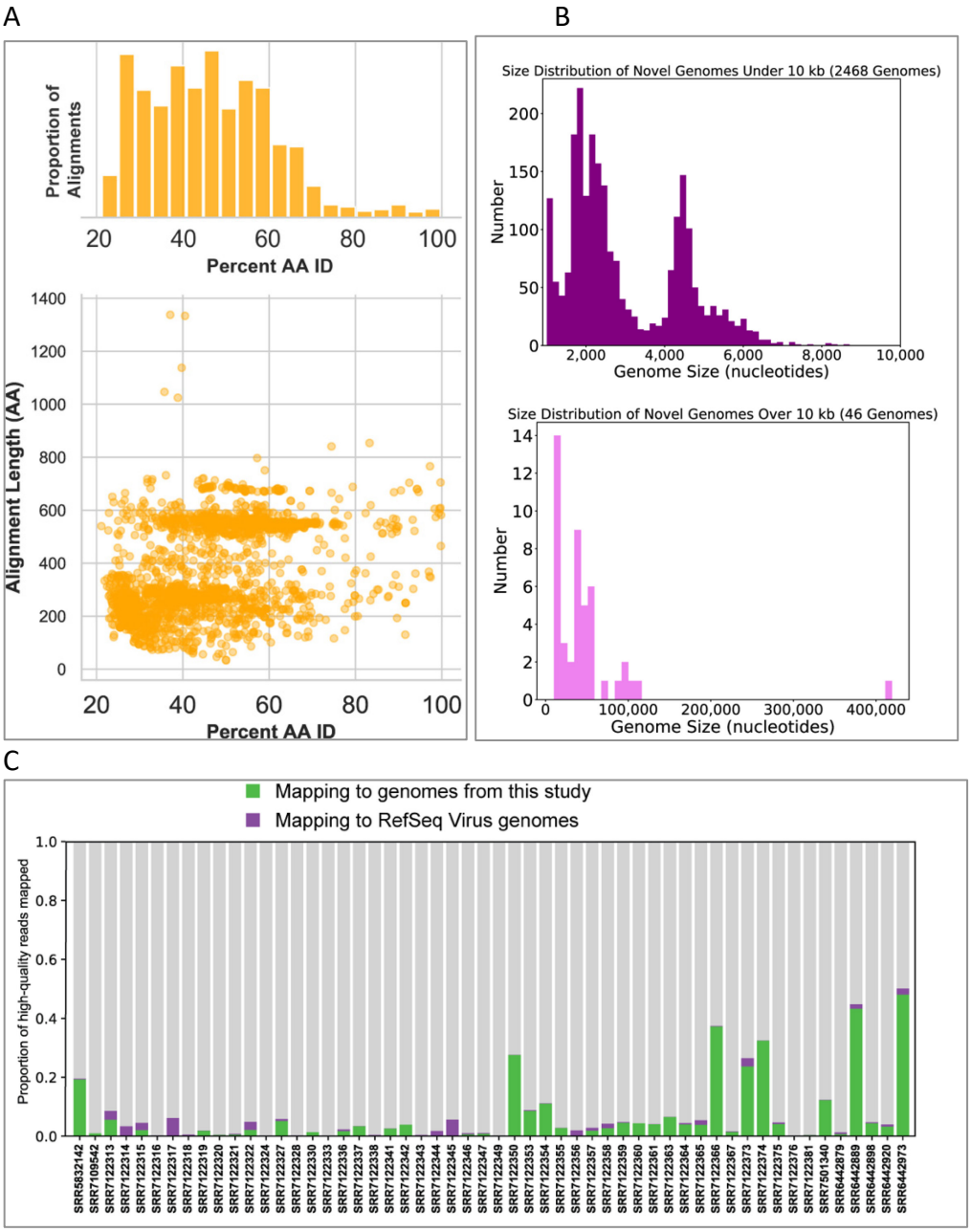


Figure 2.3 Characterization of discovered sequences

(A) BLASTX summary of each circular DNA molecule recovered from virus enriched samples. Sequences were queried against a database of viral and plasmid sequences. Only hits with E values < 10^{-5} were plotted. Here, BLASTX only reports the most significant stretch of amino acid sequence from each circular contig, and, therefore, other regions of each contig can be assumed to be equally or less conserved. (B) Size distribution of circular DNA sequences from this study. (C) Mapping reads to complete viral genome references. Quality-trimmed reads were aligned with Bowtie2 to reference genomes from RefSeq and this study. Genomes were masked for low-complexity regions.

There has been a recent renewal of interest in the hypothesis that viruses may be etiologically associated with degenerative brain diseases, such as Alzheimer's disease [69,70]. Conflicting literature suggests the possible presence of papillomaviruses in human brain tissue [71,72]. Samples of brain tissue from individuals who died of Alzheimer's disease (n=6) and other forms of dementia (n=6) were subjected to virion enrichment and deep sequencing. Although complete or partial genomes of known papillomaviruses, Merkel cell polyomavirus, and/or anelloviruses were observed in some samples, no novel complete viral genomes were recovered. No viral sequences were detected in a follow-up RNA deep sequencing analysis of the brain samples. It is difficult to know how to interpret these negative data. It is conceivable that the known viral DNA sequences observed in the Optiprep-RCA samples represent virions from blood vessels or environmental sources.

It has recently become apparent that certain nucleic acid extraction reagents are contaminated with viral nucleic acids [73]. To ensure we were not merely reporting the sequences of the "reagent virome," we performed our wet bench and bioinformatic pipeline on three independent replicates of reagent-only samples. We found no evidence of sequences of any viruses reported here or elsewhere. Further, cross-sample comparison of contigs showed that almost no sequences were found in different animal samples, aside from technical replicates. In total, six viral genomes were observed in multiple unrelated samples from at least two sequencing runs. It is unclear whether this small minority of genomes (0.24% of the genomes reported in the current study) represent reagent contamination, lab contamination, or actual presence of the sequences in different types of samples.

Given the stringent requirements for sequences to be considered as belonging to a complete viral genome, as well as the largely unexplored nucleotide space of the virome, it is unsurprising that, in most samples, most reads did not align to the genomes reported in this study or virus genomes from RefSeq (Figure 2.3C).

Assignment of hallmark genes to networks shows expansion of virus sequence space

Single stranded DNA viruses, in general, have vital genes encoding proteins that mediate genome replication, provide virion structure, and, in some cases, facilitate packaging of viral nucleic acid into the virion. Being structurally conserved, these genes also tend to be important for evolutionary comparisons and can serve as important "hallmark genes" for virus discovery and characterization. However, even structurally conserved proteins sometimes do not have enough sequence conservation as to be amenable to high confidence BLASTP searches. We therefore set out to catalog hallmark ssDNA virus genes based using protein structural prediction. Structures of hallmark genes of exemplar isolates from most established ssDNA virus families have been solved and deposited in publicly available databases such as PDB (Protein Data Bank) [65]. Using bioinformatic tools, such as HHpred, one can assign structural matches for a given gene based on the predicted potential folds of a given amino acid sequence. HHpred has been extensively tested and validated for computational structural modelling by the structural biology community [74,75]. The method proves especially useful for protein sequences from highly divergent viral genomes that have little similarity to annotated sequences in current databases.

We extracted protein sequences from our dataset and compiled nonredundant proteins from circular ssDNA viruses in GenBank and used them as queries in HHpred searches against the PDB, PFam, and CDD databases. We then grouped structurally identifiable sequences into hallmark gene categories and aligned them pairwise (each sequence was compared to all other sequences) using EFI-EST [76]. The resulting sequence similarity networks (SSNs) were visualized with Cytoscape [77], with each node representing an predicted protein sequence (Figures 2.4, 2.5, 2.7). Nodes (sequences) with significant amino acid similarity are connected with lines representing BLAST similarity scores better than a threshold E value. Sequence similarity network analyses, it has been proposed [78], represent relationships between viral sequences better than phylogenetic trees. Further, SSNs have previously been used for viral protein and genome cluster comparison [18,79-81], and can be used to display related groups of viral genes in two dimensions [82]. These clusters were also used to guide the construction of meaningful phylogenetic trees (Figure 2.4A-B, 2.6).

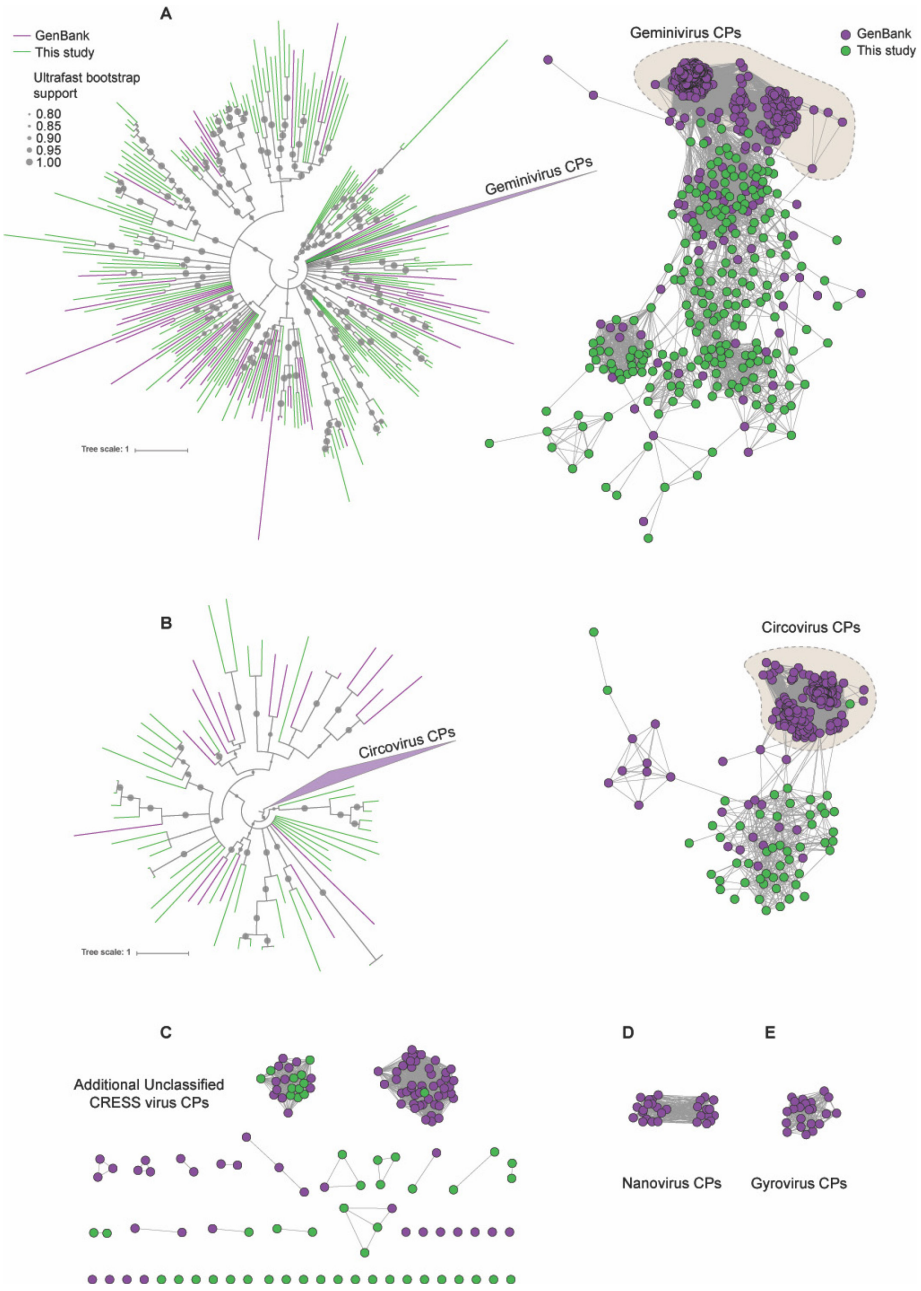


Figure 2.4: Sequence similarity network analysis of CRESS virus capsid proteins

EFI-EST was used to conduct pairwise alignments of amino acid sequences from this study and GenBank with predicted structural similarity to CRESS virus capsid proteins. The E value cutoff for the analysis was 10^{-5}. (A) Cluster consisting of proteins with predicted structural similarity to geminivirus-like capsids and/or STNV-like capsids. The phylogenetic tree was made from all sequences in this cluster. (B) A cluster consisting of sequences with predicted structural similarity to Circovirus capsid proteins. The phylogenetic tree was made from all sequences in this cluster. (C) Assorted clusters and singletons from unclassified CRESS virus proteins that were modelled to be capsids. (D) Nanovirus capsids. (E) Gyrovirus capsids.

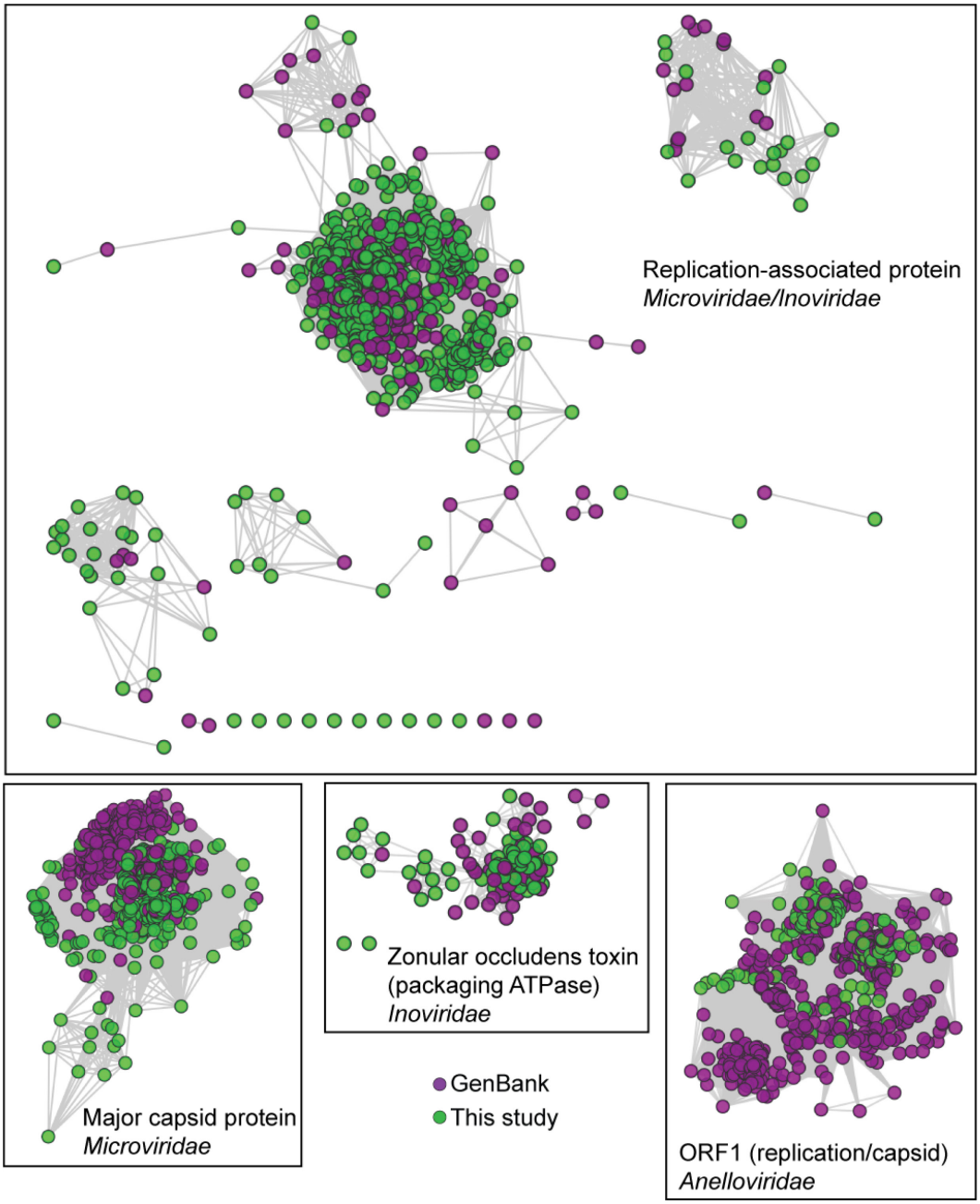


Figure 2.5: Network Analysis of additional viral hallmark genes

Depiction of additional viral hallmark genes from this study and GenBank as sequence similarity networks. E value cutoff = 10^{-5}. See Figure 2.3 and Methods.

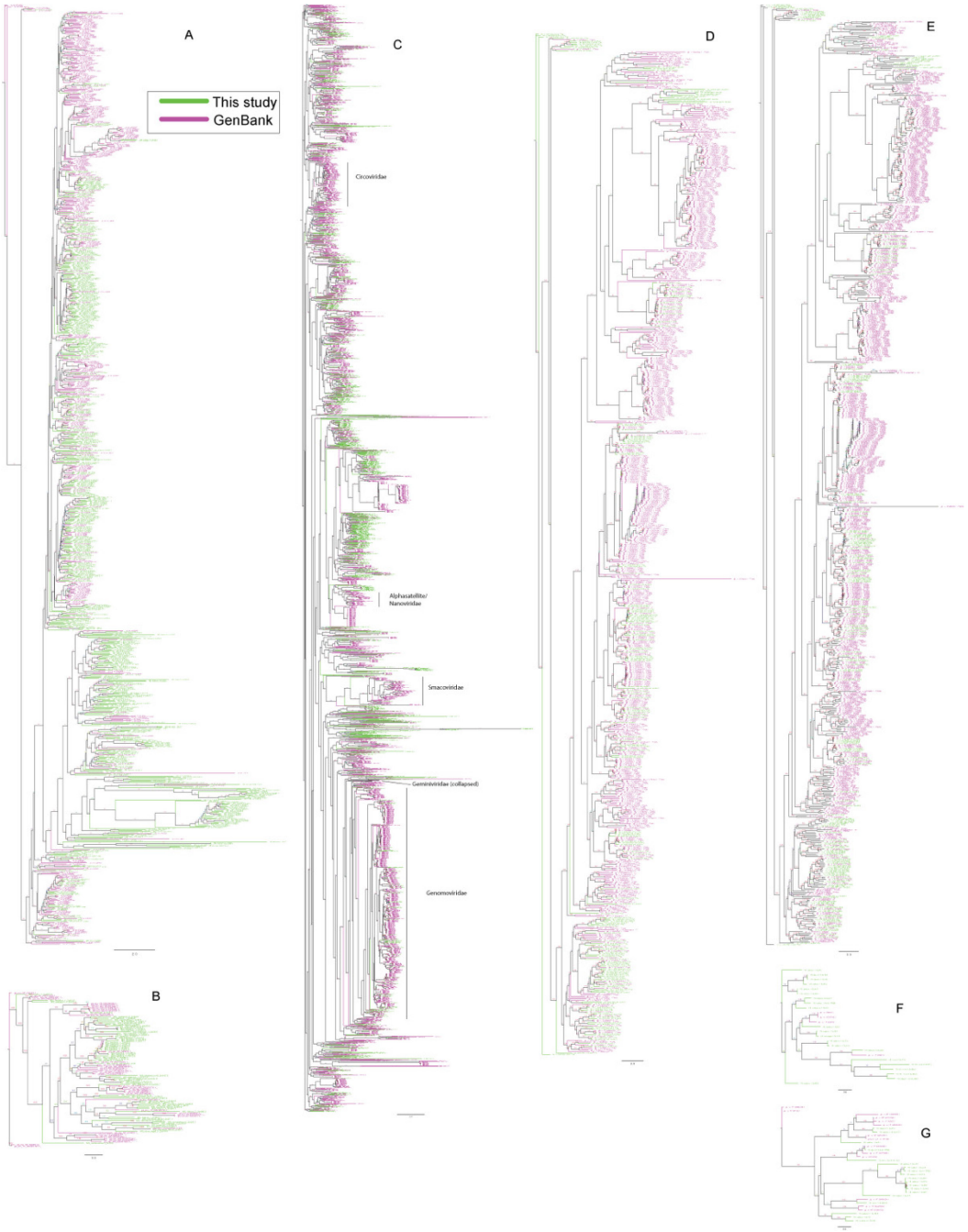


Figure 2.6: Phylogenetic trees of viral hallmark genes

Sequences were aligned with PROMALS3D using structure guidance when possible. Trees were drawn using IQ-Tree with automatic determination of substitution model. See methods. Branches are labeled with bootstrap percent support after 1000 ultrafast bootstrapping events.

(A) *Microviridae* major capsid protein. (B) *Inoviridae* zonular occludens toxin. (C) CRESS virus Rep. (D) *Anelloviridae* ORF1 (E) *Microviridae*/*Inoviridae* Replication-associated protein I. (F) *Microviridae*/*Inoviridae* Replication-associated protein II. (G) *Microviridae*/*Inoviridae* Replication-associated protein III.

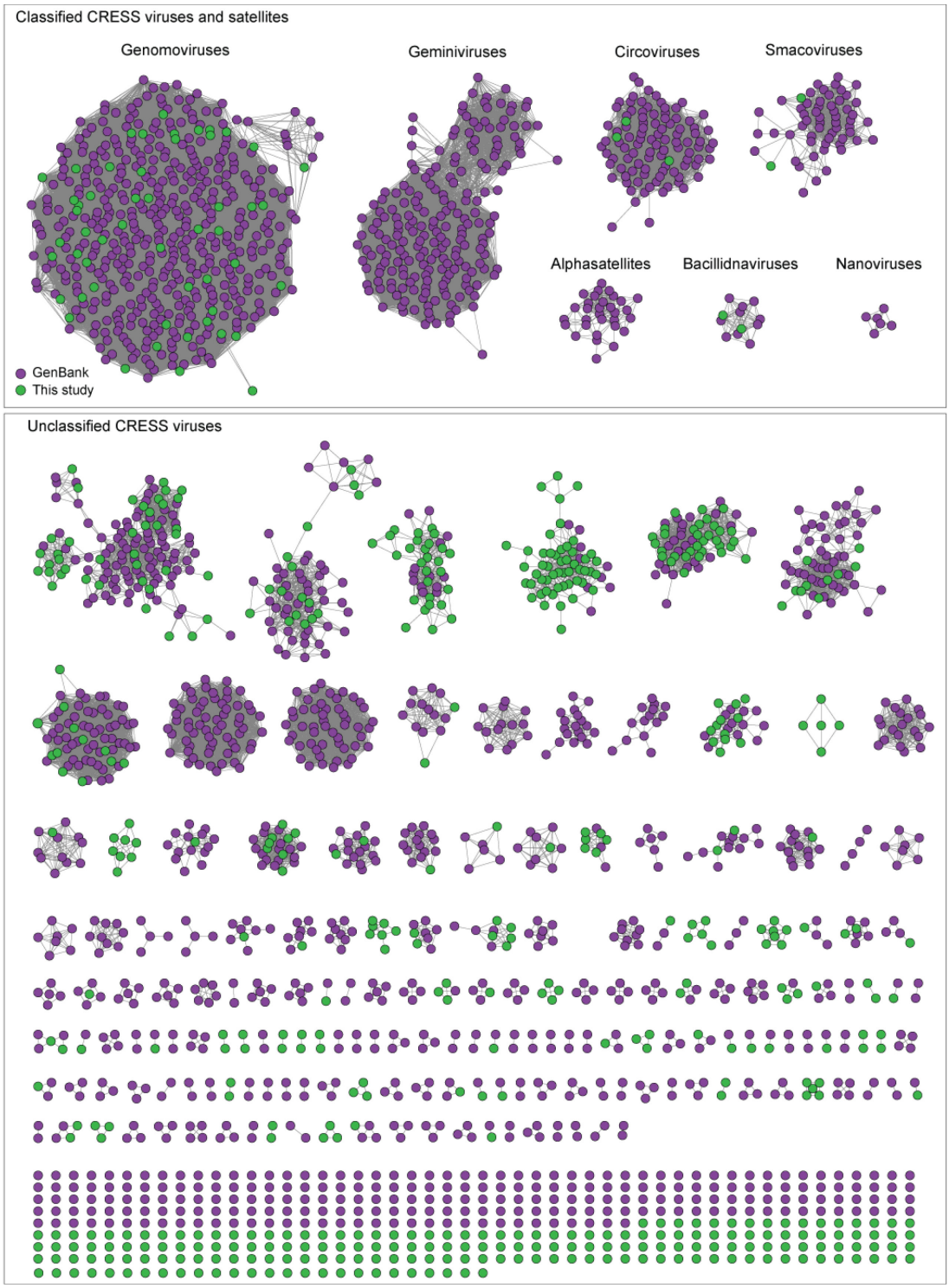


Figure 2.7: Network analysis of CRESS virus Rep proteins

EFI-EST was used to conduct pairwise alignments of amino acid sequences from this study and GenBank that were structurally modelled to be a rolling-circle replicase (Rep). The analysis used an E value cutoff of 10^{-60} to divide the data into family-level clusters.

In Figure 2.4, sequences that showed a structural match to a known eukaryotic circular ssDNA virus capsid protein are displayed as a network. This general capsid type features a single beta-jellyroll fold and assembles into T=1 virions of 20-30 nm in diameter. The network shows that sequences from this study expand and link smaller disconnected clusters of sequences found in GenBank entries (Figure 2.4A-C). Perhaps more importantly a number of previously unknown clusters were identified, providing insight into highly divergent hallmark sequences and making this capsid sequence space amenable to BLAST searches in GenBank (Figure 2.4C). Although the satellite tobacco necrosis virus (STNV) capsid protein encapsidates an RNA molecule, it has previously been noted that its structure is highly similar to the capsid proteins of geminiviruses and other ssDNA viruses [50,83-87] and was included as a model for populating this network.

A similar pattern can be seen in sequence similarity networks for the Rep genes of CRESS viruses (Figure 2.7). Rep genes have been the primary sequences used for taxonomy of CRESS viruses [68]. In this case, it was determined that a network with alignment cutoffs with E values of 1×10^{-60} could split the data neatly into "family-level" clusters [88,89], precisely mirroring ICTV taxonomy of CRESS viruses. Many additional family-level clusters can be discerned from unclassified CRESS viruses. Other eukaryotic and prokaryotic ssDNA virus hallmark gene networks are shown in Figure 2.5. Phylogenetic trees of networks are displayed in Figure 2.6.

Cytoscape files of sequence similarity networks and phylogenetic trees can be found at https://ccrod.cancer.gov/confluence/display/LCOTF/DarkMatter.

New classes of large CRESS viruses feature unconventional structural genes

Although no single family of viruses accounts for the majority of genomes in this study, these results expand the knowledge of the vast diversity of CRESS viruses, which appear to be ubiquitous among eukaryotes [90-93] and are likely to also infect archaea [18,94]. Characterized CRESS viruses have small icosahedral virions (20-30 nm in diameter) with a simple T=1 geometry [95]. This capsid architecture likely limits genome size, as nearly all previously reported CRESS virus genomes and genome segments are under 3.5 kb. Exceptions to this size rule are bacilladnaviruses, which have 4.5 - 6 kb genomes [96] and cruciviruses, which have 3.5 - 5.5 kb genomes [97]. Interestingly, the genomes of these larger CRESS viruses encode capsid genes that appear to have been acquired horizontally from RNA viruses [98]. In our dataset, eight CRESS-like circular genomes exceed 6 kb in length (Figure 2.9). Further, this study's large CRESS genomes are apparently attributable to several independent acquisitions of capsid genes from other taxa and/or capsid gene duplication events.

Notably, a large CRESS genome (CRESS virus isolate ctdh33, associated with rhabditid nematodes that were serially cultured from a soil sample) encoded three separate genes with structural homology (HHpred probability scores 97-99%) to STNV capsid (Figure 2.9G). The three predicted STNV capsid homologs in the nematode virus are highly divergent from one another, with only 28-30% amino acid similarity, but also highly divergent from other amino acid sequences in GenBank. A possible explanation for this observation is that the capsid gene array is the result of gene duplication events.

CRESS genomes ctba10, ctcc19, ctbj26, ctcd34, and ctbd1037 (ranging from 3.5 - 6.2 kb in length) also each encode two divergent capsid gene homologs (Figure 2.9A,B,C,E,H). Single

genomes encoding multiple capsid genes with related but distinct amino acid sequences have been observed in RNA viruses [99] and giant dsDNA viruses [100], but we believe that this is the first time it has been reported in ssDNA viruses.

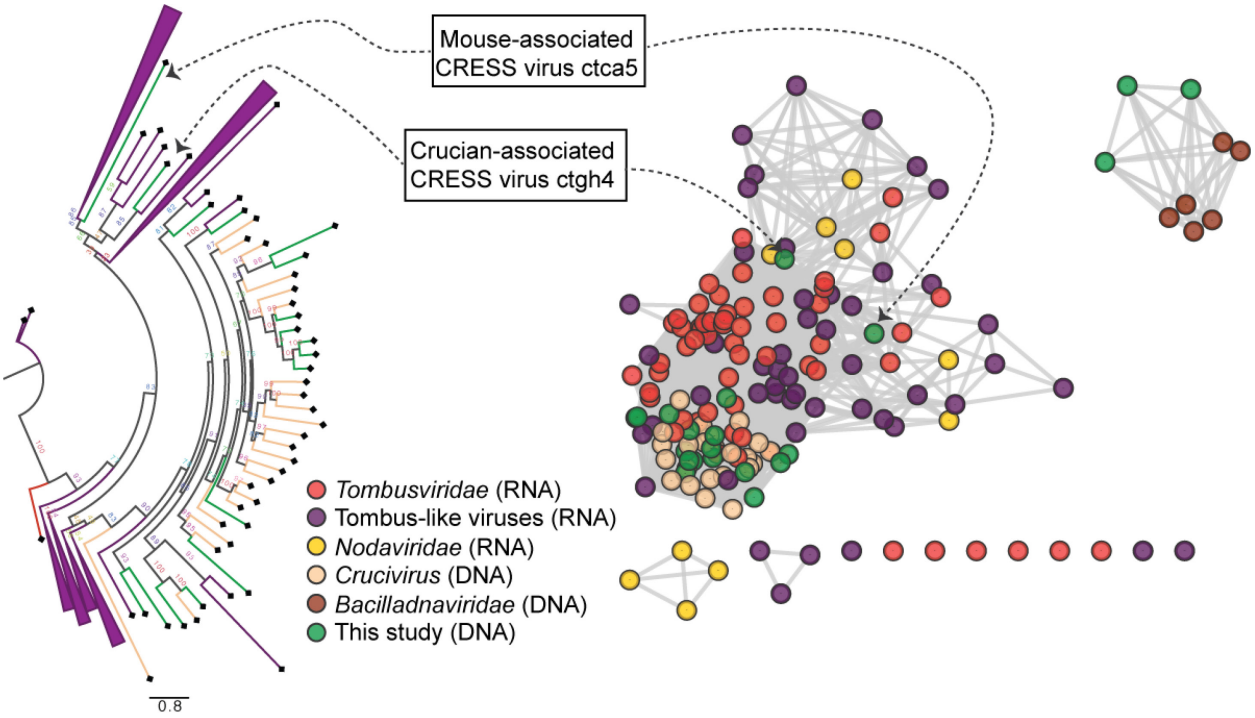


Figure 2.8: RNA virus capsid-like proteins

Sequence similarity network generated with EFI-EST (E value cutoff of 10^{-5}) showing capsid protein sequences of select ssRNA viruses (*Nodaviridae*, *Tombusviridae*, tombus-like viruses) and ssDNA viruses (*Bacilladnaviridae* and crucivirus) together with protein sequences from DNA virus genomes observed in the present study with predicted structural similarity to an RNA virus capsid protein domain (PDB: 2IZW). Predicted capsid proteins for CRESS virus ctca5 and CRESS virus ctgh4 have no detectible similarity to any known DNA virus sequences. On the left, a phylogenetic tree representing the large cluster is displayed. Collapsed branches consist of *Tombusviridae*, tombus-like viruses, and *Nodaviridae* capsid genes.

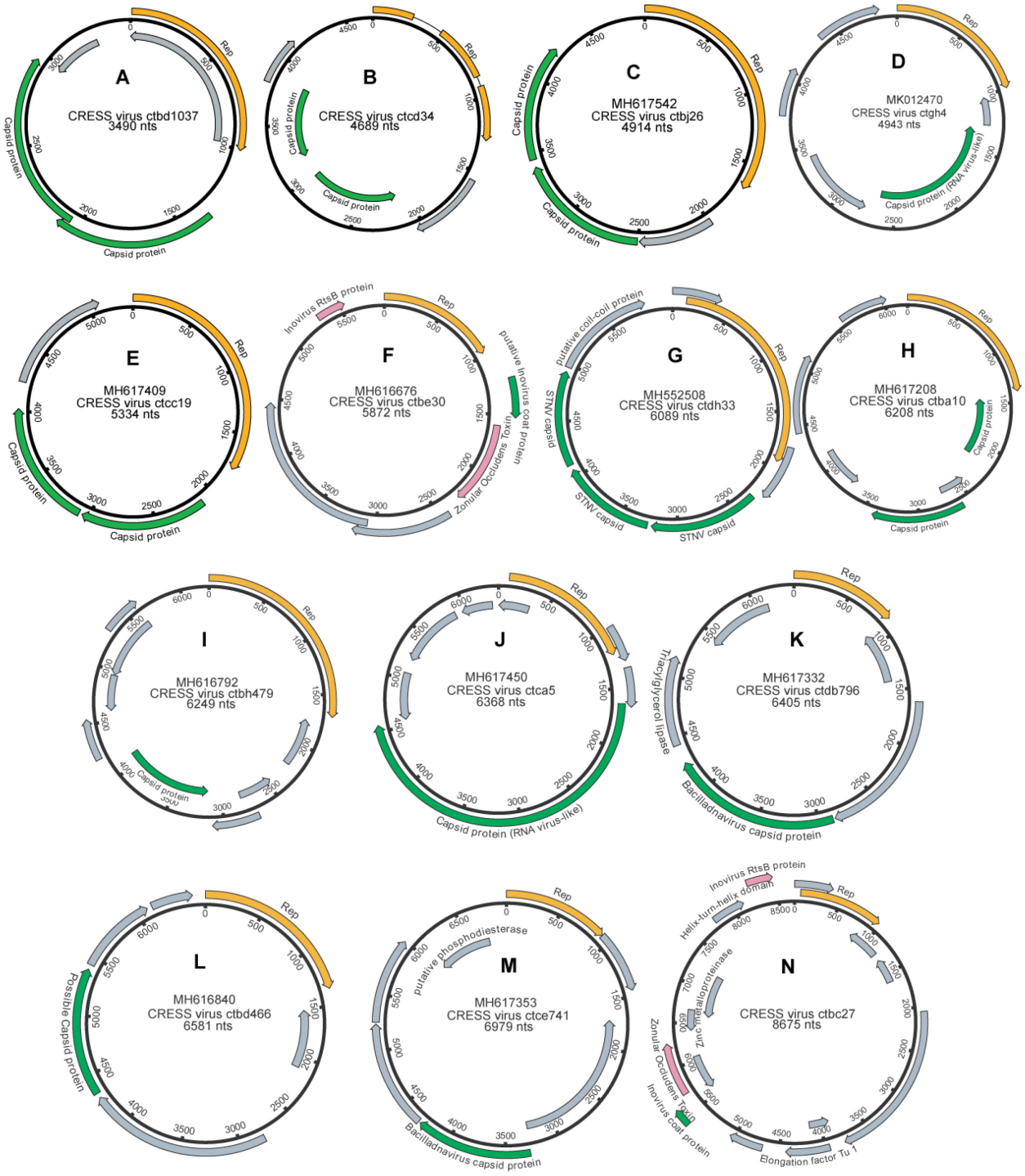


Figure 2.9: Genome maps of large CRESS virus genomes

Predicted CRESS Rep-like genes are displayed in orange, virion structural genes shown in green, other identifiable viral genes shown in pink, other genes in grey. GenBank accession numbers are displayed above the virus name.

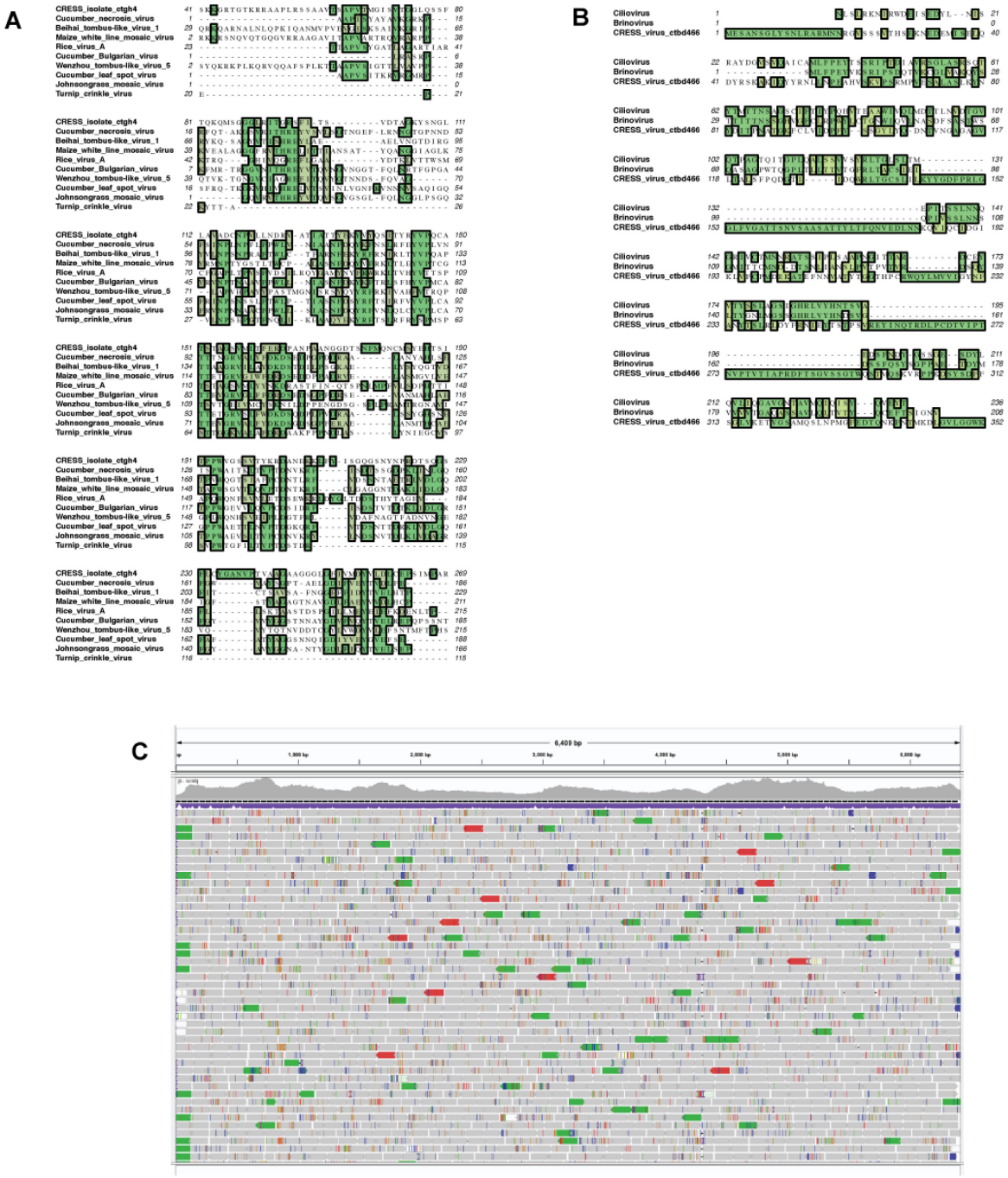


Figure 2.10: Validation of proteins with predicted similarity to RNA virus capsid proteins

(A) First order neighbors for Crucian-associated CRESS virus ctgh4 capsid protein were extracted from the network shown in Figure 2.8 and aligned using Muscle. (B) The same approach was applied to CRESS virus ctbd466 capsid protein. (C) A visualization (Integrative Genomics Viewer) of a read alignment to CRESS virus isolate ctca5. The visualization shows no evidence of artifactual chimerization in the contig assembly process.

Two related large CRESS viruses (ctdb796 and ctce741) encode capsid proteins similar to those of bacilladnaviruses (Figure 2.9K,M). Interestingly, the Rep genes of the two viruses do not show close similarity to known bacilladnavirus Reps and are instead similar to the Reps of certain unclassified CRESS viruses, suggesting that CRESS ctdb796 and CRESS ctce741 are representatives of a new hybrid CRESS virus family.

Two other CRESS virus genomes (isolates ctca5 and ctgh4) encode capsid genes that show amino acid similarity to distinct groups of icosahedral T=3 ssRNA virus capsids [101] (tombus- and tombus-like viruses), but not to cruciviruses or bacilladnaviruses (Figure 2.8, Figure 2.9D,J, Figure 2.10A). Further, a 6.6 kb CRESS virus (isolate ctbd466) (Figure 2.9L) was found to encode a gene with some similarity to the capsid region of the polyprotein of two newly described ssRNA viruses (ciliovirus and brinovirus (Figure 2.10B) [101,102]. Protein fold predictor Phyre^2 [103] showed a top hit (58% confidence) for the capsid protein of a norovirus (ssRNA virus with T=3 icosahedral capsid) for isolate ctbd466 (see GenBank: AXH73946).

Two CRESS genomes (ctbe30 and ctbc27) from separate Rhesus macaque stool samples combine Rep genes specific to CRESS viruses with several genes specific to inoviruses, including inovirus-like capsid genes, which encode proteins that form a filamentous virion (Figure 2.9F,N). The bacteriophage families *Inoviridae* and *Microviridae* are ssDNA viruses that replicate via the rolling circle mechanism, but they are not considered conventional CRESS viruses because they exclusively infect prokaryotes and do not encode Rep genes with CRESS-like sequences. Other inovirus-like genes encoded in the ctbe30 and ctbc27 genomes include homologs of zonular occludens toxin (ZOT, a packaging ATPase) and RstB (a DNA-binding protein required for host genome integration) [104] (Figure 2.9F,N). TBLASTX searches using ctbe30 and ctbc27 sequences

yielded large segments of similarity to various bacterial chromosomes (e.g., GenBank accession numbers AP012044 and AP018536), presumably representing integrated prophages. This suggests that ctbb30 and ctbc27 represent a previously undescribed bacteria-tropic branch of the CRESS virus supergroup.

Viral genomes discussed in this section were validated by aligning individual reads back to the contigs followed by visual inspection. No disjunctions were detected, indicating that illegitimate recombinations are not evident (see Figure 2.10C for an example).

Network analysis of genetic "dark matter" demonstrates conservation of gene sequence and genome structure

We defined potential viral "dark matter" in the survey as circular contigs with no hits with E values $<1\times10^{-5}$ in BLASTX searches of a database of viral and plasmid proteins. We posited that leveraging sequence similarity networks would be useful both for analyzing groups of gene homologs and for discerning which gene combinations tended to be present on related circular genomes. To categorize the 609 dark matter elements based on their predicted proteins, we used pairwise comparison with EFI-EST. A majority of translated gene sequences could be categorized into dark matter protein clusters (DMPCs) containing four or more members (Figure 2.11A). Further, groups of related dark matter elements (i.e. dark matter genome groups (DMGGs)), much like viral families, could be delineated by the presence of a conserved, group-specific marker gene. For example, DMPC1 can be thought of as the marker gene for DMGG1. Certain DMPCs tend to co-occur on the same DMGG. For instance, DMPC7 and DMPC17 ORFs are always observed in genomes with a DMPC1 ORF (i.e., DMGG1) (Figure

2.11B). This *pro tempore* categorization method is useful for visualizing the data, but we stress that is not necessarily taxonomically definitive.

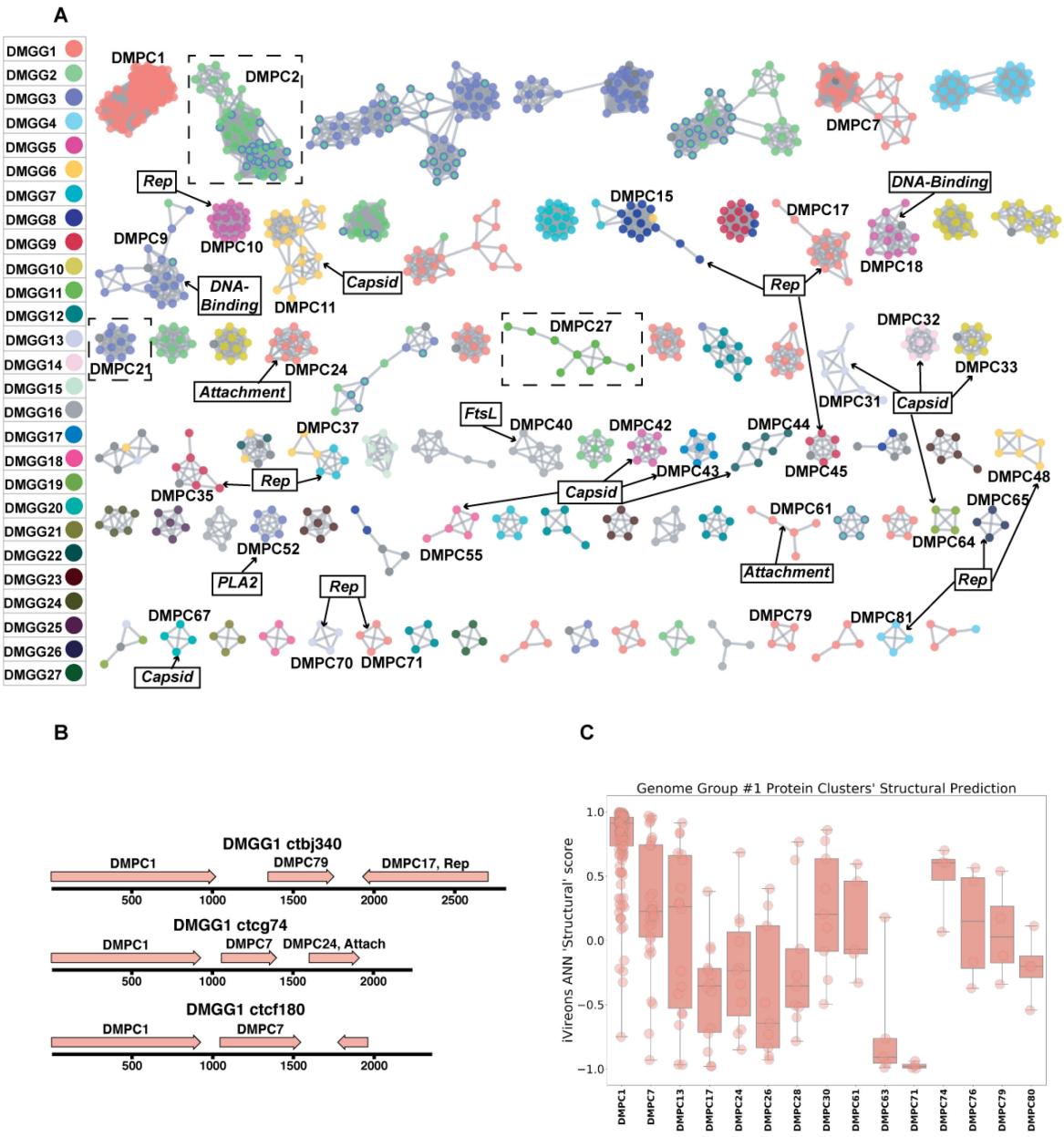


Figure 2.11: Dark matter analysis

(A) Sequence similarity network analysis for genes from dark matter circular sequences (minimum cluster size = 4). Clusters are colored based on assigned dark matter genome group (DMGG). Structural predictions from HHpred are indicated (>85% probability). *Rep* = rolling circle replicases typical of CRESS viruses or ssDNA plasmids. *Capsid* = single-jellyroll capsid protein. *Attachment* = cell attachment proteins typical of inoviruses. *DNA-Binding* = DNA-binding domain. *PLA2* = phospholipase A2. *FtsL* = FtsL-like cell division protein. Clusters that contain a representative protein that was successfully expressed as a virus-like particle are outlined by a dashed rectangle (See Figure 2.14). (B) Maps of three examples of DMGG1 with DMPCs labeled (linearized for display). (C) DMGG1 iVireons "structure" score summary by

protein cluster. Scores range from -1 (unlikely to be a virion structural protein) to 1 (likely to be a virion structural protein). Additional iVireons score summaries can be found in Figure 2.13.

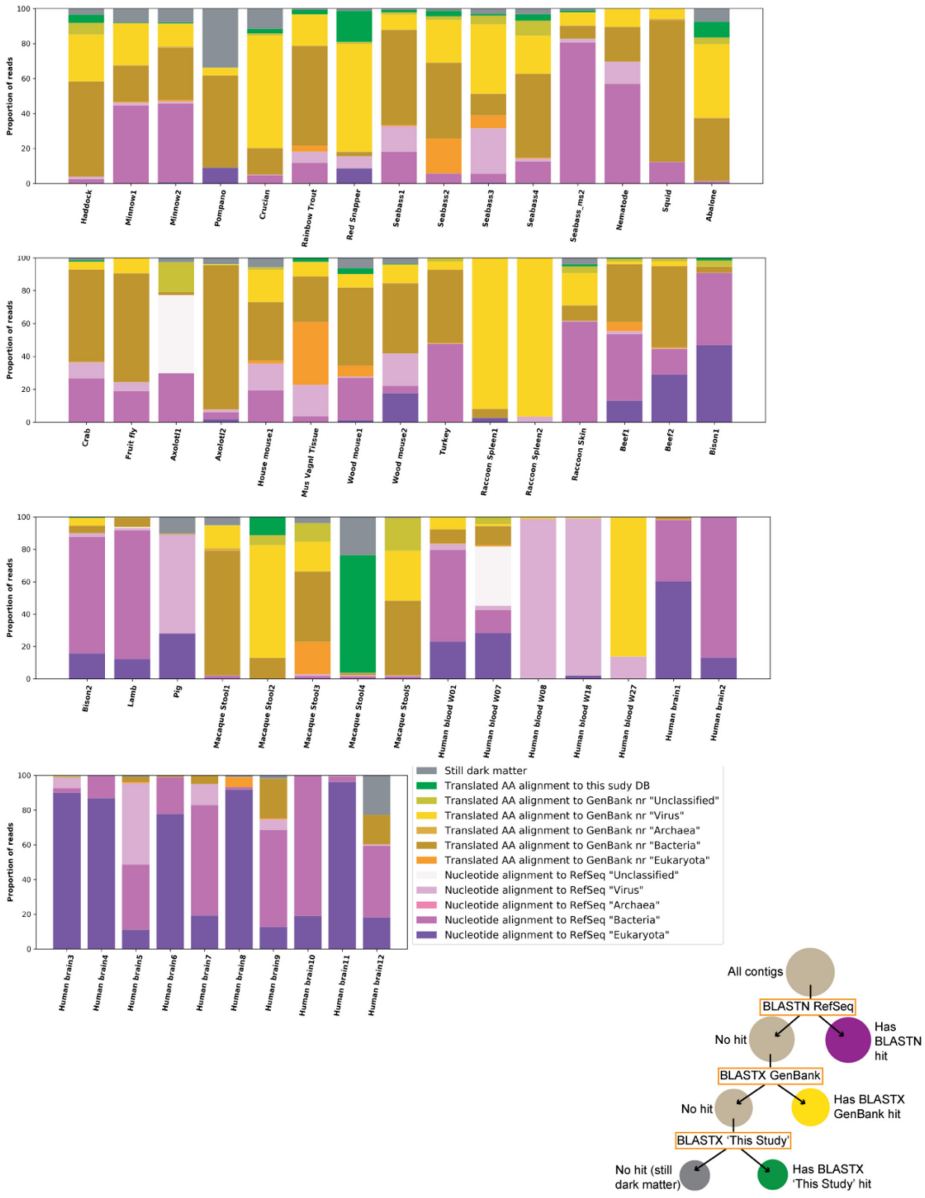


Figure 2.12: Sample characterization by iterative BLAST Searches

Contigs of over 1000 nts from each sample were subject to iterative BLAST searches. First, BLASTN was performed against the RefSeq database. Contigs without hits were then queried by BLASTX against all of GenBank 'nr' database. Contigs without hits were then queried by BLASTX against a database of proteins from genomes reported in this study. The proportion of total reads mapping to each contig was calculated and used for this plot. Individual inspection of contigs shows that most hits in the "Translated AA alignment to GenBank nr 'Bacteria'" were likely plasmid or prophage proteins. The proportions of hits in each category are sensitive to stringency settings and to which databases are chosen for the analysis. The key aims of the figure are to display the proportion of reads the current survey rendered classifiable and the fraction of remaining dark matter reads in various samples.

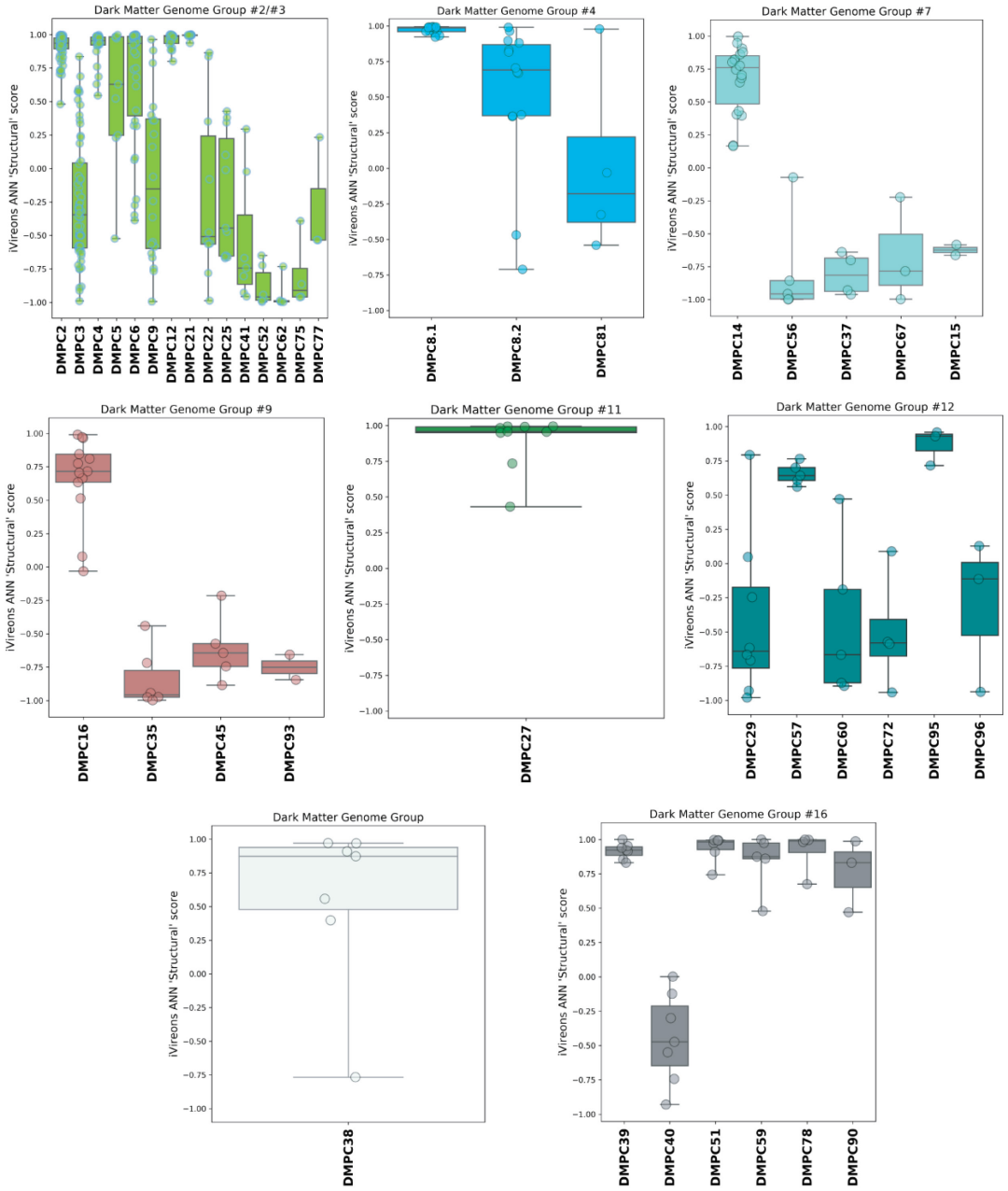


Figure 2.13: iVireons scores of DMGGs with candidate viral structural gene(s)

Box-and-whisker plots of iVireons "Structural" scores for individual DMPCs (numbers on x-axes) grouped by DMGG. Scores (y-axes) range from -1 (unlikely to be a virion structural protein) to 1 (likely to be a virion structural protein). DMGG2 and DMGG3 have been combined due to inferred chimerism.

HHpred, was again employed to make structural predictions for these data [17]. Instead of querying individual sequences, alignments were prepared using MAFFT [105] for each major DMPC to identify conserved residues and increase sensitivity. Then, each alignment was used for an HHpred query. The results indicate that ten DMPCs are likely viral capsid proteins and 11 are rolling circle replicases (Figure 2.11A).

While most of the circular dark matter in the survey could be characterized using these methods, dark matter contigs represent a small remaining fraction in some samples (Figure 2.12).

Cell culture expression of candidate "dark matter" capsids yields particles

In contrast to viral genes such as Rep, with conserved enzymatic functions, sequences of the capsid genes are often poorly conserved, even within a given viral family [106]. Moreover, it appears that capsid proteins have arisen repeatedly through capture and modification of different host cell proteins [16]. This makes it challenging to detect highly divergent capsid proteins using alignment-based approaches or even structural modelling. We therefore turned to an alignment-independent approach known as iVireons, an artificial neural network trained by comparing alignment-independent variables between a large set of known viral structural proteins and known non-structural proteins [12] (https://vdm.sdsu.edu/ivireons/).

Of the 17 DMGGs for which HHPRED did not identify capsid genes, iVireons predicted that ten contain at least one DMPC predicted to encode some type of virion structural protein (median score of cluster >0.70). This allowed us to generate the testable hypothesis that some

of these predicted structural proteins would form virus-like particles (VLPs) if expressed in cell culture.

A subset of predicted capsid proteins were expressed in human-derived 293TT cells and/or in *E. coli* and subjected to size exclusion chromatography. Electron microscopic analysis showed that several of the predicted capsid proteins formed roughly spherical particles, whereas a negative control protein did not form particles (Figure 2.14). Although the particles were highly irregular, the DMGC11 isolate ctgh70 preparation was found to contain nuclease-resistant nucleic acids, consistent with nonspecific encapsidation. The results suggest that, in multiple cases, we were able to experimentally confirm that iVireons correctly predicted the identity of viral capsid proteins.

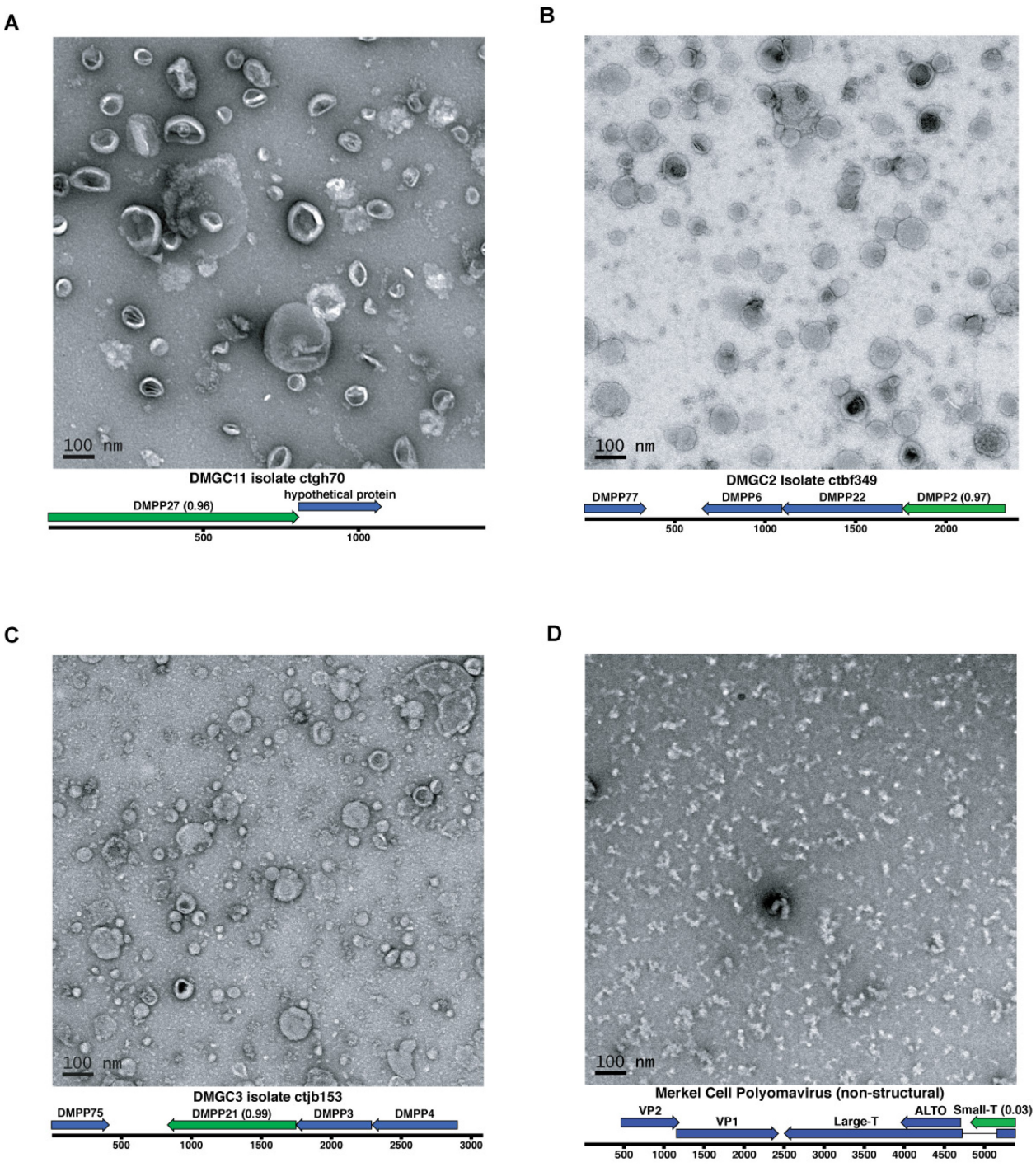


Figure 2.14: Expression of putative capsid proteins

Images taken by negative stain electron microscopy. Genome maps are linearized for display purposes. Expressed genes are colored green. iVireons scores are listed in parentheses. (A-C) Images represent virus-like particles from iVireons-predicted viral structural genes. (D) Merkel cell polyomavirus small T antigen (a viral non-structural protein) is shown as a negative control.

Discussion

Massive parallel DNA sequencing surveys characterizing microbial communities typically yield a significant fraction of reads that cannot be mapped to known genes. The present study sought to provide the research community with an expanded catalog of viruses with circular DNA genomes associated with humans and animals, as well as a means to characterize future datasets. We hope that the availability of this expanded viral sequence catalog will facilitate future investigation into associations between viral communities and disease states. Our annotation pipeline, Cenote-Taker, can be accessed via http://www.cyverse.org/discovery-environment. The CyVerse version of Cenote-Taker can readily annotate circular or linear DNA viruses. RNA viruses with polyproteins or frameshifts will require *post hoc* manual editing. Efforts could be made, for example, to apply the pipeline to previously published viromes to uncover additional viral genomes missed by other methods.

At the present time, GenBank's RefSeq database includes complete sequences for approximately 9,000 viral genomes, most of which fit into 131 families recognized by the International Committee on Taxonomy of Viruses (ICTV) [107]. Similarly, the IMG/VR database contains over 14,000 circular virus genomes from hundreds of studies, though some of these appear to be redundant with each other and are not comprehensively annotated [108]. The current study, which focused on circular DNA viruses with detergent-resistant capsids, found 2,514 new complete circular genomes. The availability of these comprehensively annotated genomes in GenBank contributes new information and understanding to a broad range of established, emerging, and previously unknown taxa. Figure 2.7 shows dozens of potential

family-level groupings within the unclassified CRESS virus supergroup. Sequences from this study contribute to 40 of such groupings and constitute the only members of seven groups. There are also 192 singleton CRESS sequences that could establish many additional family-level groups.

Although small ssDNA viruses are ubiquitous, they are often overlooked in studies that only characterize sequences that are closely related to reference genomes. In addition, ssDNA is not detected by some current DNA sequencing technologies unless second-strand synthesis (such as the RCA approach used in the current study) is conducted.

While many of the viruses discovered in this study appear to be derived from prokaryotic commensals, it is important to note that bacteriophages can contribute to human and animal diseases by transducing toxins, antimicrobial resistance proteins, or genes that alter the physiology of their bacterial hosts [109]. Furthermore, interaction between animal immune systems and bacteriophages appears to be extensive [110].

Over 100 distinct human anellovirus sequences were found in human blood. Anelloviruses have yet to be causally associated with any human disease, but this study indicates that we are likely still just scratching the surface of the sequence diversity of human anelloviruses. It will be important to fully catalog this family of viruses to address the field's general assumption that they are harmless.

Several of the CRESS viruses detected in this study are larger than any other CRESS virus genomes that have been described previously. In some cases, the larger size of these genomes may have been enabled by a process involving capsid gene duplication events. Further, CRESS virus acquisition of T=3 capsids from ssRNA *Nodaviridae* and *Tombusviridae* families has been

previously suggested as the origin of bacilladnaviruses [98] and cruciviruses [52,111-113], respectively. We present evidence of additional independent recombination events between CRESS viruses and ssRNA viruses and ssDNA bacteriophages. In light of these findings, it should be reiterated that only DNA (not RNA) was sequenced in our approach, so DNA/RNA *in silico* false recombination does not seem plausible. These data suggest that CRESS viruses are at the center of a tangled evolutionary history of viruses in which genomes change not just via gradual point mutations but also through larger scale recombination and hybridization events.

It is likely that some dark matter sequences detected in this study share a common ancestor with known viruses but are too divergent to retain discernable sequence similarity. In some cases, the dark matter circles may represent a more divergent segment of a virus with a multipartite genome. Alternatively, some of these sequences likely represent entirely new viral lineages that have not previously been recognized.

Methods

Key Resources Table				
Reagent type (species) or resource	**Designation**	**Source or reference**	**Identifiers**	**Additional information**

strain, strain background (*Escherichia coli*)	T7 Express lysY/I^q *E. coli*	NEB	Cat#: C3013I	
cell line (*Homo-sapiens*)	293TT cells	https://dtp.cancer.gov/repositories/	NCI-293TT	Deposition to ATCC in progress
recombinant DNA reagent	Dark matter capsid expression plasmids	Generated here	Lead contact	
commercial assay or kit	TempliPhi ™ 100 Amplification Kit	Sigma	Cat#: GE25-6400-10	
chemical compound, drug	Optiprep Density Medium	Sigma	Cat#: D1556-250ML	

chemical compound, drug	Sepharose 4B beads	Sigma	Cat#: 4B200-100ML	
software, algorithm	Cenote-Taker	http://www.cyverse.org/discovery-environment	Cenote-Taker 1.0.0	github: https://github.com/mtisza1/Cenote-Taker
software, algorithm	EFI-EST	https://efi.igb.illinois.edu/efi-est/	EFI-EST	
software, algorithm	NCBI BLAST	NCBI	RRID:SCR_004870	
software, algorithm	SPAdes assembler	http://cab.spbu.ru/software/spades/	RRID:SCR_000131	
software, algorithm	A Perfect Circle (APC)	https://github.com/mtisza1/Cenote-Taker/blob/m	APC	

		aster/apc_ct1.pl		
software, algorithm	EMBOSS suite (getorf)	http://emboss.sourceforge.net/	RRID:SCR_008493	
software, algorithm	Circlator	http://sanger-pathogens.github.io/circlator/	RRID:SCR_016058	
software, algorithm	HHSuite	https://directory.fsf.org/wiki/Hhsuite	RRID:SCR_016133	
software, algorithm	tbl2asn	https://www.ncbi.nlm.nih.gov/genbank/tbl2asn2/	RRID:SCR_016636	
software, algorithm	MacVector	http://macvector.com	RRID:SCR_015700	

software, algorithm	Bandage	https://rrwick.github.io/Bandage/	Bandage	

LEAD CONTACT AND MATERIALS AVAILABILITY

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Chris Buck (buckc@mail.nih.gov)

METHOD DETAILS

Sample collection and sequencing

De-identified human swabs and tissue specimens were collected under the approval of various Institutional Review Boards (Supplemental File 2). Animal tissue samples were collected under the guidance of various Animal Care and Use Committees.

Nematodes were cultured out of soil samples collected in Bethesda, Maryland, USA on OP50-Seeded NGM-lite plates (*C. elegans* kit, Carolina Biological Supply).

Viral particles were concentrated by subjecting nuclease-digested detergent-treated lysate to ultracentrifugation over an Optiprep step gradient, as previously described https://ccrod.cancer.gov/confluence/display/LCOTF/Virome [57]. Specifically, for each sample, no more than 0.5 g of solid tissue was minced finely with a razorblade. Alternatively, no more than 500 µl of liquid sample was vortexed for several seconds. Samples were transferred to 1.5ml siliconized tubes. The samples were resuspended in 500 µl Dulbecco's PBS and Triton X-100 (Sigma) detergent was added to a final concentration of 1% w/v. 1 µl of Benzonase (Sigma) was added. Samples were vortexed for several seconds. Samples were incubated in a 37°C water bath for 30 minutes, with brief homogenizing using a vortex every 10 minutes. After incubation, NaCl was added to the samples to a final concentration of 0.85M. Tubes were spun for 5 minutes at 5000g. Resulting supernatants were transferred to a clean siliconized tube. Supernatant-containing tubes were spun for an additional 5 minutes at 5000g. Resulting supernatants were added to iodixanol/Optiprep (Sigma) step gradients in ultracentrifuge tubes (Beckman: 326819) (equal volumes 27%, 33%, 39% iodixanol with 0.8M NaCl; total tube volume, including sample, ~5.1ml). Ultracentrifuge tubes were spun at 55,000rpm for 3.5 hours (Beckman: Optima L-90K Ultracentrifuge). After spin, tubes were suspended over 1.5ml siliconized collection tubes and pierced at the bottom with 25G needle. Six fractions of equal volume were collected drop-wise from each ultracentrifuge tube.

From each fraction, 200 µl was pipette to a clean siliconized tube for virus particle lysis and DNA precipitation. To disrupt virus particles, 50 µl of a 5X master mix of Tris pH 8 (Invitrogen, final conc. 50mM), EDTA (Invitrogen, final conc. 25mM), SDS (Invitrogen, final conc. 0.5%), Proteinase K (Invitrogen, final conc. 0.5%), DTT (Invitrogen, final conc. 10mM) was added and

mixed by pipetting up and down. Samples were heated at 50°C for 15 minutes. Then, proteinase K was inactivated for 10 minutes at 72°C. To the 250 µl of sample, 125 µl of 7.5M ammonium acetate was added and mixed by vortexing. Then, 975 µl of 95% ethanol was added and mixed by pipetting. This was incubated at room temperature for 1 hour. Then, the samples were transferred to a 4°C fridge overnight.

Samples were then restored to ambient temperature. Then, samples were spun for 1 hour at 20,000g in a temperature-controlled tabletop centrifuge set to 21°C. Supernatant was aspirated, and 500 µl ethanol was added to each pellet. Pellets were resuspended by flicking. Then, samples were spun for 30 minutes at 20,000g in a temperature-controlled tabletop centrifuge set to 21°C. Supernatant was aspirated, and samples were spun once more at 20,000g for 3 minutes. Remaining liquid was carefully removed with a 10 µl micropipette. Tubes were left open and air dried for at least 10 minutes.

DNA from individually collected fractions of the gradient was amplified by RCA using phi29 polymerase (TempliPhi, Sigma) per manufacturer's instructions. While we expected most viral particles to travel to the middle of the gradient based on previous experiments, RCA was conducted on individual fractions spanning the gradient, in an attempt to detect viruses with different biophysical properties [114]. Pooled, amplified fractions were prepared for Illumina sequencing with Nextera XT kits. Then libraries were sequenced with Illumina technology on either MiSeq or NextSeq500 sequencers. Contigs were assembled using SPAdes with the "plasmid" setting. Circularity was confirmed by assessing assembly graphs using Bandage [115].

Analysis of brain samples

Brain samples were initially analyzed by Optiprep gradient purification, RCA amplification, and deep sequencing, as described above. JC polyomavirus, which has previously been reported in brain samples [116], can display high buoyancy in Optiprep gradients [117]. Fractions from near the top of the Optiprep gradient were subjected to an alternative method of virion enrichment using microcentrifuge columns (Pierce) packed with 2 ml of Sepharose 4B Bead suspension (Sigma) exchanged into PBS. Fractions were clarified at 5000 x g for 1 minute, and 200 µl of clarified extract was loaded onto the gel bed. The column was spun at 735 x g and the eluate was digested with proteinase K, ethanol-precipitated, and subjected to RCA. No additional viral sequences were detected by this method.

The brain samples were also subjected to confirmatory analysis by RNA sequencing. RNA was extracted from brain tissues with Qiagen Lipid Tissue RNeasy Mini Kit and subjected to human ribosomal RNA depletion with Thermo RiboMinus. The library was prepared with NEBNext Ultra™ II Directional RNA Library Prep Kit for Illumina and subjected to massive parallel sequencing on the Illumina HiSeq platform (see BioProject PRJNA513058).

Cenote-Taker, Virus Discovery and Annotation Pipeline

Cenote-Taker, a bioinformatics pipeline written for this project and fully publicly available on CyVerse, was used for collection and detailed annotation of each circular sequence. The flow of the program can be described as follows:

(1) Identifies and collects contigs (assembled with SPAdes) larger than 1000 nts

(2) Predicts which contigs are circular based on overlapping ends

(3) Determines whether circular contig has any ORFs of 80 AA or larger or else discards sequence

(4) Uses BLASTN against GenBank “nt” database to disregard any circular sequences that are >90% identical to known sequences across a >500 bp window

(5) Uses Circlator [118] to rotate circular contigs so that a non-intragenic start codon of one of the ORFs will be the wrap point

(6) Uses BLASTX against a custom virus + plasmid database (derived from GenBank “nr” and RefSeq) to attempt to assign the circular sequence to a known family

(7) Translates each ORF of 80 AA or larger

(8) Uses RPS-BLAST to predict function of each ORF by aligning to known NCBI Conserved Domains

(9) Generates a tbl file of RPS-BLAST results

(10) Takes ORFs without RPS-BLAST hits and queries the GenBank “nr viral” database with BLASTP

(11) Generates a tbl file of BLASTP results

(12) Takes ORFs without any BLASTP hits and queries HHblits (databases: uniprot20, pdb70, scop70, pfam_31, NCBI_CD)

(13) Generates a tbl file of HHblits results

(14) Complies with a GenBank request to remove annotations for ORFs-within-ORFs that do not contain conserved sequences

(15) Combines all tbl files into a master tbl file

(16) Generates a unique name for each virus based on taxonomic results

(17) Generates properly formatted fsa and tbl files in a separate directory

(18) Uses tbl2asn to make gbf (for viewing genome maps) and sqn files (for submission to GenBank)

The source code can be found at: https://github.com/mtisza1/Cenote-Taker

This work utilized the computational resources of the NIH HPC Biowulf cluster. (http://hpc.nih.gov).

Genome maps were drawn, and multiple sequence alignments were computed and visualized using MacVector 16.

Anelloviruses

Analysis of linear contigs in the survey found many instances of recognizable viral sequences. One noteworthy example were anelloviruses, where many contigs terminated near the GC-rich stem-loop structure that is thought to serve as the origin of replication. This segment of the anellovirus genome is presumably incompatible with the short read deep sequencing technologies used in this study. Nearly complete anellovirus genomes, defined as having a complete ORF1 gene and at least 10-fold depth of coverage, were also deposited in GenBank (Supplementary File 2).

GenBank Sequences

Amino Acid sequences from ssDNA viruses were downloaded in June 2018 based on categories in the NCBI taxonomy browser. As many sequences in GenBank are from identical/closely related isolates, all sequences were clustered at 95% AA ID using CD-HIT [119].

Sequence Similarity Networks

Amino acid sequences from GenBank (see above) and this study were used as queries for HHsearch (the command-line iteration of HHpred) against PDB, PFam, and CDD. Sequences that had hits in these databases of 80% probability or greater were kept for further analyses. Note that capsid protein models for some known CRESS virus families have little, if any, similarity to other capsid sequences and have not been determined (e.g. *Genomoviridae* and *Smacoviridae*) and were therefore not displayed in networks. Models used: (CRESS virus capsids network:5MJF_V, 3R0R_A, 5MJF_Ba, 4V4M_R, 4BCU_A, PF04162.11, 5J37_A, 5J09_C, 3JCI_A, cd00259, PF04660.11, PF03898.12, PF02443.14, pfam00844); (CRESS virus Rep network:4PP4_A, 4ZO0_A, 1M55_A, 1UUT_A, 1U0J_A, 1S9H_A, 4R94_A, 4KW3_B, 2HWT_A, 1L2M_A, 2HW0_A, PF08724.9, PF17530.1, PF00799.19, PF02407.15, pfam08283, PF12475.7, PF08283.10, PF01057.16, pfam00799); (*Microviridae*/*Inoviridae* replication-associated protein: 4CIJ_B, 4CIJ_C, PF05155.14, PF01446.16, PF11726.7, PF02486.18, PF05144.13, PF05840.12); (*Microviridae* capsid: 1M06_F, 1KVP_A, PF02305.16); (*Anelloviridae* ORF1: PF02956.13); (*Inoviridae* ZOT: 2R2A_A, PF05707.11).

Phylogenetic Trees

Sequences from this study and GenBank were grouped by structural prediction using HHpred. Then, sequences were compared by EFI-EST to generate clusters with a cut-off of $1x10^{-5}$. Sequences from these clusters were then extracted and aligned with PROMALS3D [120] using structure guidance, when possible. Structures used: (*Microviridae* MCP: 1KVP); (CRESS virus capsid STNV-like: 4V4M); (CRESS virus capsid circo-like: 3JCI); (*Inoviridae* ZOT: 2R2A); (CRESS virus Rep: 2HW0) (CRESS virus/RNA virus S Domain capsid: 2IZW). The resulting alignments were used to build trees with IQ-Tree with automatic determination of the substitution model and 1000 ultrafast bootstraps [121]. Models used: (*Microviridae* MCP: Blosum62+F+G4); (*Microviridae* Rep I: Blosum62+I+G4); (*Microviridae* Rep II: LG+I+G4); (*Microviridae* Rep III: VT+I+G4); (CRESS virus/RNA virus S Domain capsid: Blosum62+F+G4); (*Circoviridae* capsid: VT+F+G4); (CRESS virus capsid STNV-like: VT+F+G4); (*Inoviridae* ZOT: VT+I+G4); (*Anelloviridae* ORF1: VT+F+G4). Trees were visualized with FigTree (http://tree.bio.ed.ac.uk/software/figtree/) and iTOL [122].

Expressing Potential Viral Structural Proteins in human 293TT cells

293TT cells were transfected with potential viral structural protein expression constructs for roughly 48 hours. Cells were lysed in a small volume of PBS with 0.5% Triton X-100 or Brij-58 and Benzonase (Sigma). After several hours of maturation at neutral pH, the lysate was clarified at 5000 x g for 10 min. The clarified lysate was loaded onto a 27-33-39% Optiprep gradient in PBS with 0.8 M NaCl. Gradient fractions were collected by bottom puncture of the tube and screened by PicoGreen nucleic acid stain (Invitrogen), BCA, and SDS-PAGE analysis. Electron microscopic analysis was then performed. Expression in 293TT cells of some "dark matter" virus

capsids was attempted but not successful in any case. 293TT cells were generated in-house for the previous paper [123], and passages from original stocks were used. Authentication testing has not been conducted. Mycoplasma testing is conducted annually using MycoScope PCR Mycoplasma Detection Kit (Genlantis).

Expressing Potential Viral Structural Proteins in *E. coli*

Several genes that were identified by iVireons as being potential viral structural proteins were cloned into plasmids with a T7 polymerase-responsive promoter. Plasmids were transfected into T7 Express lysY/I^q *E. coli*, which express T7 polymerase under the induction of IPTG. Bacteria were grown at 37°C in LB broth until OD600 = 0.5. Flasks were cooled to room temperature, IPTG was added to 1 mM, and cultures were shaken at room temperature for approximately 16 hours. Cells were then pelleted for immediate processing.

Total protein was extracted with a BPER (Pierce) and nuclease solution. Then, virion-sized particles were enriched from the clarified lysate using size exclusion chromatography with 2% agarose beads https://ccrod.cancer.gov/confluence/display/LCOTF/GelFiltration. Fractions were analyzed using Coomassie-stained SDS-PAGE gels for presence of a unique band corresponding to the expressed protein. Fractions of interest were analyzed using negative stain electron microscopy.

Electron Microscopy

Five µl samples were adsorbed onto a carbon-deposited copper grid for one minute. Sample was then washed 5 times on water droplets then stained with 0.5% uranyl acetate for 1 second.

The negatively stained samples were examined on a FEI Tecnai T12 transmission electron microscope.

ViromeQC

ViromeQC was run on reads from each sample corresponding to an SRA run. The "human" setting was used, and the diamond alignment to "31 prokaryotic single-copy markers" was reported.

Mapping reads to reference genomes

Viral genomes from RefSeq were downloaded from NCBI. On RefSeq and "This study" genomes, RepeatMasker was used with "-noint" and "-hmmer" settings to mask low-complexity regions to prevent nonspecific mapping. However, this likely led to some degree of under-mapping. Reads were trimmed with fastp and aligned with Bowtie2 using default settings.

Sequencing

Illumina sequencing was conducted at the CCR Genomics Core at the National Cancer Institute, NIH, Bethesda, MD 20892.

DATA AND CODE AVAILABILITY

All reads and annotated genomes associated with this manuscript can be found on NCBI BioProject Accessions PRJNA393166 and PRJNA396064.

Cenote-Taker, the viral genome annotation pipeline, can be used by interested parties on the Cyverse infrastructure: http://www.cyverse.org/discovery-environment.

ADDITIONAL RESOURCES

Relevant protocols on lab website: https://ccrod.cancer.gov/confluence/display/LCOTF/Virome

Acknowledgements

This research was supported [in part] by the Intramural Research Program of the NIH, National Cancer Institute.

We would like to acknowledge the GenBank team at NCBI for productive discussion about their viral genome submission requirements and facilitation of annotated genome deposition.

Declaration of Interests

The authors declare no competing interests.

Contributions

Michael J Tisza: Conceptualization, Resources, Data curation, Software, Formal analysis, Validation, Investigation, Visualization, Methodology, Project administration

Diana V Pastrana: Conceptualization, Formal analysis, Investigation, Methodology

Nicole L Welch: Formal analysis, Investigation, Methodology

Brittany Stewart: Formal analysis, Investigation

Alberto Peretti: Formal analysis, Investigation

Gabriel J Starrett: Data curation, Software, Formal analysis

Yuk-Ying S Pang: Formal analysis, Investigation

Siddharth R Krishnamurthy: Software, Formal analysis

Patricia A Pesavento: Formal analysis, Investigation

David H McDermott: Resources, Investigation

Philip M Murphy: Resources, Investigation

Jessica L Whited: Resources, Investigation

Bess Miller: Resources, Investigation

Jason Brenchley: Resources, Investigation

Stephan P Rosshart: Resources, Investigation

Barbara Rehermann: Resources, Investigation

John Doorbar: Resources, Investigation

Blake Ta'ala: Conceptualization, Resources

Olga Pletnikova: Resources, Investigation

Juan C Troncoso: Resources, Investigation

Susan M Resnick: Resources, Investigation

Ben Bolduc: Resources, Investigation, Software

Matthew B Sullivan: Conceptualization, Software

Arvind Varsani: Conceptualization, Data curation, Formal analysis, Supervision, Investigation, Visualization, Methodology

Anca M Segall: Conceptualization, Data curation, Formal analysis, Supervision, Investigation, Visualization, Methodology

Christopher B Buck: Conceptualization, Resources, Data curation, Formal analysis, Supervision, Funding acquisition, Investigation, Visualization, Methodology, Project administration

3 Bibiviruses are a New, Unusual Virus Family Common in the Human Gut

Abstract

Humans are covered in a complex network of microbial life, including bacteria, eukaryotes, and viruses. Uncovering these lifeforms and discerning which are important for health and disease has been the topic of much research, especially since metagenomic sequencing has enabled detection of organisms without having to propagate them in culture. Viruses still remain largely mysterious, however, and virus-enriched samples typically contain an abundance of unrecognizable sequences. Some of these unknown sequences are likely to be viruses of previously unrecognized types. In this report we show that one group of previously unrecognizable sequences are an unusual class of viruses infecting *Bacteroidetes* bacteria. Computational and experimental evidence suggests that the emerging "bibiviruses" encode a major capsid protein completely different from any known virion structural protein. Bibiviruses are among the most abundant virus taxa in human gut samples.

Introduction

The network of complex microbial life residing in and on humans is known as the microbiome[36]. It is estimated that each person contains at least as many bacterial cells as human cells[124]. Many human cells have active or latent viral infections, and each bacterial species likely hosts multiple viral species[4]. Because of this, there may be more viral genome copies within humans than all the cellular genome copies combined[125].

The effects of the bacterial component of the human microbiota on human health and disease has been under intense study. It is clear that bacteria can have a variety of positive effects on us[21,124] and, at the same time, associations between specific bacteria and a variety of diseases have been established[126-131]. However, just as infection with certain viruses can have drastic effects on humans, bacterial viruses (phages) can also have drastic effects on their host. The predator/prey relationship that defines many phage/bacterium dynamics can alter the population density of certain bacteria[23], effectively regulating the "dose" of that bacterium. On the other hand, many phages will be retained as an integrated or episomal prophage within the host bacterium, rarely lysing their host cells[132]. These temperate phages often contain genes that can dramatically alter the phenotype of the bacteria, such as toxins[133], virulence factors[24], antibiotic resistance genes[134], photosystem components[135], other auxiliary metabolic genes[136], and CRISPR-Cas systems[137], along with countless genes of unknown function. To understand how humans achieve a successful or unsuccessful balance as a "holobiont"[138], the effects and mechanisms of phages on their bacterial hosts and, in turn, on us will have to be understood.

In general, the phage knowledge base, including genome sequences and structure of virion proteins, stems from phages that cause dramatic plaques of lysed bacteria on lawns of easily culturable bacterial species[9]. Although this has been a powerful approach for identifying and describing a diverse range of phages, it remains unclear how many additional phage classes have been missed with the classic methods.

In many individuals, bacteria from the order *Bacteroidales* are the most abundant taxon in the gut. *Bacteroidales* are also the source of much interpersonal variation[30]. These bacteria, like many other anaerobes, are challenging to grow in culture, often precluding them from

biochemical interrogation or use in cell culture studies. Despite this, recent advances have demonstrated roles for specific members of this order in immune regulation[139], cardiomyopathy[140], and alleviation of obesity[141]. The phage diversity of *Bacteroidales* is largely unaccounted for. For example, only in 2014 was the *Bacteroidales* phage crAssphage discovered, and determined to be by far the most abundant known phage in the human gut[14]. The discovery relied entirely on computational assessment of metagenomic data and, while it is a tailed dsDNA phage like other caudoviruses, it was so divergent from any previously characterized virus sequence that it was proposed to represent a new viral family. It took several years before a related "crAss-like" phage was able to be cultured[142] and the initially discovered crAssphage has still not been cultured.

The problem of a limited assay being the basis for almost all phage knowledge is apparent when looking at metagenomic sequences from "viromic" samples enriched for nuclease-resistant sequences associated with virion-sized particles. Typically, 50-70% of sequences from massively parallel sequencing of viromes are not discernably similar to known virus types[67]. Some of these unknown sequences may be mobile genetic elements that parasitize viruses by packaging themselves into viral capsids (for example, phage-inducible chromosomal islands (PICIs))[143]. Other unrecognizable sequences may be viruses that are too divergent from known virus types to be detected by current bioinformatic methods. A theoretical third category of unrecognizable elements could be viruses that have no shared ancestry with known virus families.

In this study, we describe bibviruses, a new type of virus infecting *Bacteroidales* bacteria. Bibiviruses encode a major capsid protein that is not recognizably similar to previously

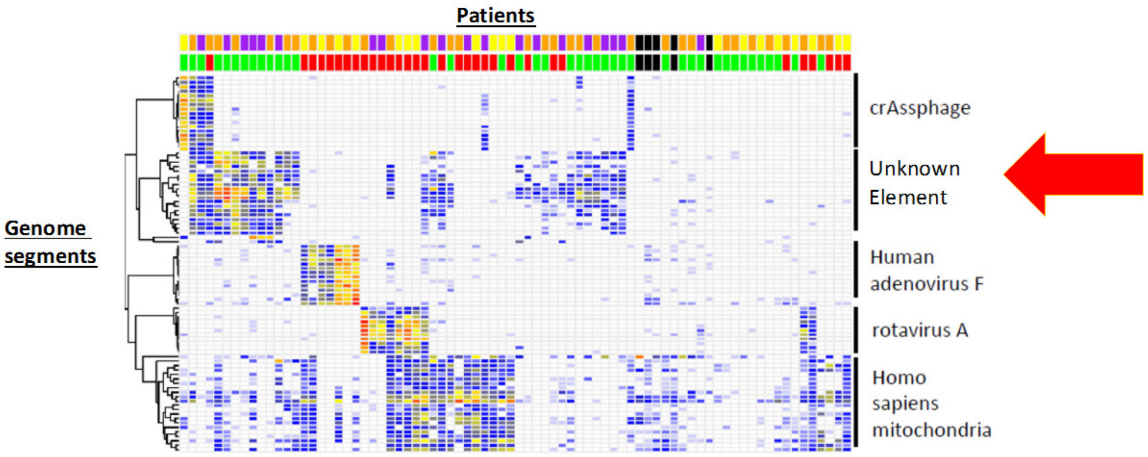
identified capsid proteins. Virome analysis reveals that at hundreds of unique bibiviruses exist and are present in human gut samples from around the world, comprising up to 36% percent of reads in viromic datasets. A strain of *Parabacteroides distasonis* with an integrated bibivirus (prophage) was cultured, and bibivirus particles were isolated.

Results

Identification of a common unknown element in stool

As part of the Global Enteric Multi-center Study[144], i.e. "GEMS", Viromic preps of filtered stool from 78 children, some with diarrheal disease, living in The Gambia were sequenced and analyzed (see Methods). Unlike many analyses in which reads or contigs are only considered if they fall into recognizable bins based on similarity to known viruses, the most prevalent sequences across all samples were determined by creating an occurrence profile for each unique gene[14]. As expected, sequences corresponding to adenovirus, rotavirus, and crAssphage were all detected, alongside human mitochondrial DNA (Fig. 3.1A). The most common non-human sequences (present in about half of patients) corresponded to an unknown element of just over 16kb. This element encodes a superfamily 2 DNA helicase gene (replicase), a Xer-like tyrosine recombinase (integrase), and a lysozyme gene distantly related to those of dsDNA phages and other transposable mobile genetic elements. The elements were not found to encode detectable virion structural genes based on BLAST and HHpred searches. This unknown viromic element, shown graphically in Fig 3.1B, shows high nucleotide similarity over half its length to a chromosomal region from *Alistipes megaguti* (GenBank LR027382), a bacterium in the order *Bacteroidales*, as well as translated nucleotide TBLASTX similarity across ~60% of its length to a handful of *Bacteroides* and *Parabacteroides* chromosomes.

A



B

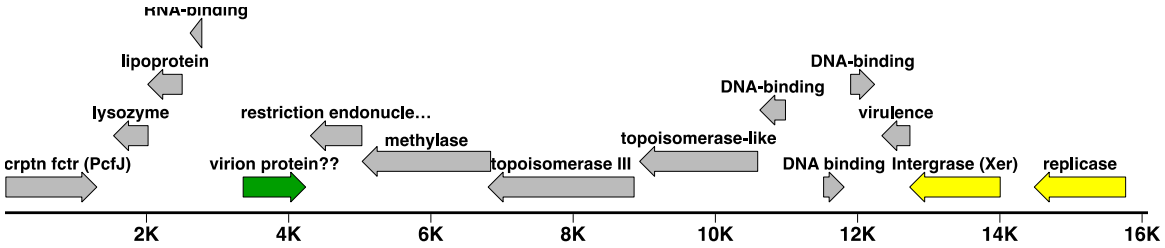


Figure 3.1 Identification of an unknown element in virome samples of children in Gambia
(A) Heatmap of highly abundant genes from viromic preps of fecal samples from Gambian children. The top bar indicates subject age (purple=1 year, orange=2 years, yellow=3-4 years, black=unavailable). Second row of bars indicate diarrheal status (green=healthy, red=diarrheal, black=unavailable). (B) The genome map for the unknown viromic element.

* Figure 3.1A Courtesy of Mathieu Almeida and Laura Tanase (U. Maryland)

Analysis and expression of a novel putative major capsid protein

In order to detect potential virion structural genes encoded by this element that lacked discernable homology known sequences and structures, an alignment-independent prediction strategy, iVireons, was used to analyze all genes[12]. To reduce noise from the signal of iVireons, homologues of each gene were recruited via BLASTP of GenBank nr and fed together into iVireons using the major capsid protein (MCP) model. One conserved ORF gave uniformly high iVireons MCP scores (Fig 3.2A) while other predicted proteins generally showed low or non-uniform iVireons scores (data not shown).

Computational analyses of the candidate MCP predicted that the secondary structure consists of several alpha-helical domains and a disordered C-terminus (Fig 3.2C). This structure is conserved in all homologues (data not shown). Overall, it is likely that this protein encodes multiple coil-coil domains (Fig 3.2B). MCPs of dsDNA phage that form icosahedral virions typically fold into beta jellyroll structures, but the lack of beta sheets in the unknown viromic element likely precludes that possibility. Some MCPs, such as those from Hepadnaviruses are composed primarily of alpha helices with a C-terminal disordered domain, but these share no amino acid similarity to the Gambia virome element protein.

A codon-optimized version of the candidate MCP was synthesized and expressed in *E. coli*. The expressed proteins assembled into roughly spherical ~100 nm particles similar to immature capsid particles (Fig 3.2E).

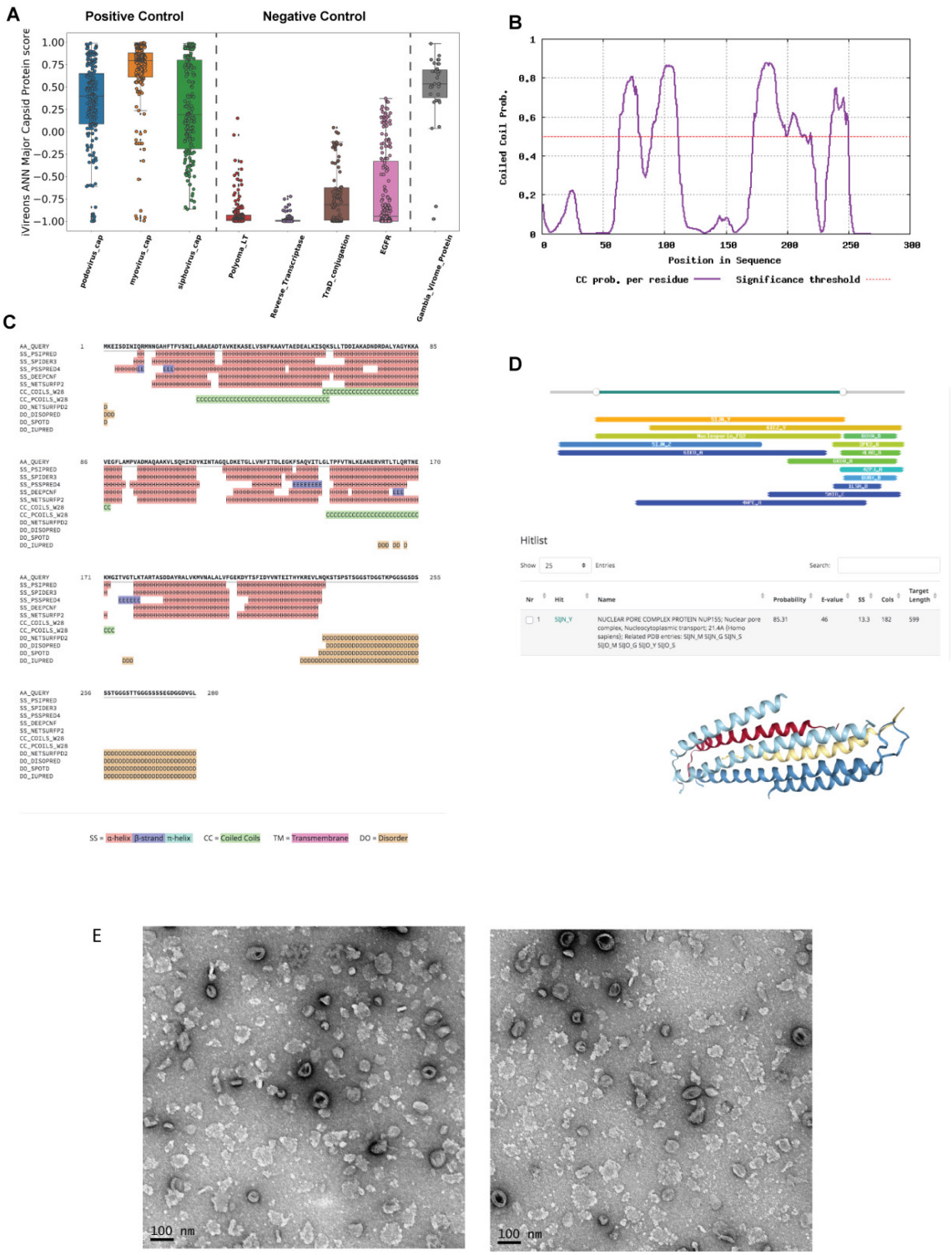


Figure 3.2 Analysis and Expression of prospective capsid protein from unknown viromic element

(A) iVireons scores of different gene families from 1 (likely a capsid protein) to -1 (unlikely a capsid protein). (B) Coil-coil probability plot along the length of an example of the candidate MCP. (C) Secondary structure analysis. (D) Structural homology prediction and illustration of the

coil-coil domain of nuclear pore complex nup155 protein. (E) Negative stain electron microscopy image of the unknown viromic element putative capsid protein structures from exogenous expression and virus-like particle enrichment in *E. coli*.

Induction of virion production with bile salts

To establish whether the novel element represents a class of bona fide viruses, sequences of bacteria from order *Bacteroidales* were scanned for genes with similarity to the candidate virion protein (Fig 3.2) with the goal of finding a culturable isolate carrying an intact prophage. An in-house *Parabacteroides* isolate, *P. distasonis* APC 919/143, from a healthy volunteer was found to have an element integrated into the main chromosome with a gene with ~30% amino acid identity to the candidate MCP of the unknown viromic element.

Traditional prophage induction methods were used to try to induce prophage from this isolate, including Mitomycin C, UV light and H_2O_2. These methods were successful at inducing a caudovirus prophage but did not induce the unknown element (data not shown). However, when bile salts were added to the growth media of mid-log growth bacteria, replication of the unknown element was induced at appreciable levels. Filtrate of the bacteria supernatant was collected, and nucleases were added to remove non-encapsidated DNA. Then, the supernatant was spun down an Optiprep buoyant density gradient to enrich for virus like particles. Individual Optiprep fractions were sequenced, and reads were aligned back to the *P. distasonis* APC 919/143 reference genome. The analysis showed a discrete ~22kb element with a peak abundance at the ~27% Optiprep fraction (fraction 3) (Fig. 3.3A). Negative stain microscopy images of peak fractions show spherical particles of 100 - 120nm each (Fig. 3.3B). We thus suggest it is appropriate to refer to this class of elements as viruses and suggest the name "bibiviruses," a portmanteau representing bile-inducible *Bacteroidales*-infecting viruses.

Note to committee: Mass Spec of particles is delayed by COVID-19 lab shutdown

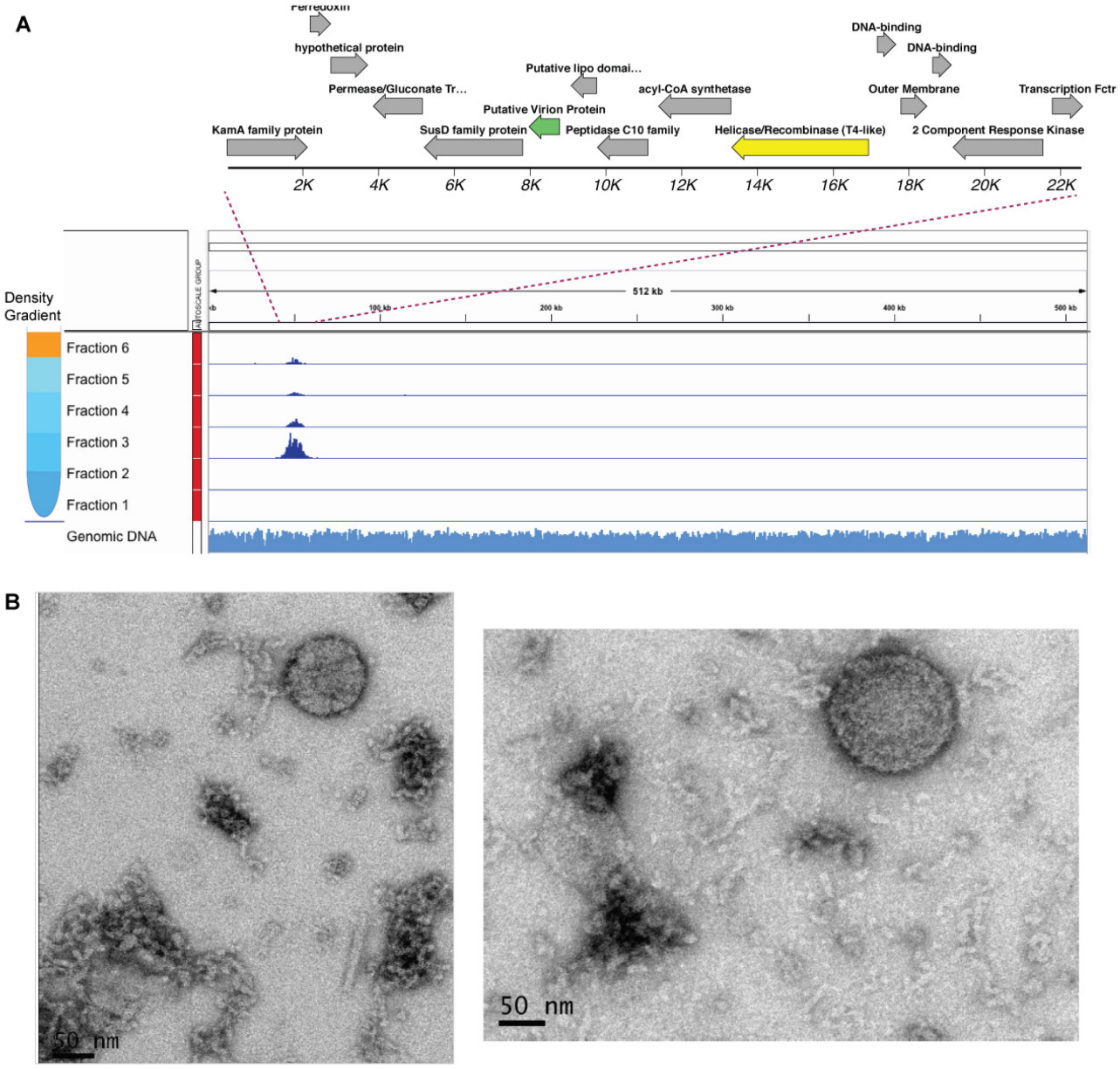


Figure 3.3: Induction and Isolation of a Bibivirus from *Parabacteroides distasonis*

(A) Top: Genome map of *P. distasonis* APC 919/143 bibivirus (accession number ###). Bottom: read alignment to scaffold from *P. distasonis* APC 919/143 reference genome. Sequences were derived from an optiprep gradient fractionation and scaled to Fraction 3. Unenriched genomic DNA shown for comparison. (B) Putative Bibivirus particles from Fraction 3 in (A).

Bibiviruses can be found in viromics datasets from people around the globe

To investigate the diversity and general prevalence of bibiviruses, *de novo* assembly was conducted on 389 publicly available virome deep sequencing samples from several labs. Predicted open reading frames from contigs over 2000 nt were compared to a Hidden Markov Model composed of a sequence alignment of 30 bibivirus MCPs[145]. This identified 530 bibivirus contigs representing 290 unique virus taxa with a cutoff of 95% average nucleotide identity. A phylogenetic tree was drawn based on the putative virion protein (Fig. 3.4). CRISPR spacers were found for 35 bibiviruses, almost entirely from genomes of bacteria from genera *Parabacteroides* and *Prevotella*. This is a relatively frequent spacer acquisition considering many contigs in the bibivirus dataset appear to be partial viral genomes, and the same analysis on a comparable dataset of crAss-like phage contigs (also mostly incomplete) found spacers for 7 of 220 contigs, with 6 of 7 spacer sets derived from *bacteroidales* bacteria. Some virome contigs with CRISPR spacer matches to *Parabacteroides* (Fig 3.4) (e.g., SAMN10290196_a1_ct2444 and others) showed genome organization similar to the bibivirus isolated from *P. distasonis* APC 919/143, while others showed very different gene content and arrangement.

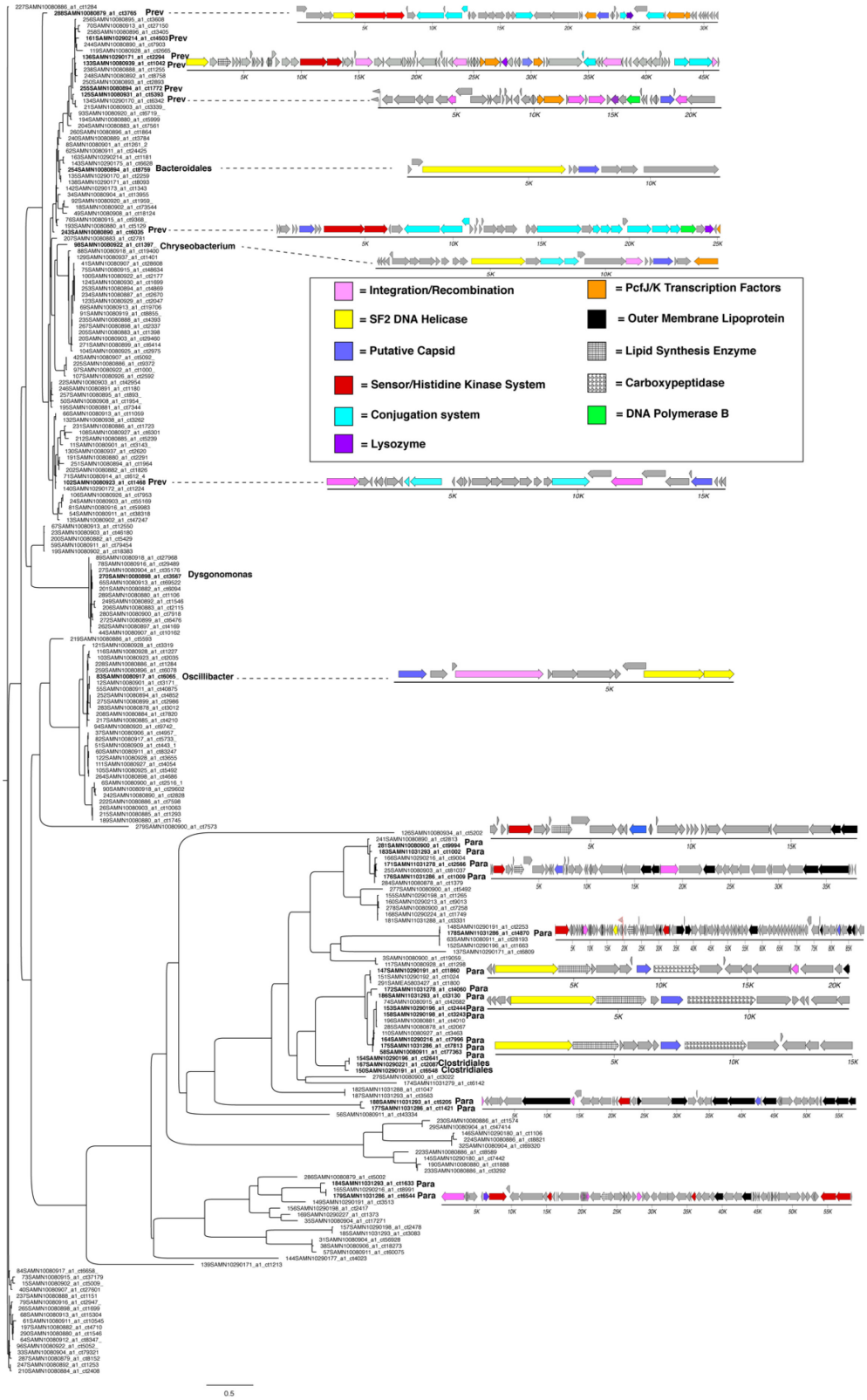


Figure 3.4: Overview of bibivirus phylogeny and host association

Phylogenetic tree from putative capsid proteins and CRISPR-determined host abbreviated in bold. Prev=*Prevotella*, Para=*Parabacteroides*.

Reads from several publicly available gut virome sequencing projects were mapped to the 290 unique bibivirus contigs (Figure 3.5). In a majority of samples, 1-36% of reads mapped to bibivirus contigs. The analysis gives a general impression that bibiviruses may be more abundant in fecal samples from African and South American individuals than European individuals (Figure 3.5), but it is impossible to compare between studies that use somewhat different methods for virus-like particle enrichment and library preparation.

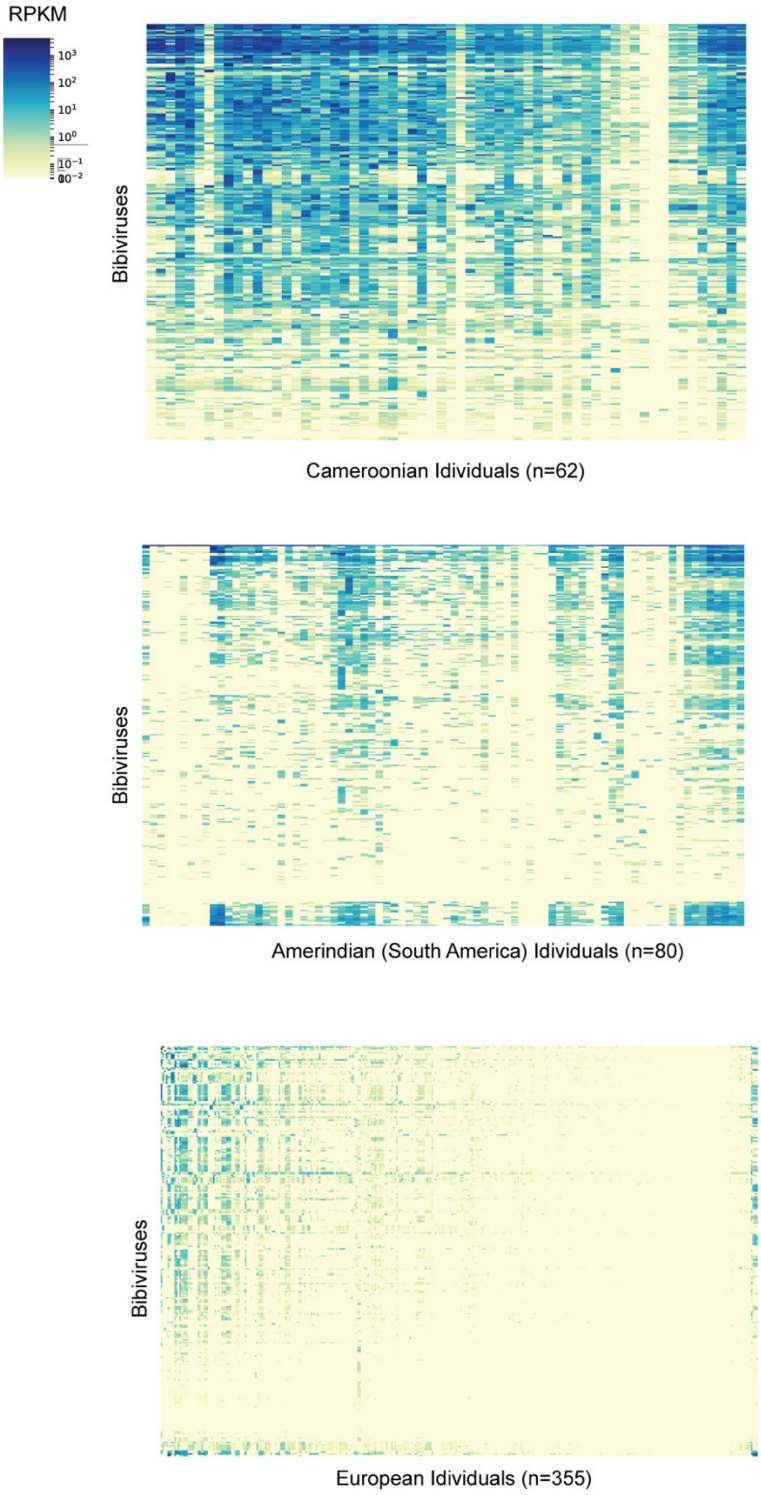


Figure 3.5: Overview of bibivirus prevalence in different regions

Heatmap colors are scaled the same across plots. Cameroon samples derive from PRJNA491626. Amerindian samples derive from PRJNA418044. European samples derive from PRJNA407341, PRJNA385126, and PRJEB29491.

Discussion

Many studies have generated and analyzed metagenomic data to discover new viruses. Most of these studies only consider genes with recognizable sequence similarity to known virus hallmark genes. This is understandable, as it allows high-throughput analysis and confident determination of viral sequences. A few studies have taken alternative approaches to discovering highly divergent virus groups. Notably, Seguritan et al.[12] used artificial neural networks to uncover sequences resembling virion structural genes that were not detectable with primary sequence alignment. Obbard et al.[146] looked at insect anti-viral defense systems, which acquire and synthesize small RNA molecules from viral invaders as an RNAi system, to discover a family of segmented RNA viruses with a highly divergent RNA-dependent RNA polymerase gene and a still-unidentifiable capsid gene. Dutilh et al.[14] used occurrence profiles to find a sequence represented in twelve out of twelve virus-like particle preparations from human stool before identifying virus-like features within the sequence.

Similarly, this study used an occurrence profile on viromic preps of fecal samples from children in The Gambia to detect a high-prevalence group of elements that were initially unrecognizable. The elements, which we named bibiviruses, encode a protein that was predicted by artificial neural network analyses to be a potential MCP. Expression of the candidate MCP in E. coli led to assembly of roughly spherical ~100 nm particles. Further, an endogenous bibivirus of *Parabacteroides distasonis* was induced with bile salts (but not with more traditional prophage-induction methods, such as mitomycin C). Virions recovered from bile salt-treated cultures were slightly over 100 nm in size and were patterned with low knobby protrusions. The appearance of the particles is reminiscent outer membrane vesicles[147]. It is

conceivable that bibivirus particles have been routinely seen in electron microscopy images of virus-like particle preparations and have been incorrectly discounted as membrane vesicles.

Based on CRISPR spacer analysis, it appears that bibiviruses primarily infect bacteria within the order *Bacteroidales*, which is often the most abundant taxon in the human gut. Some *Bacteroidales*, such as certain strains of *Bacteroides fragilis*, can be pathogenic to humans[148]. As bibiviruses often encode outer membrane proteins and other accessory genes, it will be interesting to see if bacteria with endogenous bibivirus prophage have different virulence profiles than those without.

Methods

Identification of an unknown element in virome of GEMS cohort

Stool samples were previously collected from Gambian children, aged from several days to 59 months, by the Global Enteric Multicenter Study initiative (GEMS) in 2008[144]. Hospital staff at specified sites in the Gambia collected case samples from children with moderate to severe diarrhea and who exhibited symptoms including inelastic skin, bloody stool, or the need for IV fluids, whereas caretakers collected control samples from healthy children according to the methods in the paper by Kotloff et al[144]. The participants' guardians gave permission for the participants' involvement in the study, as the participants themselves were underage. GEMS passed the Institutional Review Boards (IRB) from each country where sample collection occurred.

Via GEMS, stool samples were refrigerated and tested within 24 hours of collection. Viral-like particles (VLP) were separated and collected from the sample by centrifugation at

16,000 g for 1 minute. Iterations of sample washing and centrifugation were performed to enrich virus particles in the supernatant. Reverse transcription was performed so RNA (in the form of cDNA) and DNA from the virus-enriched prep were available for sequencing. DNA was extracted with a Qiagen QIamp stool extraction kit and sequenced using Roche 454 Sanger sequencing.

Roche 454 sample reads were subjected to de novo assembly using Newbler 2.9. To create the gene catalog (Fig. 3.1A), 454 reads were assembled into contigs using Newbler 2.9. Prodigal 2.6 was used to predict genes on the contigs, which were then passed through cd-hit 4.6.1 to create a non-redundant gene catalog. The gene nucleotide sequences were clustered with CD-HIT using a minimum overlap of 80% and an identity of 95%. Only the genes contained in clusters of 4 genes or more were kept for further analysis. Using the genes as a reference, reads were aligned using BLASTn with the same parameters used for virus detection. Genes were clustered using the pheatmap and Euclidean distance method.

Computational analysis of candidate major capsid protein

The iVireons Major Capsid Protein model was accessed via https://vdm.sdsu.edu/ivireons/. Positive and negative control proteins for iVireons comparison were identified by key word search using GenBank.

Secondary structure, DeepCoil, and structural homology prediction were conducted the with MPI bioinformatics toolkit[17]: https://toolkit.tuebingen.mpg.de/.

Major capsid protein expression in *E. coli*

Predicted MCP sequences were codon optimized with IDT optimization tool and cloned into plasmids with a T7 polymerase-responsive promoter. Plasmids were transfected into T7 Express

lysY/I^q *E. coli*. Bacteria were grown at 37°C in LB broth until OD600 = 0.5. Flasks were cooled to room temperature, IPTG was added to 1 mM, and cultures were shaken at room temperature for approximately 16 hours. Cells were then pelleted for immediate processing.

Total protein was extracted with a BPER (Pierce) and nuclease solution. Then, virion-sized particles were enriched from the clarified lysate using Optiprep gradient ultracentrifugation. Fractions were analyzed using Coomassie-stained SDS-PAGE gels for presence of a unique band corresponding to the expected size of the candidate MCP. Fractions of interest were analyzed using negative stain electron microscopy.

Induction and enrichment of bibivirus from *Parabacteroides distasonis* APC 919/143

Parabacteroides distasonis APC 919/143 was isolated from the stool of a healthy volunteer, and the genome was sequenced in-house. The strain was grown in Cooked Meat Medium from Hardy (Cat#: K19) under anaerobic conditions. All inductions, including bile salt (Sigma: B8756-10G), were done by growing the bacteria to mid-log phase in liquid broth. Then, the candidate induction agent was added directly to the broth and cultures were incubated overnight for approximately 18 hours. Then, cultures were put on ice for 10 minutes and spun twice at 5000 x g for 10 minutes, with the supernatant being collected for further processing. Benzonase and $MgCl_2$ to 2mM were added to the clarified supernatant and tubes were incubated at 37°C for 30 minutes to digest unencapsidated DNA. The nuclease-digested supernatant was then subjected to filtration through 0.45 µm filters twice to remove large particulate material. Finally, the filtrates were overlayed onto Optiprep step gradients of 21%/27%/33%/39% w/v and spun at 50,000rpm for 3.5 hours in an ultracentrifuge (Beckman: Optima L-90K Ultracentrifuge). Six

equal-volume fractions were drip extracted from the ultracentrifuge tubes by piercing the bottom with a 25-gauge needle.

To disrupt virus particles, 50 µl of a 5x master mix of Tris pH 8 (Invitrogen, final conc. 50 mM), EDTA (Invitrogen, final conc. 25 mM), SDS (Invitrogen, final conc. 0.5%), Proteinase K (Invitrogen, final conc. 0.5%), DTT (Invitrogen, final conc. 10 mM) was added to 200 µl of each fraction and mixed by trituration. Samples were heated at 50°C for 15 minutes. Then, proteinase K was inactivated for 10 minutes at 72°C.

DNA was extracted with NEB Monarch DNA Gel Extraction Kit. Sequencing was conducted using Illumina MiSeq instrumentation at the NCI sequencing core.

Reads were trimmed and quality controlled with fastp and reads were aligned to the *Parabacteroides distasonis* APC 919/143 genome with Bowtie2 and alignment plots were visualized with Integrative Genome Viewer.

Contributions

Michael J Tisza: Conceptualization, Resources, Data curation, Software, Formal analysis, Validation, Investigation, Visualization, Methodology, Project administration. (All figures except Figure 3.1A)

Mathieu Almeida: Conceptualization, Data curation, Methodology, Project administration

Laura Tanase: Conceptualization, Data curation, Methodology, Investigation, Visualization. (Figure 3.1A)

GEMS Consortium: Sample collection and administration

David Wang: Conceptualization, Project administration

Colin Stein: Conceptualization, Project administration

Christopher B Buck: Conceptualization, Project administration

4 Cenote-Taker2 Democratizes Virus Discovery and Sequence Annotation

Abstract

Viruses, despite their great abundance and significance in biological systems, remain largely mysterious. Indeed, the vast majority of the perhaps hundreds of millions of viral species on the planet remain undiscovered. Many viruses deposited in central databases like GenBank and RefSeq are littered with genes annotated as “hypothetical protein” or the equivalent. Cenote-Taker2, a virus discovery and annotation tool available on command line and with a graphical user interface with free high-performance computation access, utilizes highly sensitive models of hallmark virus genes to discover familiar or divergent viral sequences from user-input contigs. Additionally, Cenote-Taker2 uses a flexible set of modules to automatically annotate the sequence features of contigs, providing more gene information than comparable tools. The outputs include readable and interactive genome maps, run summary tables, and files that can be directly submitted to GenBank. We expect Cenote-Taker2 to facilitate virus discovery, annotation, and expansion of the known virome.

Introduction

Virus hunters have a challenging signal-to-noise problem to consider. For example, animals and bacteria share homologous genes with more amino acid identity than even the

most-conserved genes in some virus families (for example, GenBank sequences: polyomavirus Large T antigen [NP_043127.1 vs. YP_009110677.1] and 60S ribosomal protein L23 [CUU95522.1 vs. NP_000969.1]). Further, there are no universal genes found in all viral genomes that could be used to probe complex datasets for viruses in the same way cellular life can be detected through PCR targeting ribosomal sequences. Finally, at least hundreds of millions of virus species are likely to exist on earth[4], but sequences for only tens of thousands of virus species are deposited in the central GenBank virus database. Fewer than 10,000 virus species exist in the authoritative RefSeq database. Several tools have been developed to detect virus sequences in complex datasets. Strategies include detection of hallmark genes conserved within known virus families[46], detection of short nucleotide sequences believed to be enriched in viruses (deepvirsorter, arXiv:1806.07810), or the ratio of genes common to virus genomes to genes common to non-viral sequences[149]. However, each of these tools has pitfalls that can lead to false positives or false negatives and some tools are limited by minimum sequence length or are only geared to detect a limited range of virus families.

Beyond discovery and detection, *de novo* annotation of contigs representing viruses presents a number of challenges. To list a few, determination of genome topology, accurate calling of open reading frames, determining the virus-chromosome junction in integrated proviruses, resolution of taxonomy, and, especially, accurate annotation of highly divergent homologs of known genes all present technical hurdles. An even deeper problem is the misannotation of genes in GenBank entries, including the authoritative RefSeq virus database.

This manuscript presents version 2.0 of our Cenote-Taker pipeline, which was originally geared toward annotation of viruses with circular DNA genomes[150]. This flexible tool enables

the discovery and annotation of all virus classes and is available for use on Linux terminal and as a graphical user interface (GUI) with free compute cluster usage on CyVerse. Cenote-Taker 2 outpaces comparable tools in gene annotation by providing information for a higher percentage of genes with a higher degree of accuracy, especially for virus hallmark genes. Additionally, Cenote-Taker 2 performs better in discovery of viral sequences in complex datasets, with lower false positive and false negative rates than other available tools.

Results

Cenote-Taker2 process overview

A basic run of Cenote-Taker2 requires only a file of contigs and a file with metadata that enables easy submission of annotated sequences to GenBank. A number of optional settings allow users to customize the pipeline. In-depth discussion of the options can be found at the Cenote-Taker2 GitHub repo and wiki. Figure 4.1 provides a visual of Cenote-Taker2 workflow. First, Cenote-Taker2 analyzes contigs above a user-determined length and attempts to detect two possible hallmarks of some types of virus genomes - circularity or the presence of inverted terminal repeats (ITRs). Circles are rotated to a position where no open reading frames (ORFs) overlap the wrap-point. An optional step calculates the read depth of each contig. All input contigs are then scanned for the presence of a curated set of hallmark genes specific to known virus families. For users who wish to use Cenote-Taker2 to discover novel viruses in complex datasets, only contigs containing the minimum user-determined number of virus hallmark genes are kept for further analysis. For users who have indicated that their input contigs are pre-filtered to only contain viral contigs, all contigs are kept and annotated.

Many viral genomes are integrated into bacterial chromosomes. In datasets likely to contain cellular chromosomes, a single contig might thus contain a virus sequence flanked on one or both sides by a cellular sequence. Users can choose to let Cenote-Taker2 prune flanking cellular sequences and generate a genome map for the viral portion of the contig. Another optional module conducted at this step queries a nucleotide database, such as GenBank nt, with BLASTN[58], and sorts contigs with at least 90% average nucleotide identity to an entry in the database.

Next, candidate tRNA genes are detected and annotated[151]. A tentative taxonomy of each contig is then guessed using BLASTX against a custom database containing Refseq virus and plasmid sequences from GenBank. This taxonomy is used to determine the best ORF-caller (PHANOTATE for putative bacteriophage[152], Prodigal for other viruses[153]). ORFs are then functionally annotated based on validated datasets using tools for detection of remote homologs (i.e. hmmscan[145], RPS-BLAST[154], HHblits/HHsearch[74]). In these steps, only carefully curated databases (CDD, PFam, PDB, Cenote-Taker2 hallmark database) are queried to avoid propagation of mis-annotated sequences in databases such as GenBank nr. All annotation, taxonomy information, and metadata are combined to generate several outputs. Each contig is represented as an interactive genome map file (.gbf), a gene feature file (.gtf), and a file that can be used for GenBank submission (.sqn). Finally, key information on all annotated contigs is provided in a single summary table (.tsv).

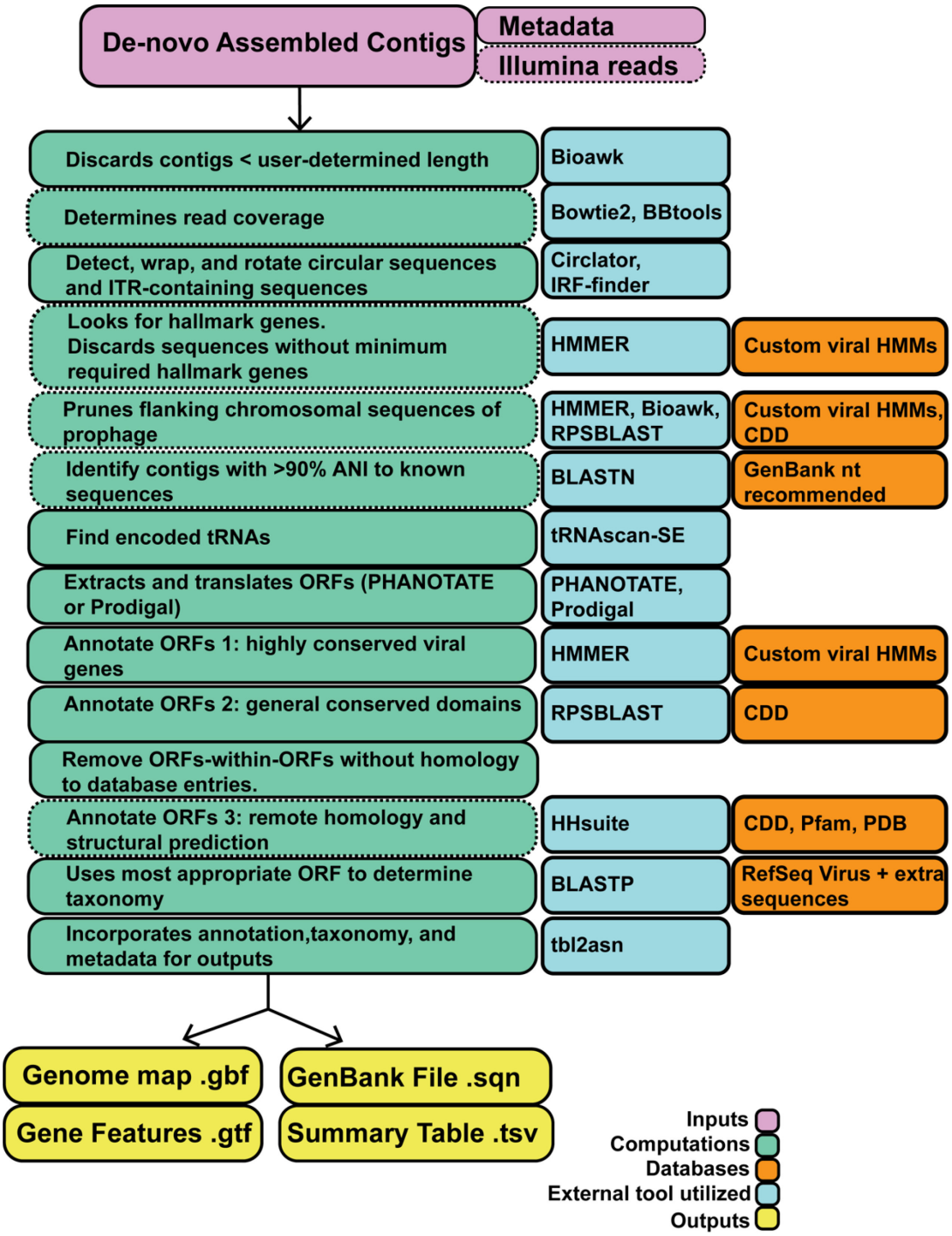


Figure 4.1: Schematic of Cenote-Taker2 Processes

Visual representation of Cenote-Taker2 virome analysis. Boxes with hashed lines represent optional inputs or processes.

Cross-comparison of Virus Annotation Modules

VirSorter and VIGA (https://github.com/EGTortuero/viga) are virus genome annotation tools comparable to Cenote Taker 2. A comparison of features is shown in Table 1. Two arbitrarily chosen “challenging” viral genomes were chosen as initial case studies for comparing the three pipelines (Fig. 4.2). For the newly described Yaravirus (doi: 10.1101/2020.01.28.923185), only Cenote-Taker2 could discern an annotation for any genes, with the major capsid protein (MCP), packaging ATPase, and replicative helicase all being annotated. For a previously undiscovered crAss-like phage, Cenote-Taker2 again annotates more genes than the other tools. VirSorter maps are not shown as they VirSorter didn't improve upon annotation of any gene as compared to VIGA or Cenote-Taker2.

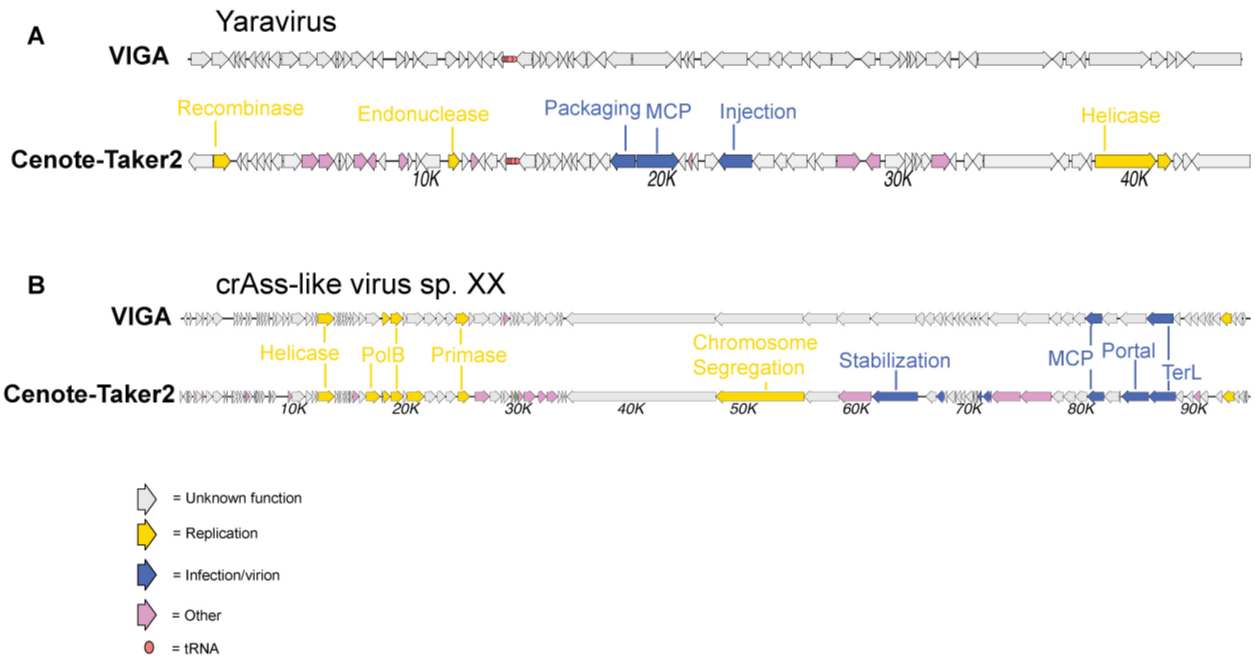


Figure 4.2: Comparison of genome maps from VIGA and Cenote-Taker

Annotation pipelines were run with optimal options. (A) Yaravirus is a newly reported megavirus-like virus. (B) crAss-Like virus sp. XX is a tailed phage discovered in a human gut metagenome (SRR6128032).

Table 1: comparison of annotation pipelines

	Annotates ORFs	Finds tRNAs	GenBank submittable	Accepts pre-filtered contigs	Viral Discovery	Takes Metadata	Computes read coverage	Determines circularity	Determines ITRs	Determines Ori	Determines DTRs	Finds matches in GenBank	Makes taxonomy call	Prunes Prophage	Does Plasmids	Summary Table
Cenote-Taker2	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	Y	Y	Y	Y	Y
VIGA	Y	Y	N[1]	Y	N	Y[2]	N	N	Y[3]	Y	Y	N	N	N	N	N
VirSorter	Y	N	Y	N	Y	N	N	Y	N	N	N	N	N	Y	N	Y

[1] It gets you most of the way there
[2] A metadata and taxonomy file must be provided for each sequence
[3] Finds inverted repeats but doesn't determine if they're terminal

Generation of Virus Hallmark Gene Hidden Markov Models

Proteins from various public databases, including virus RefSeq and assemblies of virus-enriched datasets, were clustered using EFI-EST[76] (pairwise E value cutoff < 1e-10). Clusters were visualized in Cytoscape[77], and multi-lobed clusters were manually divided or discarded. Each cluster of three or more proteins was aligned using MAFFT[105]. The resulting multiple sequence alignments (MSAs) were used as queries for HHsearch structural prediction and distant homology detection searches against PDB, CDD, and Pfam. MSAs without confident alignment to any models in this search were again used as queries for HHblits against UniProt. Each MSA with a hit in either search was used to generate a Hidden Markov Model (HMM) using Hmmer. All HMMs were kept for further consideration if the name corresponded to a possible viral hallmark gene (e.g. major capsid protein). All Putative Hallmark HMMs were tested for specificity with a two-step validation by first querying against a negative control database, namely, human proteins from RefSeq, using Hmmer. Second, protein sequences from a variety of human and environmental metagenome-derived contigs were queried against the database of the remaining HMMs using Hmmer and any proteins with "hits" to the database were then cross-queried using HHsearch against PDB, CDD, and Pfam. If these "hit" proteins had similarity to models in these databases that were qualitatively different from the identity of the putative Hallmark HMM, the Hallmark HMM was discarded. Some replication-related Hallmark HMMs were later removed because they were similar to genes typically found on plasmids or conjugative transposons. Finally, HMMs from pVOGs (http://dmk-brain.ecn.uiowa.edu/pVOGs/) and PFAM were considered and validated in the same manner.

Comparison of Virus Discovery Module

Cenote-Taker2 was compared to three leading virus discovery tools, each with its own method for detecting viral sequences. Like Cenote-Taker2, VirSorter uses a virus hallmark gene detection approach. One limitation is that it is only designed to detect bacteriophages. DeepVirFinder uses a machine learning approach to find short nucleotide motifs common in viral sequences. An additional pipeline, Non-Targeted, compares predicted protein sequences encoded by a contig to a curated set of known viral proteins. A limitation of Non-Targeted is that it only considers contigs greater than 5 kb, while the other tools have no minimum length. The main types of datasets that might be searched are contigs derived from DNA samples enriched for viral sequences (DNA virome), RNA samples enriched for viral sequences (RNA virome), DNA from unenriched samples (genomes and metagenomes), RNA from unenriched samples (transcriptomes and metatranscriptomes). An additional parameter that can aid detection of ssDNA viruses is use of a second strand synthesis step, often from multiple displacement amplification. Five datasets, one of each type, were assembled, and contigs greater than 1000 nucleotides were analyzed. Cenote-Taker outperformed all other discovery tools for finding contigs with viral structural or replication genes for each type of dataset (Fig. 4.3, Fig. 4.4, Fig. 4.5, Fig. 4.6, Fig. 4.7). In particular, Cenote-Taker2 had the highest number of total hits from a dataset consisting primarily of ssDNA viruses (Fig. 4.4) RNA virome dataset (Fig. 4.6). Although DeepVirFinder produced more hits for some datasets, it is unclear whether the unique DeepVirFinder hits are really viral.

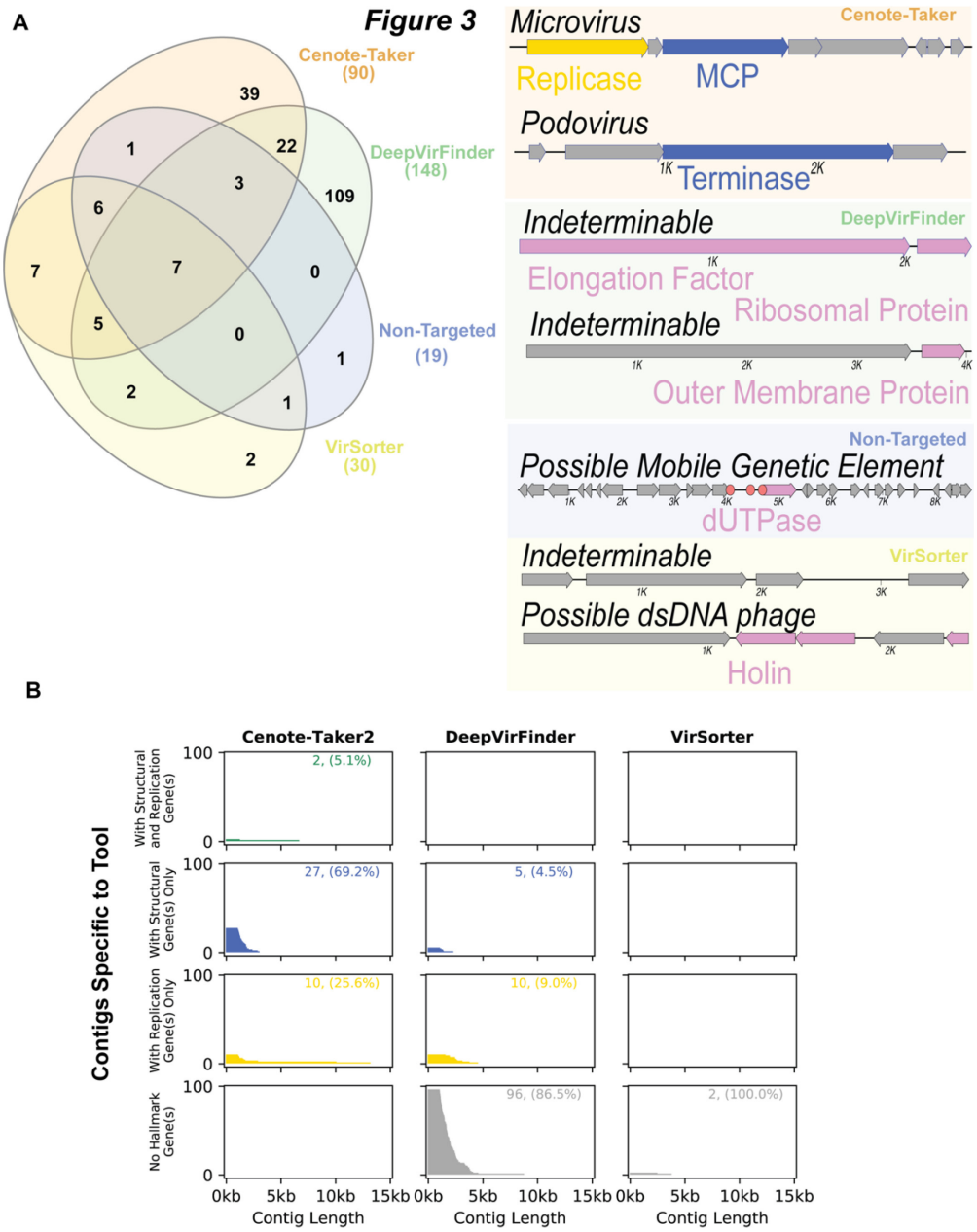
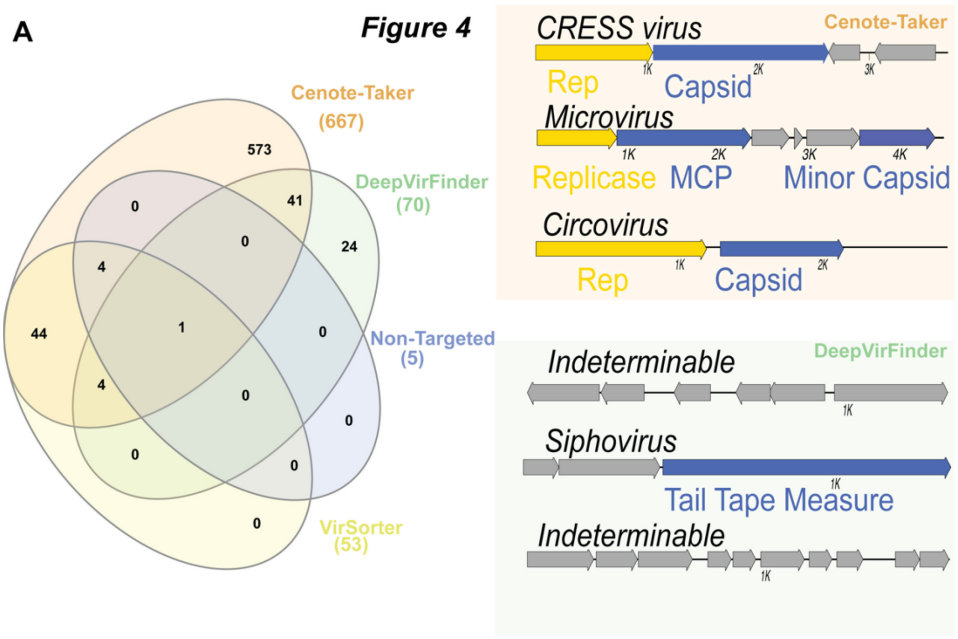


Figure 4.3: Comparison of virus discovery tools for DNA virome from human stool

A dataset for human stool enriched for nuclease-resistant DNA in virus-sized particles (SRR6128021) was assembled into contigs. Contigs > 1000 nucleotides were then analyzed using four virus detection/discovery pipelines. (A) Comparison of the overlap of contigs the various pipelines designated as viral. Maps of representative examples of contigs the indicated pipeline uniquely called as viral are shown on the right side of the panel. (B) Contig attribute chart showing only contigs called uniquely by Cenote-Taker2, DeepVirFinder, and VirSorter. Each contig is displayed as a horizontal line with the line length corresponding to the sequence length (x-axis).



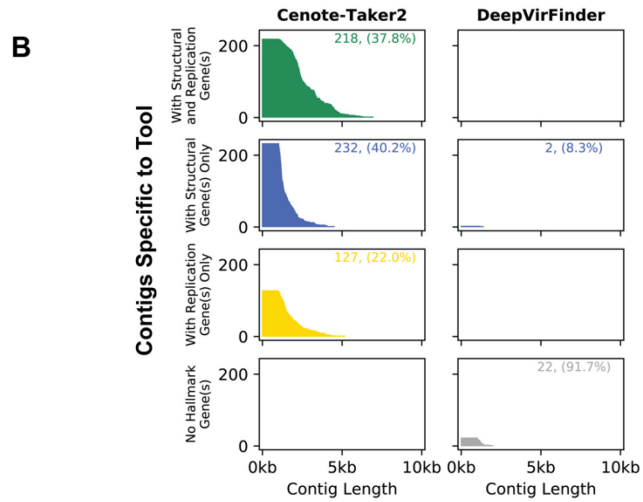
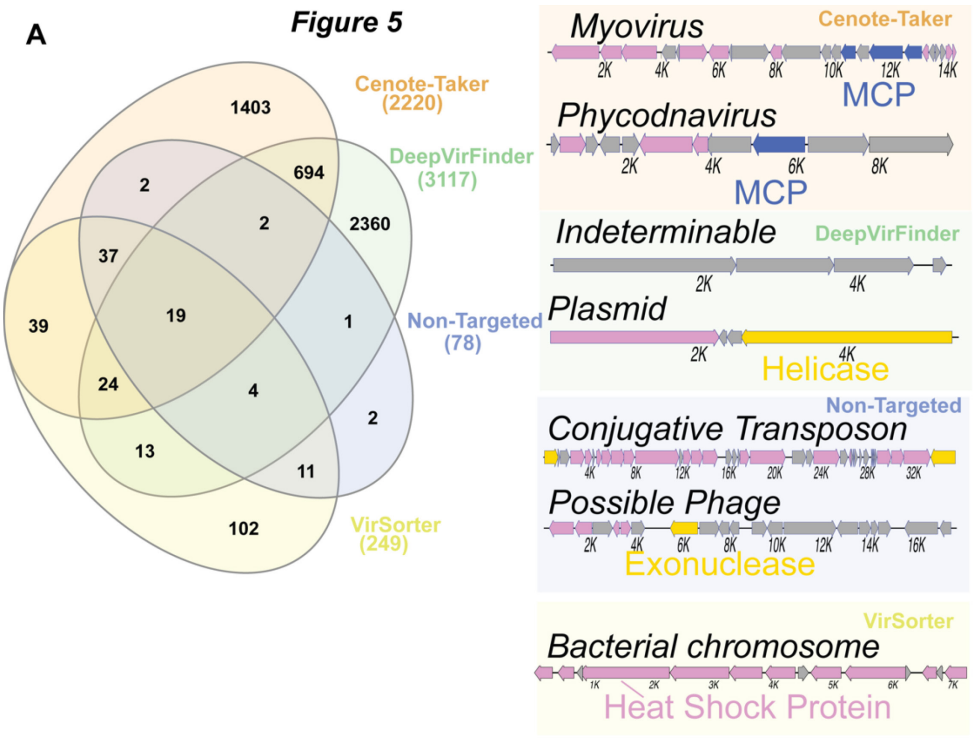


Figure 4.4: Comparison of virus discovery tools for ssDNA virome from wastewater plant

This dataset (SRR3580070) consists of contigs greater than 1000 nucleotides from wastewater enriched for virus-like particles followed by rolling circle amplification with DNA sequencing. (A) Comparison of the overlap of four different virus discovery/detection tools. (B) Contig attribute chart showing only contigs called uniquely by Cenote-Taker2, DeepVirFinder, and VirSorter. Each contig is displayed as a horizontal line with the line length corresponding to the sequence length.



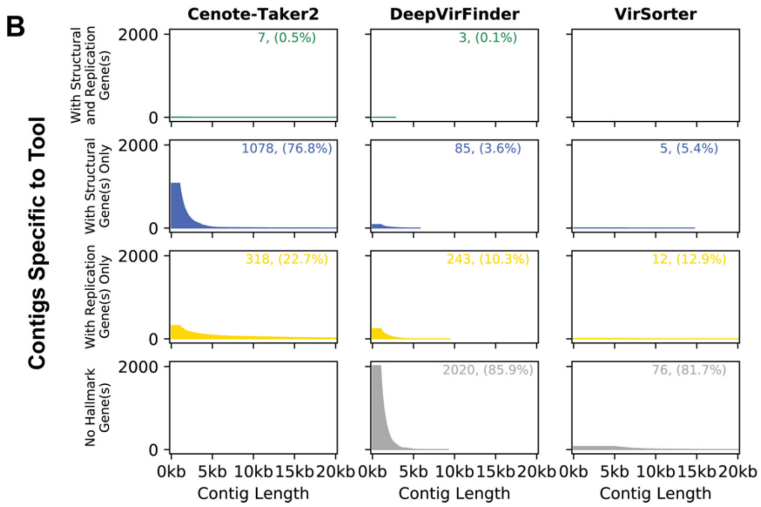


Figure 4.5 : Comparison of virus discovery tools for DNA metagenome from Amazon River water

This dataset (ERR2338392) consists of contigs greater than 1000 nucleotides from water from the Amazon River with DNA sequencing. (A) Comparison of the overlap of four different virus discovery/detection tools. (B) Contig attribute chart showing only contigs called uniquely by Cenote-Taker2, DeepVirFinder, and VirSorter. Each contig is displayed as a horizontal line with the line length corresponding to the sequence length.

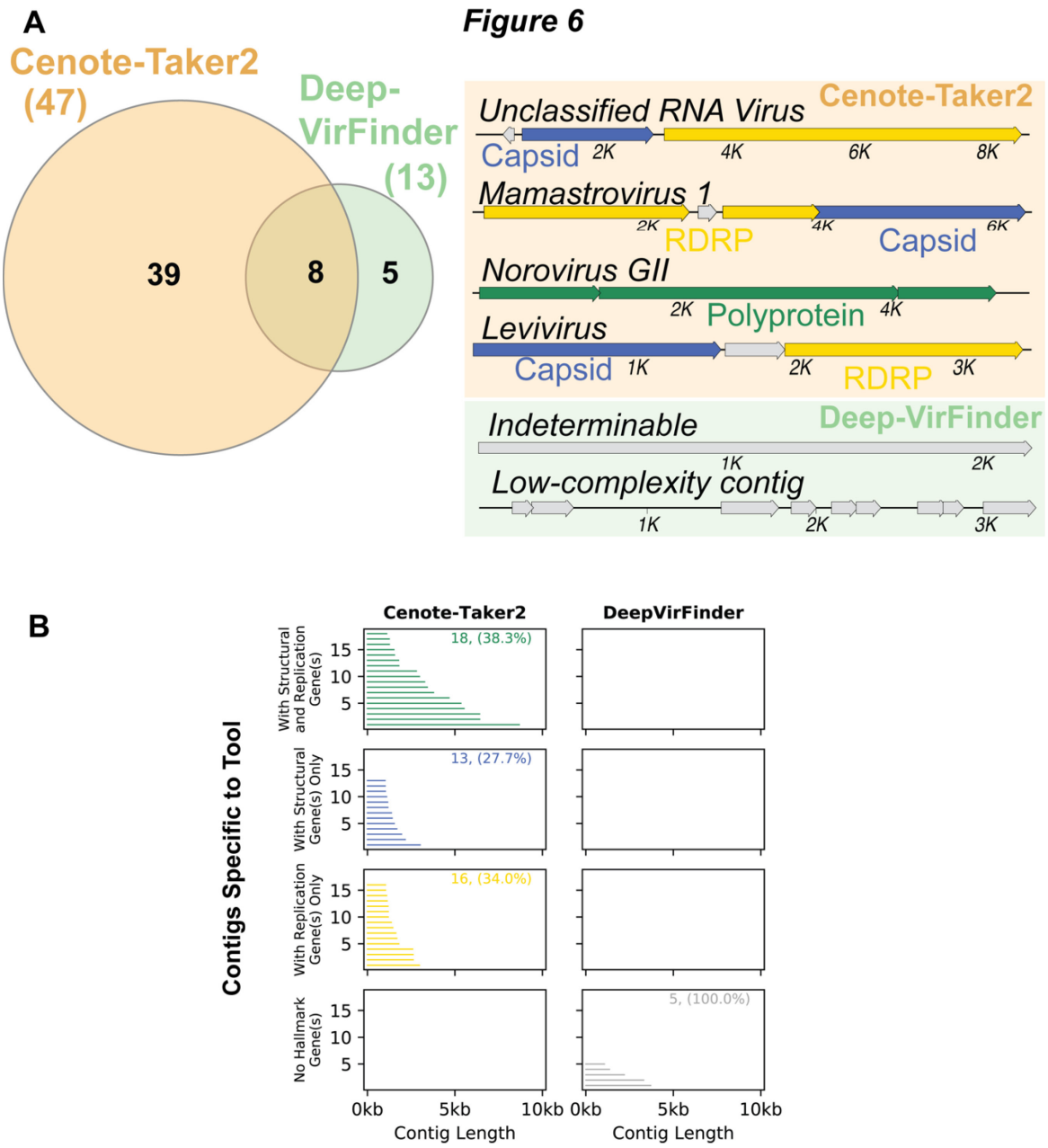
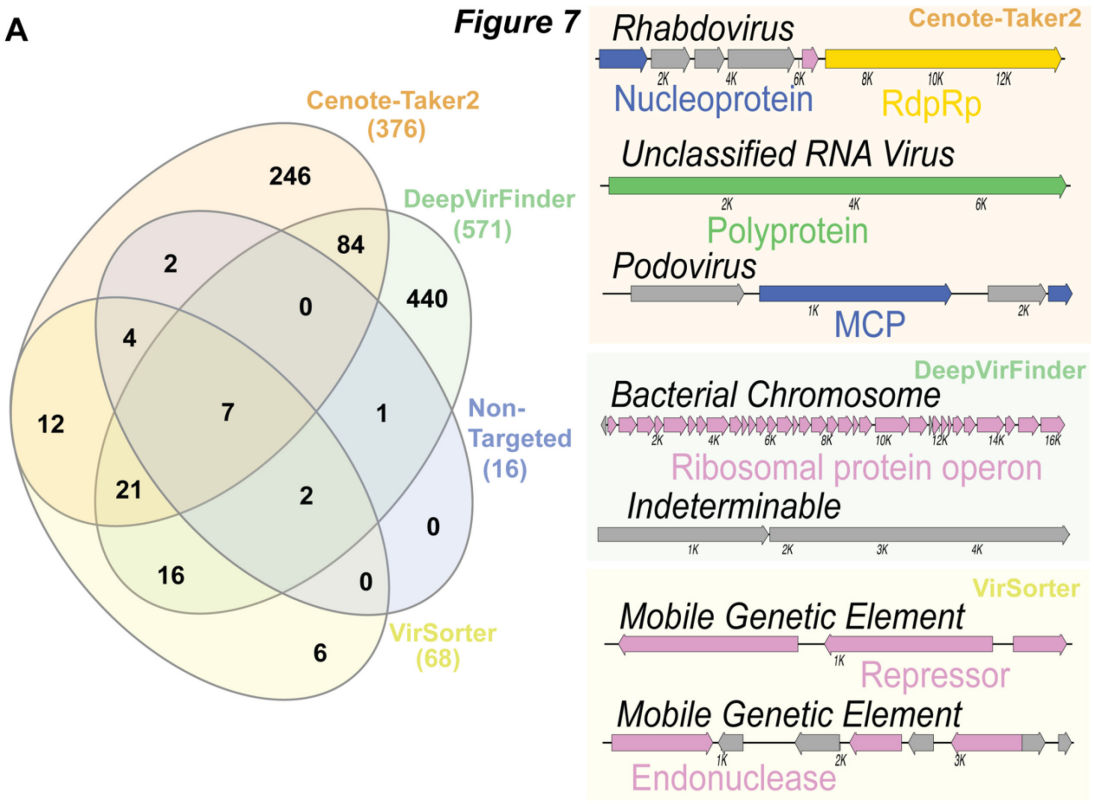


Figure 4.6: Comparison of virus discovery tools for RNA virome from sewage

This dataset (ERR3201762) consists of contigs greater than 1000 nucleotides from sewage enriched for virus-like particles with RNA sequencing. (A) Comparison of the overlap of four different virus discovery/detection tools. (B) Contig attribute chart showing only contigs called uniquely by Cenote-Taker2, DeepVirFinder, and VirSorter. Each contig is displayed as a horizontal line with the line length corresponding to the sequence length.



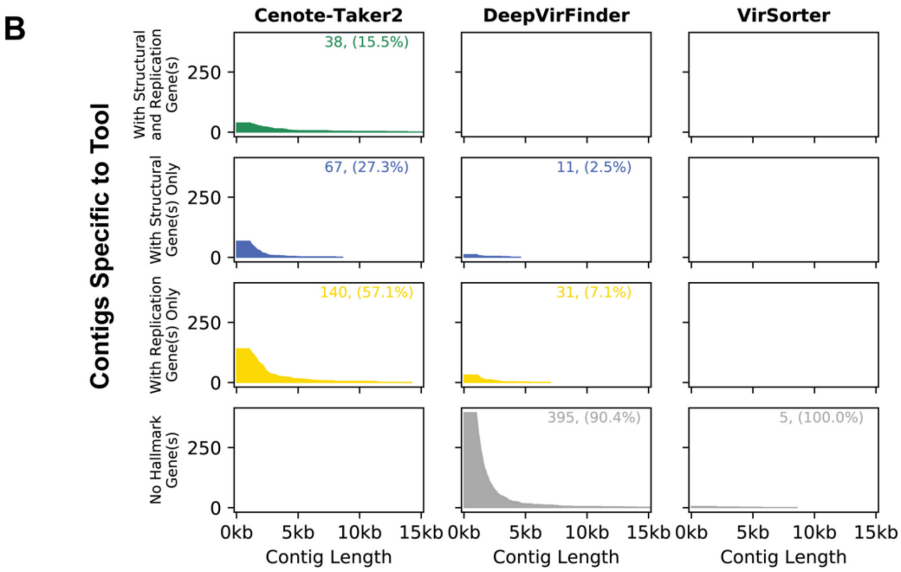


Figure 4.7: Comparison of virus discovery tools for RNA metatranscriptome from Tasmanian devil stool

This dataset (SRR8048121) consists of contigs greater than 1000 nucleotides from *Sarcophilus harrisii* stool with RNA sequencing. (A) Comparison of the overlap of four different virus discovery/detection tools. (B) Contig attribute chart showing only contigs called uniquely by Cenote-Taker2, DeepVirFinder, and VirSorter. Each contig is displayed as a horizontal line with the line length corresponding to the sequence length.

Prophage Pruning Module

When this option is selected, linear contigs will get ORF calls via Prodigal, then ORFs will be iteratively searched with 1) HMMSCAN of the custom virus hallmark gene database, 2) HMMSCAN of the custom common virus gene database, and 3) RPS-BLAST of CDD. Each gene is then considered to be 1) a virus hallmark gene, 2) a common viral gene (hit in the custom common virus gene database or hit in CDD of a domain found in 10 or more RefSeq Caudovirales genomes or hit in CDD with "PHA0" prefix), 3) a common chromosomal gene (all other CDD hits), or 4) an unknown gene (no hits in any of these databases). Then, based on the coordinates of the ORFs and their categorization, each nucleotide position in the contig is scored. Bases within virus hallmark or common viral genes are scored as 10. Bases within unknown genes are scored as 5 (bacteriophage are enriched for these genes). Bases in intergenic regions are scored as 0. Finally, bases within known bacterial genes are scored as -3. The sum of 5 kb windows tiled every 50 bases is calculated then scores are smoothed based on the scores of adjacent windows. Contig segments of 1 or more consecutive windows with a positive score are resolved, and segments containing virus hallmark genes are designated as viruses or virus fragments. Example prophage calls and virus genome maps from a *Bacteroides xylanisolvens* genome (GenBank: ASM654696v1) are shown in Figure 4.8.

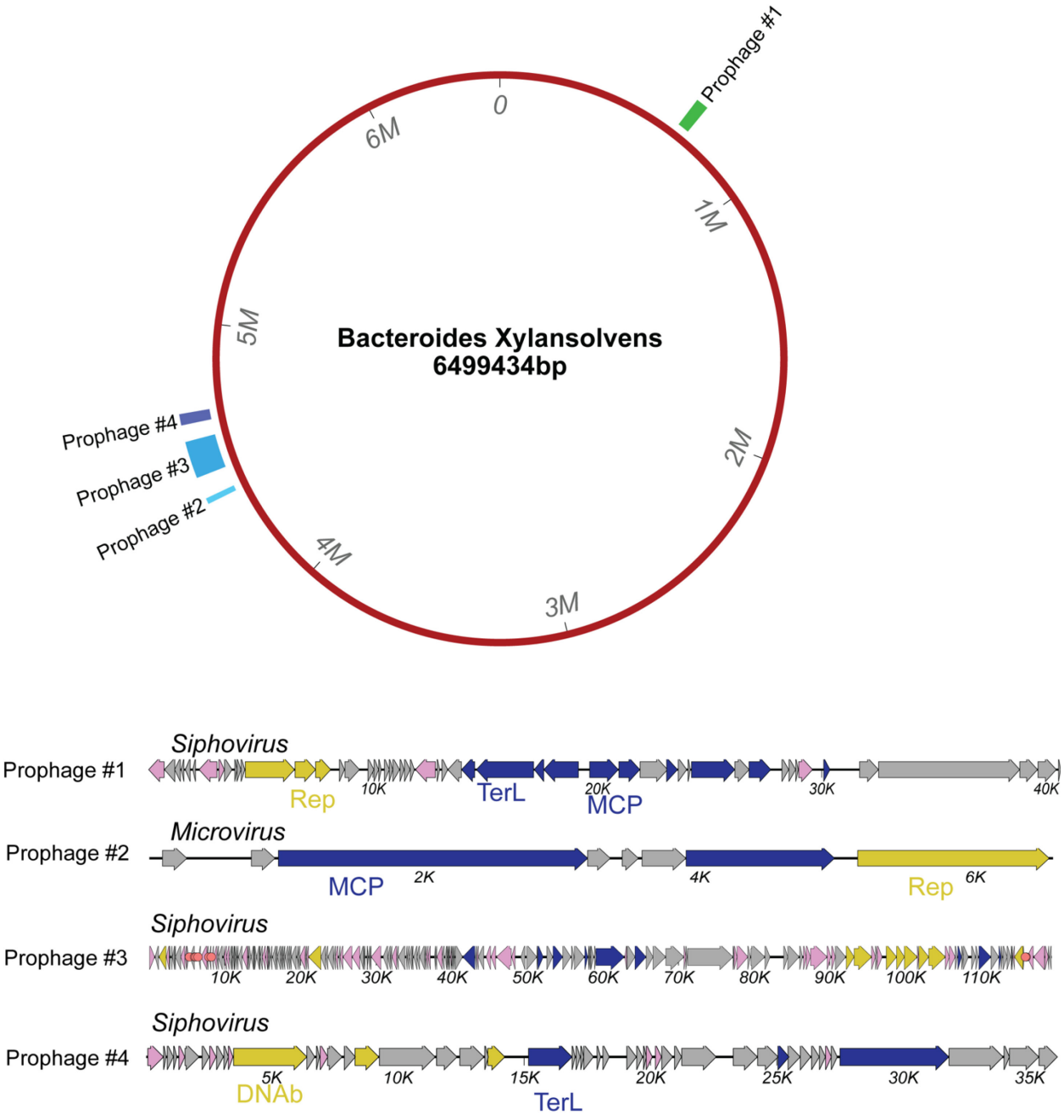


Figure 4.8: Cenote-Taker2 analysis of Bacteroides xylansolvens genome (ASM654696v1)
The circular map represents the *B. xylansolvens* genome annotated with coordinates of the prophage called with Cenote-Taker2. A map of each prophage is shown.

Discussion

We expect Cenote-Taker2 will prove useful to all microbiologists interested in analyzing viruses in their sequencing data. Both the ability to easily discover viruses as well as the ability to confirm putative viruses by visualizing genome maps, should allow scientists to confidently move from largely unintelligible contig .fasta sequences to meaningful analysis of their data. Furthermore, because Cenote-Taker2 eases submission of annotated genomes to GenBank, even those who don't use Cenote-Taker2 will indirectly benefit by having a larger, better-annotated, central sequence database.

Two annotation challenges of viral coding regions that are not resolved with Cenote-Taker2 are frame-shifting, which is documented in some RNA viruses and dsDNA bacteriophage, and intron-containing genes, which occur in many eukaryotic viruses. The authors are not aware of a way to automate resolution of these features.

Cenote-Taker2 outperforms other virus discovery pipelines for a variety of reasons. While both VirSorter and Non-Targeted employ hidden Markov models of viral genes to some extent, it's likely that the models developed for Cenote-Taker2 represent more of the diversity of viral hallmark genes. Further, since contigs are penalized by Non-Targeted if they contain common chromosomal genes, contigs representing a virus sequence flanked by a chromosomal sequence might be discarded instead of pruned. DeepVirFinder uses a fundamentally different approach, looking for nucleotide k-mers of different lengths to determine if a contig is a virus. Two reasons why this approach can fall short are: (1) nucleotide sequence space may be unable

to adequately capture the vast diversity of virus genomes (2) DeepVirFinder was trained on "virome" assemblies. Physical enrichment of virus-like particles is notoriously difficult, so some training datasets may have been contaminated with cellular chromosomes. Moreover, it is known that some sequences, even in very clean virus-like particle preparations, are not viruses but mobile genetic elements that parasitize viral capsid machinery[143].

While there are likely new "types" of yet-to-be discovered viruses encoding novel capsid and replication genes, Cenote-Taker2 can readily be updated to include new hallmark gene models. For example, a new model was made for the replication gene of the proposed new family Quenyaviruses[146].

Methods

Cenote-Taker2 Code

Cenote-Taker2 was written in Bash, Perl and Python. All scripts can be accessed on GitHub. In-depth discussion of use-cases and considerations can be found on the Wiki . Installation uses Conda to manage packages[155]. BLAST and Hmmer databases developed for this tool can be found on Zenodo.

Annotations of Challenging Viral Genomes

Cenote-Taker2 was fed these genomes with default settings except " --hhsuite_tool hhsearch" was used. VIGA default settings are particularly stringent, therefore a several custom options were used to improve annotation: " --diamondevalue 1e-04 --diamondidthr 30 --hmmeridthr 30." Genome maps in all figures were visualized with MacVector 16.

Virus Discovery Comparison

Reads from each sequencing run were trimmed with Fastp, assembled with Megahit, and scaffolded with SOAPdenovo2. Cenote-Taker2 hallmark gene Hmmer database (updated April 21st, 2020) was used with viral hits having one or more detected hallmark gene. The Cenote-Taker2 script requires 1-e08 p value as a minimum threshold for structural genes and 1e-15 for replication genes. VirSorter was used with "virome" settings and categories 1,2,4, and 5 were kept. DeepVirFinder was used with the default training set and p value threshold of 0.005. Non-Targeted Pipeline was used with default settings.

Hallmark Gene Calls

Putative viral contigs from all sources were annotated with Cenote-Taker, using RPS-BLAST with the CDD database and HHsearch with CDD, PFam, and PDB. All annotated genes were scanned for names of viral replication or structural genes and domains.

Contributions

Michael J Tisza: Conceptualization, Resources, Data curation, Software, Formal analysis, Validation, Investigation, Visualization, Methodology, Project administration

Anna Belford: Conceptualization, Software

Ben Bolduc: Software

Matthew B Sullivan: Conceptualization, Software

Igor Tolstoy: Methodology, Data curation

Christopher B Buck: Conceptualization, Data curation, Methodology

5 The Human Virome: Over Eighty Thousand Distinct Viruses and Specific Associations with Chronic Diseases

Abstract

While scientists have made remarkable strides in microbiome research, the viral component of the microbiome has generally presented a more challenging target than the bacteriome. This is despite the fact that thousands of shotgun sequencing runs from human metagenomic samples exist in public databases, and all of them encompass large amounts of viral sequences. The lack of a definitive database for human-associated viruses and insufficient methods to confidently identify divergent viruses in metagenomic data has stymied efforts to characterize virus sequences in a comprehensive way. In this study, a high specificity and sensitivity bioinformatic tool, Cenote-Taker2, was applied to thousands of human metagenome datasets, uncovering over 80,000 unique complete or high quality viral sequences. Publicly available case-control studies were reanalyzed, and strong disease associations were found for over a thousand specific viruses.

Introduction

The human virome is the collection of viruses that live in and on people and their genomes. This includes viruses that directly infect human cells[35], but mostly consists of viruses infecting resident bacteria, i.e. phages[29]. While the large majority of microbiome studies have focused on the bacteriome, revealing numerous important functions for bacteria in human physiology[124],

information about the human virome has lagged. However, a number of important studies have made inroads into characterizing the virome[27,156-159].

Just as human-tropic viruses can have dramatic effects of us, phages are able to dramatically alter bacterial physiology and regulate host population size. A variety of evolutionary dynamics can be at play in the phage/bacterium arena, including Red Queen[125], arms-race[160], and Piggy-back the Winner[161] relationships, to name just a few. In the gut, many phages enter a lysogenic or latent state and are retained as integrated or episomal prophages within the host bacterium. In some instances, the prophage can buttress host fitness (at least temporarily) rather than destroy the host cell. Prophages often contain genes that can dramatically alter the phenotype of the bacteria, such as toxins[133], virulence factors[24], antibiotic resistance genes[134], photosystem components[135], other auxiliary metabolic genes[136], and CRISPR-Cas systems[137], along with countless genes of unknown function.

There have been a few documented cases where phages have been shown to be mechanistically involved in increased bacterial virulence[133] or resistance to antibiotics[162], demonstrating the complex roles phage can play in human health. In addition, several studies have conducted massively parallel sequencing on virus-like particles derived from human stool samples, finding differential abundance of some phages in disease conditions[26-28,163]. A major issue encountered by these studies is that there is not a concise database of annotated virus genome sequences and *de novo* prediction of virus sequences from metagenomic assemblies is a daunting challenge[29]. Therefore, most of the sequence data from these studies remains unevaluated. For example, one study of twelve individuals was able to recruit over 80% of virus-

like-particle-derived reads to potential viral contigs, but most of these contigs were unclassified and a large majority were incomplete[125].

The current study sought to overcome the traditional challenges of sparse viral databases and detection of highly divergent viral sequences by using Cenote-Taker2, a virus discovery and annotation tool. The pipeline was applied to sequencing data from nearly 6000 human-associated metagenome samples. Strict criteria identified over 180,000 viral contigs representing 83,681 unique viral taxa. This curated database allowed read-alignment-based abundance profiling of the virome in human metagenomic datasets, and several case-control studies were reanalyzed to find significant associations between chronic diseases and the presence or absence of specific virus species.

Results

Characteristics of the Human Virome

Read data was downloaded from NCBI's Sequence Read Archive (SRA), from the Human Microbiome Project[30], and from several other bioprojects involving deep sequencing of human metagenomic samples. A subset of the projects performed enrichment for viral sequences. Almost all of the projects pursued DNA sequencing, but a small number of metatranscriptomic RNAseq samples were analyzed for RNA viruses[164]. Read data were binned and assembled by biosample rather than by individual run to combine read sets from the same individual. A total of 5996 samples were analyzed. Assemblies were conducted using Megahit[165].

This study aimed to keep only high-quality long viral contigs or complete viral genomes. Cenote-Taker2 (Chapter 4) was used to check contigs of >1500 nt for two common end-features of complete viral genomes: circularity or inverted terminal repeats (ITRs). Circular sequences

>1500 nt with at least one viral hallmark gene, ITR-containing contigs >4000nt with at least one viral hallmark gene, and linear (no discernable end features) contigs >12,000 nt with two or more viral hallmark genes were kept as putative viruses. Since phages are sometimes integrated into bacterial chromosomes, each linear contig was pruned with the Cenote-Taker2 prophage pruning module to remove flanking chromosomal sequences. This analysis resulted in over 180,000 high-quality putative viral sequences. Redundant sequences were clustered at >95% average nucleotide identity over 80% of the shorter contig length. A final library of 83,681 nonredundant sequences was generated (Fig 5.1). 13,173 viruses were complete circular or ITR-flanked sequences and an additional 4858 viruses were deemed complete because they were flanked on both sides by chromosomal sequences. Lack of circularity or flanking chromosomal regions does not necessarily mean that a given contig is incomplete, and it can be difficult to detect many kinds of viral genome ends using short read assemblies. Although it is also challenging to obtain single contigs for very large viruses, 194 phages over 200 kb were detected in the survey, with the largest being Siphoviridae species ctpHQ1, at 501 kb. Only about half of the >200 kb phage contigs were bounded by direct repeats, suggesting a complete circular genome. Thirty-eight family- or order-level taxa were observed, and 3474 viral sequences were “dark matter” that could not be classified. However, as viral taxonomy, especially taxonomy of dsDNA phage, is in flux to increase resolution of diverse viruses[4], these numbers will likely change in the future. The vast majority of observed sequences represent dsDNA phages in the order *Caudovirales*. Relatively small amounts of human-tropic viruses were uncovered, including adenoviruses, anelloviruses, circoviruses, herpesviruses, norovirus

(caliciviridae), papillomaviruses, and polyomaviruses. Each of the human-tropic viruses mapped to previously reported virus species.

Figure 5.1 presents a summary of observed virus taxa. One taxon, designated "Phyco-like" viruses, encompasses 123 contigs. This is an interesting group of sequences defined by Cenote-Taker2 as phycodnaviridae due to similarity of the terminase/packaging gene of these viruses to the packaging gene of phycodnaviruses (~30% AA similarity). However, most of the virion structural gene models that are pinged by phyco-like viruses are from crAss-like phage, so the sequences probably represent phages, not eukaryotic viruses. This and the fact that most of the 3474 "Unclassified" viral sequences have hallmark genes corresponding to dsDNA phage models, supports the idea that much phage diversity is still unclassified and undescribed.

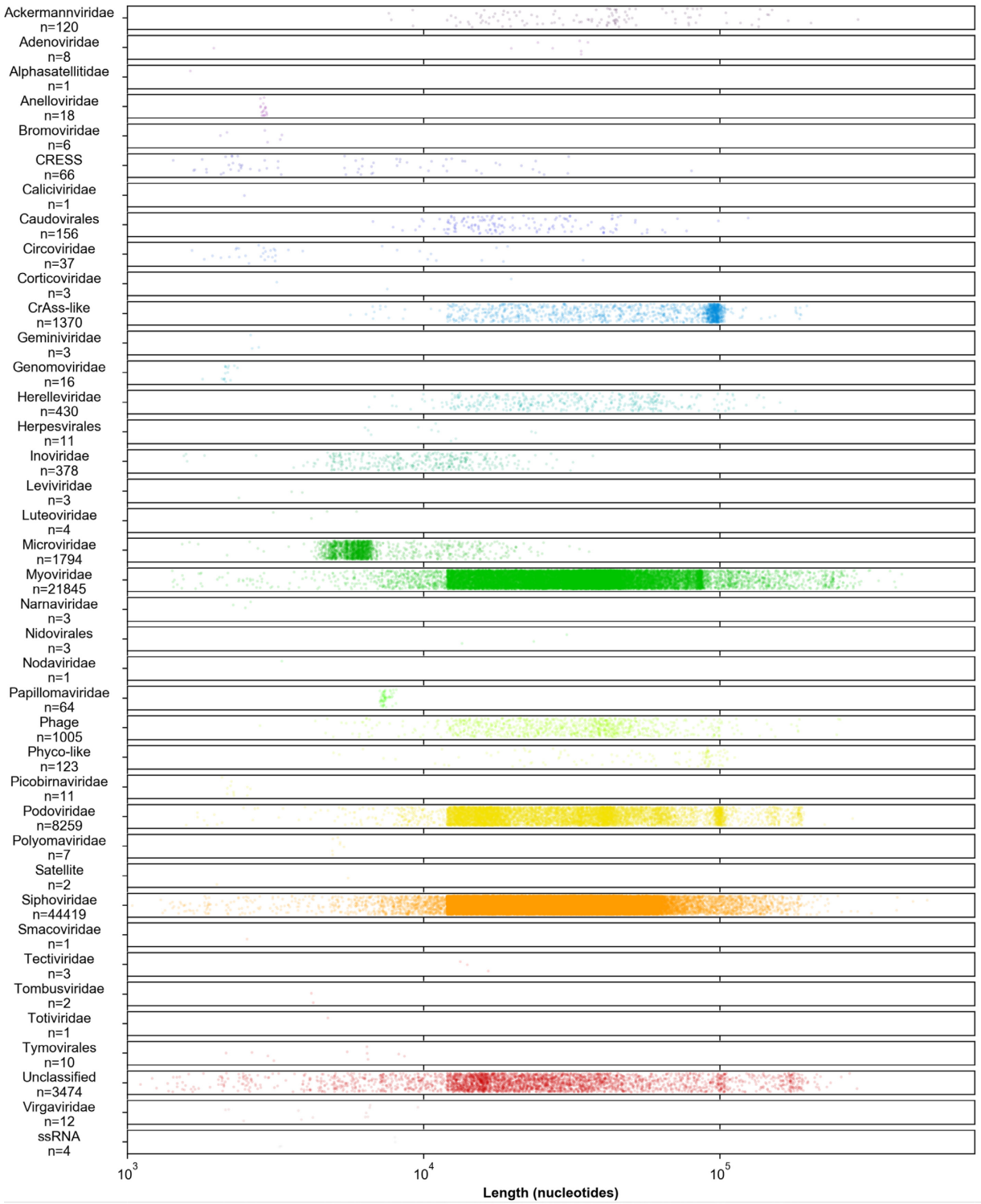


Figure 5.1: Summary of virus contig taxonomy and length

Contigs were split based on Cenote-Taker2 taxonomy calls. Each contig is represented as dot with the X-axis value representing contig length. Contigs smaller than the arbitrary 12 kb cutoff are either circular, bounded by ITRs, or the result of trimming of bacterial chromosomal sequences.

CRISPR spacer analysis reveals host for most phages

Bacteria encode CRISPR-Cas systems, which contain CRISPR spacer arrays of short (~32 nt) sequences copied from and used against invading mobile genetic elements, especially phages. Matching bacterial CRISPR spacers to phage genomes is one way to determine if a bacterium has previously been exposed to a particular phage. Advances in cataloging of CRISPR spacers from bacterial genomes and optimization of phage/host matching pipelines allowed the association of most of the phages discovered in this project to bacterial hosts (http://crispr.genome.ulaval.ca/). Specifically, 61,886 of the 83,681 virus sequences had at least one CRISPR spacer match from a known bacterium or multiple bacteria, with 675,750 total spacers matched to the database of viruses (Fig 5.2). Organized by bacterial genus, interesting trends regarding CRISPR spacer acquisition become apparent. For example, genera *Bifidobacterium* and, to a lesser extent, *Neisseria* often encode dozens of spacers specific to individual phages while *Clostridium, Porphyromonas,* and *Leptotrichia* typically encode one or only a handful of spacers per phage.

crAss-like phages seem to be the target of relatively few spacers per genome, despite the fact that many collected sequences are >100kb circular genomes. In contrast, ssDNA phages of family *Microviridae*, despite their small size, seem to be frequently targeted by many spacers in *Bacteroides* and *Parabacteroides* CRISPR systems but not CRISPR systems in other bacteria.

Figure 5.2: Summary of CRISPR spacer matches to bacterial taxa

Each plot is data from a different bacterial genus (or higher taxonomy when genus not defined) with CRISPR spacer matches to 200 or more viruses, each dot is a virus taxon. Y-axis values represent number of bacterial CRISPR spacer hits each virus had from the host bacterium. Only viruses with one or more spacer matches are displayed. Myo = *Myoviridae*, Sipho = *Siphoviridae*, crAss = crAss-like viruses, Caudo = Other *Caudovirales*, Micro = *Microviridae*, Ino = *Inoviridae*, Unclass = Unclassified Viruses

The most abundant viruses on several body sites

With this library of viruses and the large sampling effort from the Human Microbiome Project, the question of "which viruses are the most common" for multiple body sites can be answered more confidently than previously possible. It should be noted that the Human Microbiome Project data is from healthy Americans between 18 and 40 years of age, and the conclusions here may not be generalizable to other populations. Data was downloaded from SRA and analyzed for hundreds of patients at six body sites (anterior nares, buccal mucosa, posterior fornix, tongue dorsum, supragingival plaque, and stool) (Fig 5.3, 5.4, 5.5), and viral abundance was sorted by median reads per kilobase pre million (RPKM). Each body site had a different set of common viruses, in line with the observation that microbial populations are discriminated by body geography. Mostly, the most common viruses were phages from prevalent bacteria, such as the genus *Bacteroides*. In one noteworthy exception, anterior nares (nasal cavity) samples contained human papillomavirus type 38 in high abundance and prevalence.

The first described crAssphage has been considered the most abundant virus in the human gut for several years[14,166,167], but the data in the current report show 23 other phages are more abundant in human gut samples (Fig 5.5)(the original crAssphage is labeled CrAss-like_virus__sp._ctBhb420 in this figure). This is a testament to the lack of a comprehensive virus catalog for human metagenomes rather than the paucity of crAssphage in these datasets, as this phage class is still seen in most datasets at lower abundance.

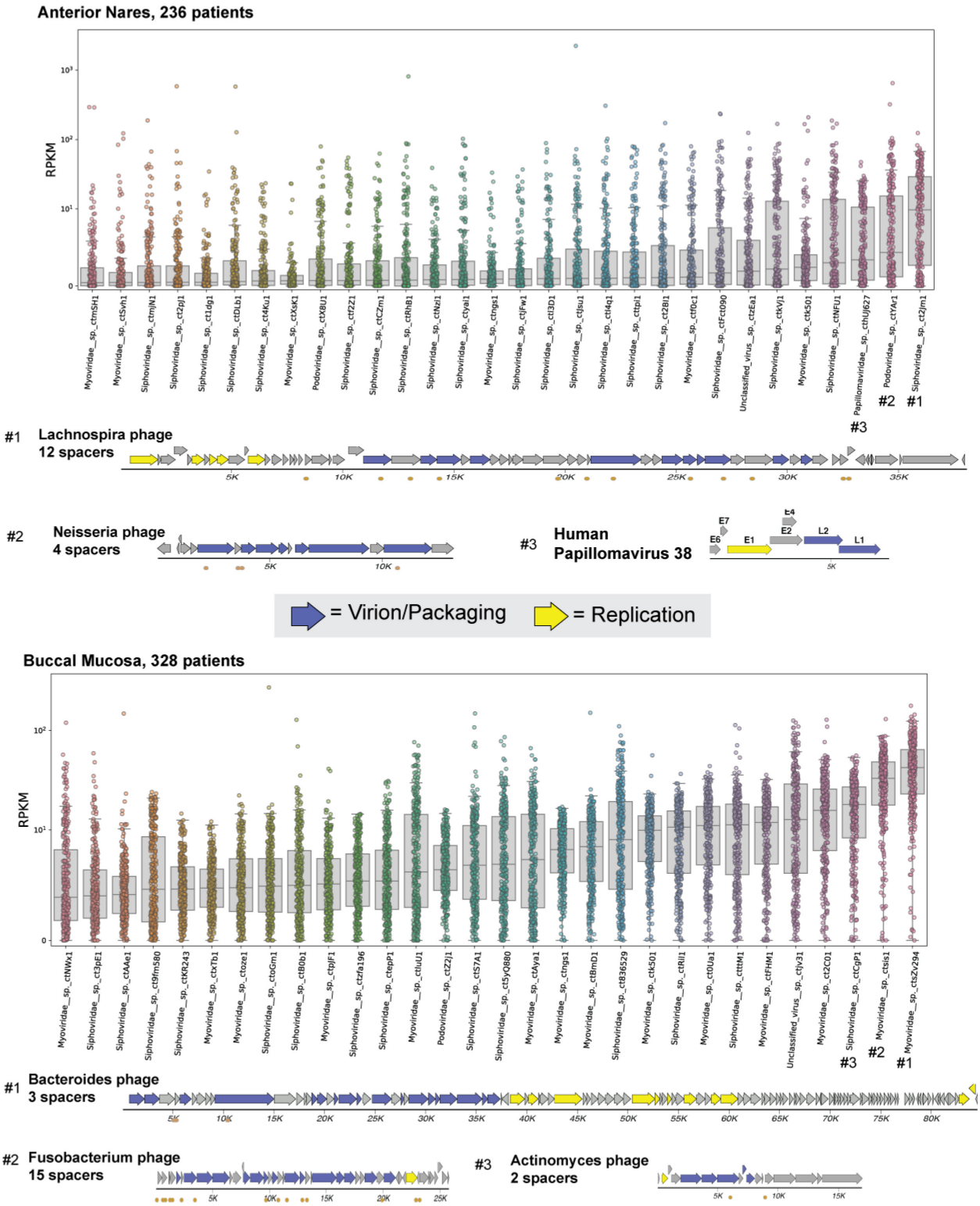


Figure 5.3: Most Common Viruses, Anterior Nares and Buccal Mucosa

The top thirty virus taxa for each body site are quantified. Genome maps are shown for the top three taxa with virion/packaging genes in blue, replication genes in yellow, and CRISPR spacer matches as orange dots (underneath).

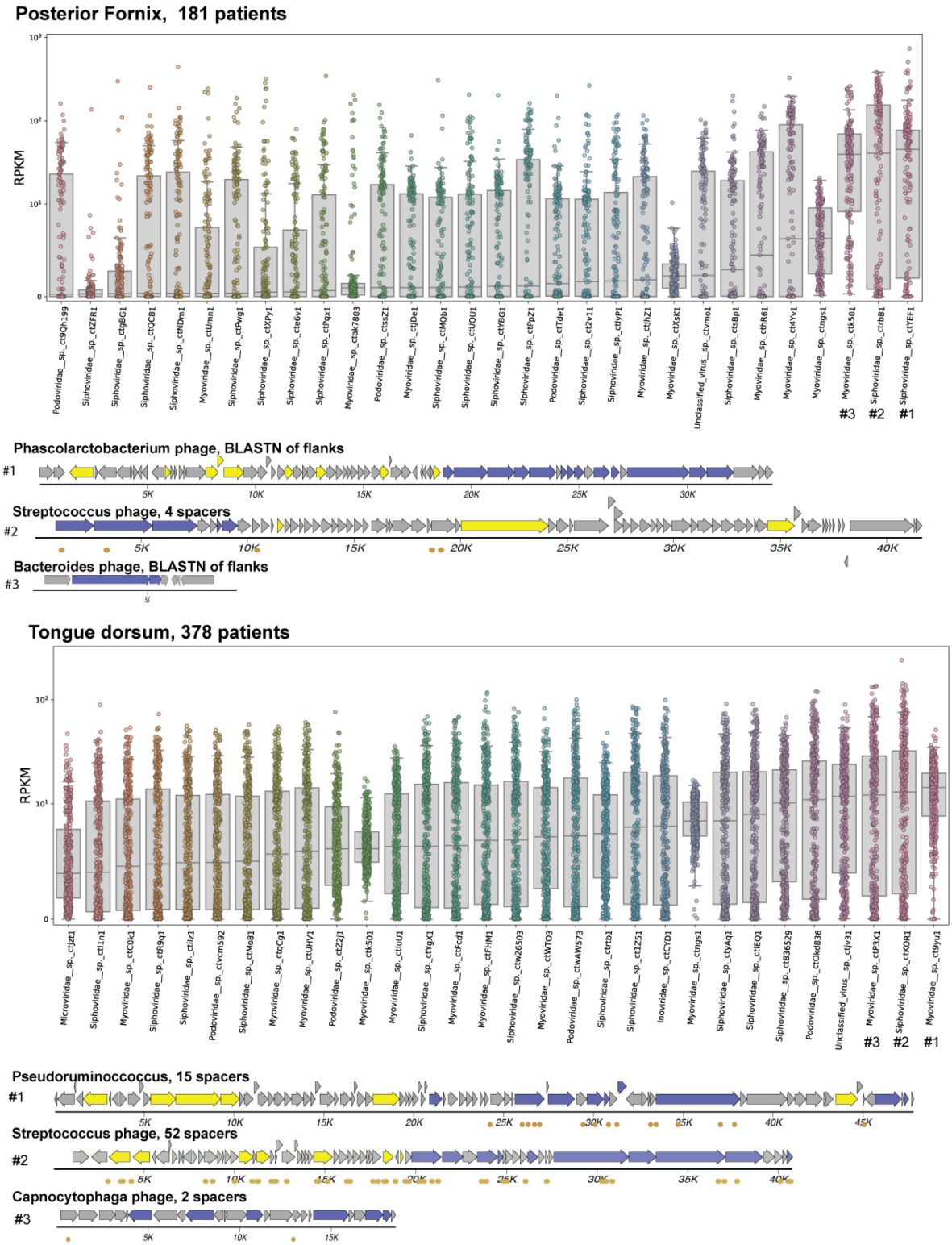


Figure 5.4: Most Common Viruses, Posterior Fornix and Tongue Dorsum

The top thirty virus taxa for each body site are quantified. Genome maps are shown for the top three taxa with virion/packaging genes in blue, replication genes in yellow, and CRISPR spacer matches as orange dots (underneath).

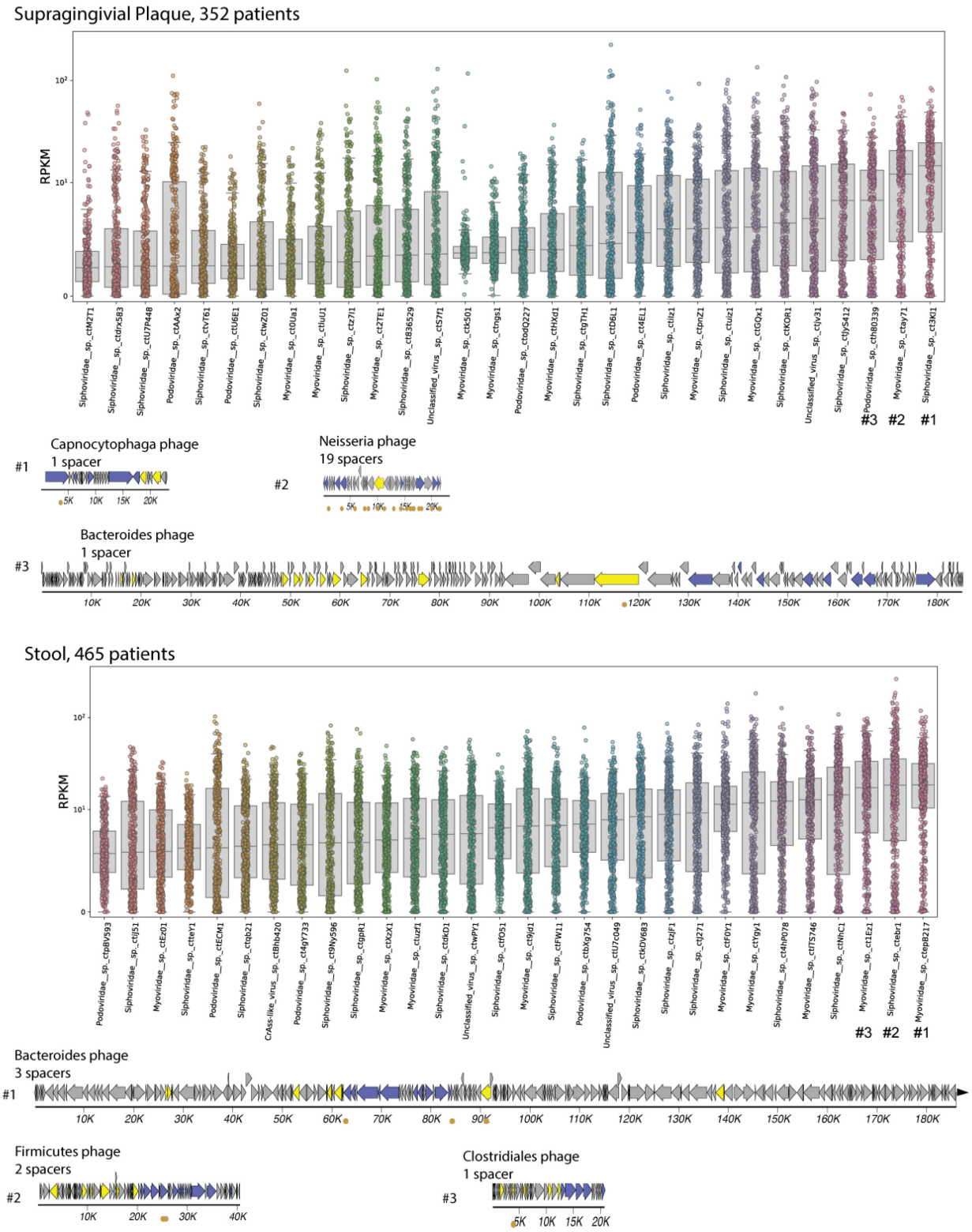


Figure 5.5: Most Common Viruses, Supragingival Plaque and Stool

The top thirty virus taxa for each body site are quantified. Genome maps are shown for the top three taxa with virion/packaging genes in blue, replication genes in yellow, and CRISPR spacer matches as orange dots (underneath).

Specific virus taxa are associated with human disease

Other studies have looked for associations between the virome and human diseases[26,28,163]. However, these studies were limited by a lack of a thorough virus reference database, and almost every study only examined sequences from viral particles[168]. Viral particles may not be the best reflection of the total viral population, especially in human digestive tracts, where most phage are believed to exist in lysogenic (non-lytic) states[169]. Furthermore, it is possible that the most important phages for human physiology are those that are integrated and expressing accessory genes, not those that are actively lysing their bacterial hosts. Thus it may be ideal to examine total DNA sequencing of samples that are not enriched for viruses. Further, viral particle preparations are notoriously difficult, with user error effects being higher than from whole genome shotgun (WGS) preparations[170].

This study looked at publicly available sequencing data from large case-control studies with stool and/or saliva WGS samples[127,130,171-177]. By comparing the abundance of each virus taxon between case and control cohorts, strong associations were seen. RPKM was used to measure abundance, and 100 bootstraps were conducted for each virus test to estimate the p-value (Fig 5.6A "Virome"). Confidence intervals for each p-value were calculated, but the table is too large to attach to this document. The analysis was compared to bacterial species-level single-copy marker gene abundance from the same data using IGGsearch[22] (Fig 5.6A "Bacteriome"). More statistically significant taxa were found for the virome than the bacteriome, even after multiple test correction. However, since more tests were performed for the virome, this is not entirely unexpected. Swarm plots of data from individual virus taxa (top twenty most-significant p values) are shown for reference (Fig. 5.6B).

The importance of considering effect size when reporting microbiome associations has become apparent in recent years[178]. Effect size (Cohen's d absolute value) for the virome and bacteriome are shown in Figure 5.6C, with values 0.2 - 0.5 implying a small effect size, values of 0.5 - 0.8 implying a medium effect size, and values of > 0.8 implying a large effect size[179]. While changes in the virome are expected to reflect changes of bacteriome to some extent, it is interesting that the effect sizes for virus taxa are generally larger. Furthermore, the predicted hosts for significant viruses are generally similar to the list of significant bacterial taxa (Fig. 5.6D). Since phages can infect multiple species within a bacterial genus, and sometimes in multiple genera, it is not clear how to correlate each virus with a specific host or hosts.

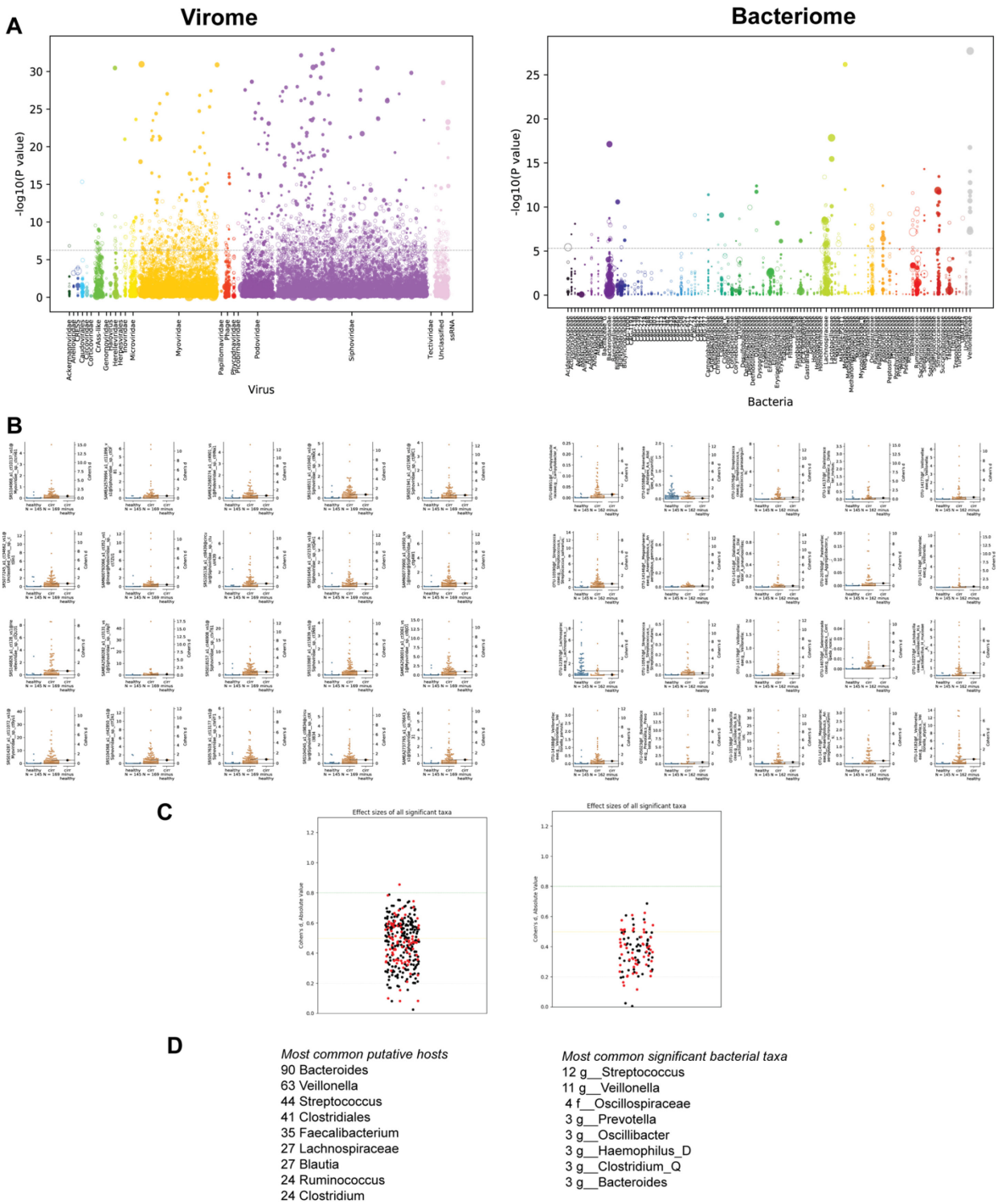


Figure 5.6 Association of the Gut Virome and Bacteriome with Liver Cirrhosis

Read data from PRJEB6337. (A) Differential abundance of viral (left) and bacterial (right) taxa in stool between cirrhosis patients (n=169) and healthy controls (n=145). Each taxon is represented as a dot along the X-axis, with the Y-axis being -log10 of the p-value. The size of

each dot corresponds to the median abundance of the taxon in the disease cohort and solid dots meaning increased abundance in the diseased state and hollow dots meaning decreased abundance in the diseased sate. The grey dotted line represents the Bonferroni-corrected significance threshold. (B) Swarm plots with Cohen's d effect sizes of the top 20 most significant taxa. (C) Plots of Cohen's d effect size (absolute value, black dots are positive and red dots are negative) from all taxa exceeding the Bonferroni-corrected significance threshold. Small effect size = 0.2 - 0.5 ; Medium effect size = 0.5 - 0.8 ; Large effect size = > 0.8. (D) Most common putative host of significant viruses is based on CRISPR spacers.

This same analysis was repeated for nine other case-control studies (Fig 5.7, 5.8, 5.9 [just virome shown]), and, in most cases, strong associations were found between the abundance of specific virus taxa and the disease state. While many more case-control studies comparing patient microbiomes exists, virtually all others use bacterial 16s amplicon sequencing, which is not suitable for virome analysis.

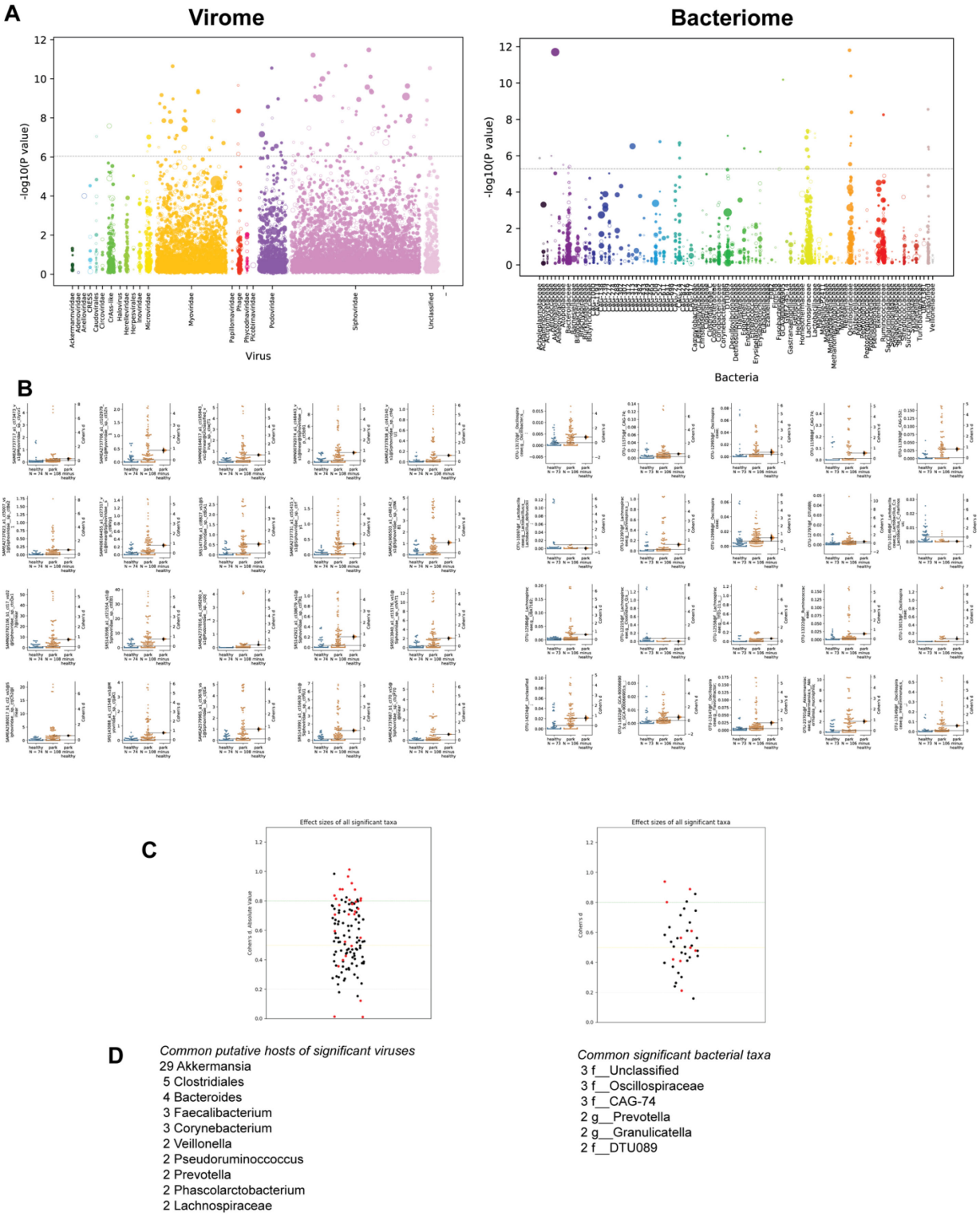


Figure 5.7 Association of the Gut Virome and Bacteriome with Parkinson's Disease

Read data from PRJEB17784. See Figure 5.6

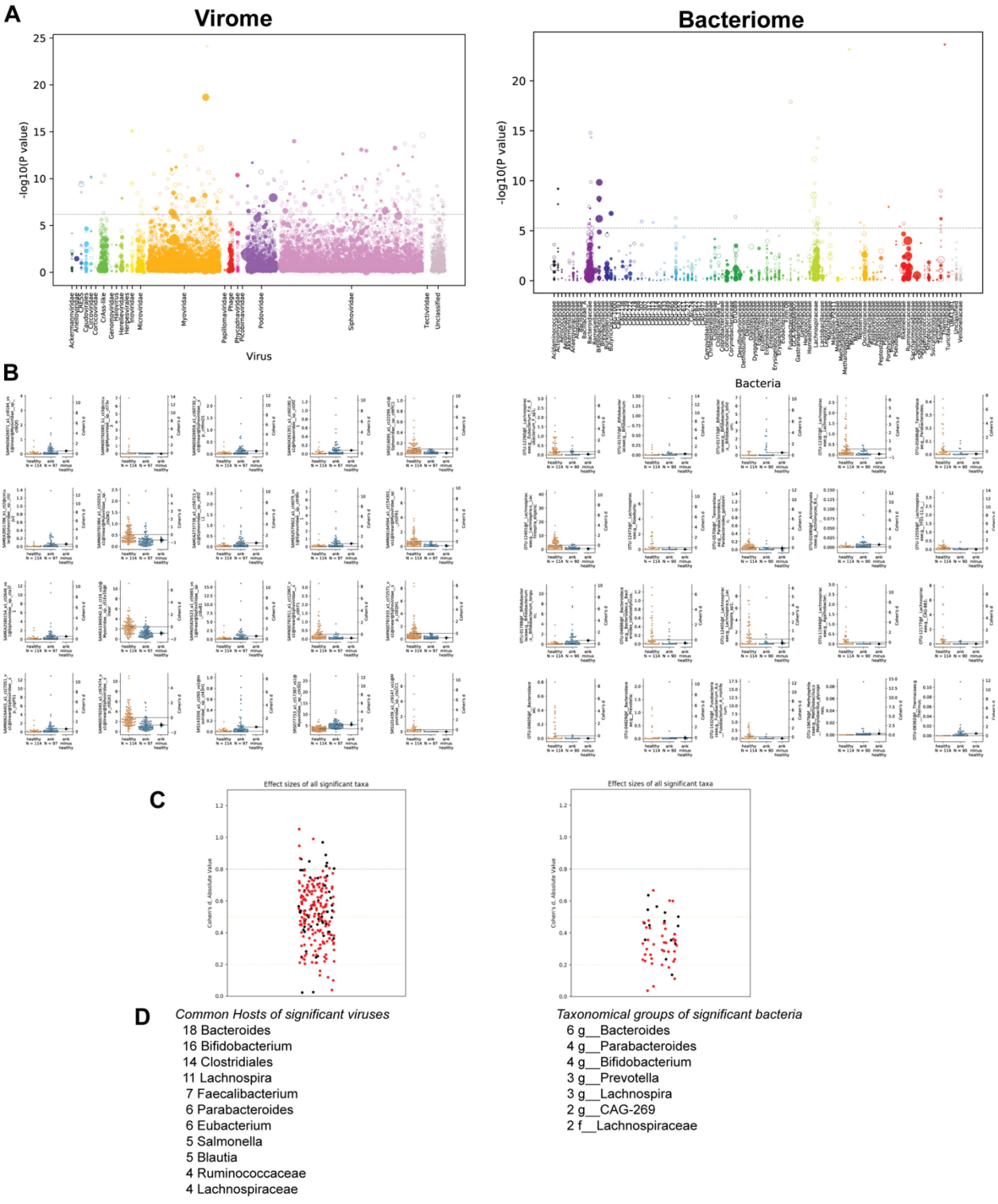


Figure 5.8 Association of the Gut Virome and Bacteriome with Ankylosing Spondylitis

Read data from PRJNA375935. See Figure 5.6

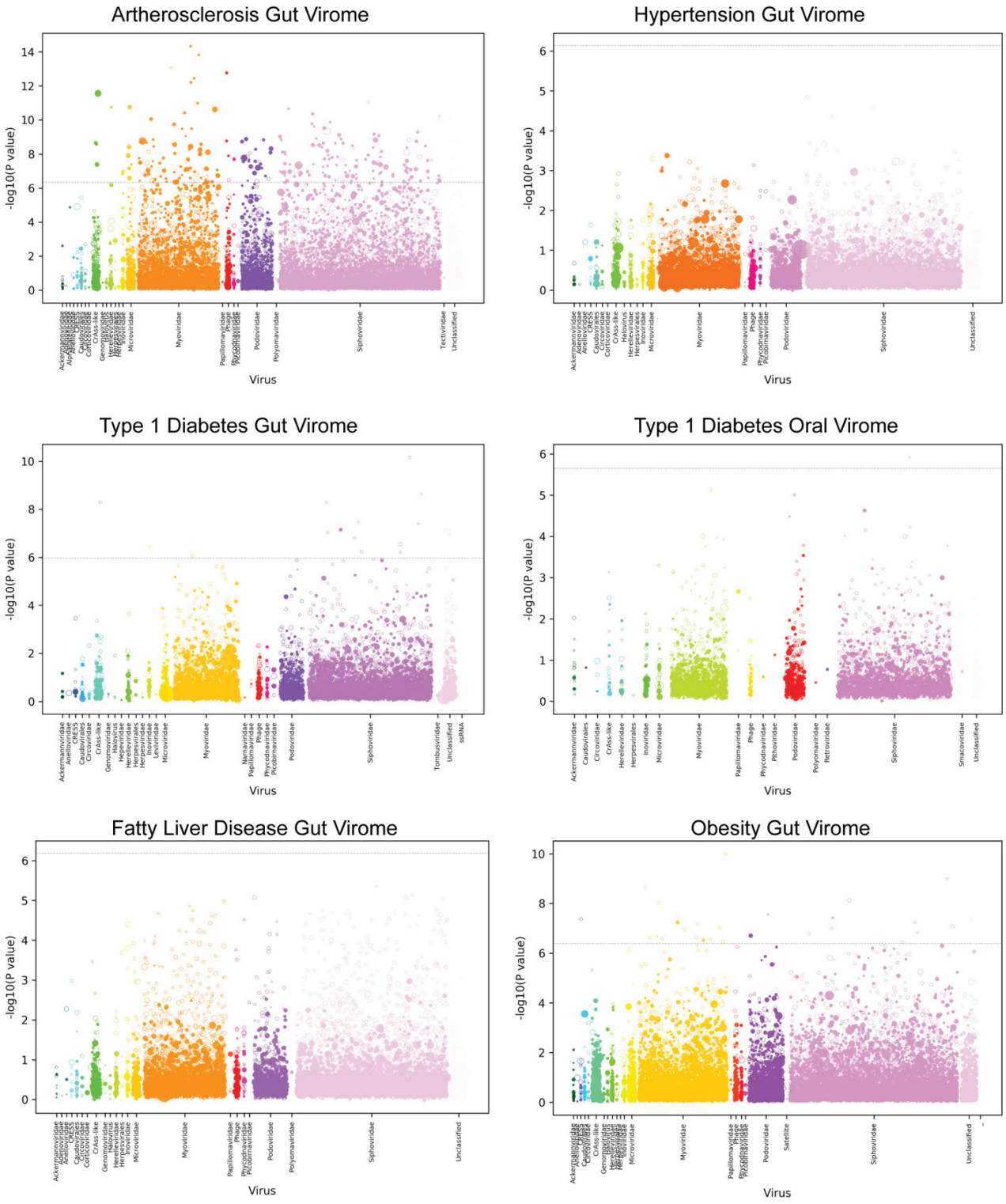


Figure 5.9 Association of the Virome with Other Diseases

Atherosclerosis reads from PRJEB21528. Hypertension reads from PRJEB13870. Type 1 diabetes reads from PRJNA289586. Fatty liver disease reads from PRJNA373901. Obesity reads from PRJEB4336.

Discussion

This study has demonstrated that, by leveraging virus-specific hallmark genes, it is possible to mine human metagenomic data at a large scale to create a comprehensive database that includes previously unknown viral genomes. This advance in turn allowed associations to be discovered between a variety of chronic disease states and specific virus taxa. It should be stressed that association is not the same as causation, and a variety of associative relationships between viruses and a given disease state are possible. To name a few: virus abundance might simply be an epiphenomenon reflecting bacterial host abundance, the human genetics that predispose people to a disease might also provide a more favorable environment for the virus, the external causes of a disease may create a more favorable environment for the virus, or the virus may contribute to the disease presentation in some way but ultimately does not cause the disease in isolation from other factors. Of course, verifying any association with independent studies of the same disease will be key to understanding how much of these findings are generalizable.

A limitation of the case-control studies that were analyzed was that they generally only had data for a single timepoint for each subject. Viromes can be "noisy", and longitudinal data on individual patients can be much more effective at discerning stable viral populations[125]. This may have been partly offset by use of large cohort sizes (mostly over 150 total patients). Another consideration is that the case-control data were all DNA WGS sequencing, whereas RNA sequencing of metatranscriptomes would provide more functional data on expression of specific genes, potentially leading to more hypotheses on possible mechanisms of action.

Even with relatively inclusive criteria used by Cenote-Taker2 (discernable amino acid similarity of a viral hallmark gene to a protein the RefSeq virus database), thousands of viruses that live on humans from this dataset could not be classified, suggesting that additional families of as-yet-uncultured viruses await formal discovery.

Methods

Identification of viral contigs in assemblies

Studies with human metagenome sequences were chosen somewhat arbitrarily, though the Human Microbiome Project data and many studies referenced in the metastudy from Nayfach et al[22] were used. For each bioproject, run tables were downloaded from SRA and unique biosamples were delineated. All runs from a given biosample were downloaded concurrently, trimmed with Fastp[180], and co-assembled with Megahit[165] using default settings. Subsequent contigs were fed to Cenote-Taker2, with settings to consider circular contigs of at least 1500 nt, ITR-containing contigs of at least 4 kb, and linear contigs of at least 12 kb. These contigs were scanned for genes matching viral hallmark models, and circular and ITR-containing contigs with one or more viral hallmark genes were kept as well as linear contigs with two or more viral hallmark genes. Cenote-Taker2 hallmark gene database was the April 21, 2020 version. While Cenote-Taker2 does take steps to remove plasmids and conjugative transposons, extra precautions were taken by removing ~4000 putative viral sequences from the non-redundant database that contained replication-associated but not virion or packaging viral hallmark genes.

Clustering similar contigs

RedRed (https://github.com/kseniaarkhipova/RedRed), a circularity-aware algorithm, was used to cluster all circular contigs (95% ANI, 80% length), and these non-redundant circles were then clustered with linear contigs using cd-hit (95% ANI, 80% length).

CRISPR spacer analysis

CrisprOpenDB was used (commit 04e4ffcc55d65cf8e13afe55e081b14773a6bb70) to assign phages to hosts based on CRISPR spacer match. Three mismatches were allowed for hits. For hits to bacteria without a currently assigned genus, family-level or order-level taxonomical information was pulled from the output table.

Determining abundance of individual virus contigs/genomes

The final database of viral sequences was processed by RepeatMasker to remove low-complexity regions that recruit reads non-specifically[181]. Additionally, linear contigs were pruned by 3 kb on the 5' and 3' ends, as even a short sequence from flanking chromosome could dampen the true viral signal. Finally, a representative genome from all 410 viruses reported to infect humans was downloaded from NCBI, and processed by RepeatMasker, and added to the final database for read alignment. Bowtie2 was used to align reads to the database, and samtools idxstats was used to calculate read coverage for each contig.

Comparing virus abundance in case-control studies

Wilcoxon rank-sum test was computed with 100 bootstraps using Python, NumPy and SciPy for each virus in a given study where at least one sample had an RPKM of 1. Cohen's d effect size was calculated using DaBest Python package with 5000 bootstraps.

Contributions

Michael J Tisza: Conceptualization, Resources, Data curation, Formal analysis, Validation, Investigation, Visualization, Methodology, Project administration

Christopher B Buck: Conceptualization, Resources, Project administration

6 Conclusions and Future Directions

This dissertation has presented work that has advanced the field of virology, specifically showing innovation in virus discovery and the role of the virome in human health.

Chapter 2, *Discovery of several thousand circular DNA viruses*, showed the power of virus particle enrichment strategies paired with bioinformatics, and over 2500 new complete circular DNA virus genomes from humans and other animals were discovered and analyzed in detail. The vast majority of the circular sequences were quite different from previously described viruses. However, most of these sequences also had genes that were detectably similar to known sequences at some level. It will be interesting to see whether, as in this study, the ~20% of circular sequences in other viromics datasets likewise have no discernable similarity to known viruses. Additionally, one could look at similar datasets generated by other scientists to see what percentage of circular contigs have similarity to any of the "dark matter genome groups" documented in this thesis. Potential cellular hosts of some of these elements could be deciphered with more thorough searching of public datasets, especially RNA sequencing experiments of cultured cells or model organisms. If transcripts of a viral gene are expressed in a given host it is likely that the virus is tropic for that host.

Chapter 3, *Bibiviruses are a New, Unusual Virus Family Common in the Human Gut*, continued on the theme of identifying unknown sequences in viromic datasets. This chapter shows how orthogonal computational and wet-bench methods can synergize to yield interesting discoveries. Although bibivirus sequences can be found in many human fecal

viromics datasets, they appear to have been ignored, just like the many other sequences (typically 50%-70%)[67] that are unidentifiable with conventional methods.

A direction of future research inspired by this project would be to take a panel of bacteria isolated from human samples and use a variety of prophage induction methods on each one. Nuclease-resistant nucleic acids could then be collected and sequenced. Mapping reads from these preparations back to the host bacterium genome would be an effective method to identify novel viruses as well as other mobile genetic elements capable of packaging themselves into viral capsids in a way that doesn't have many of the same biases as the traditional plaque assays or sequence similarity-based approaches. Further, once identified, novel elements would have a readily available culture system for further experimentation.

In Chapter 4, *Cenote-Taker2 Democratizes Virus Discovery and Sequence Annotation*, the nuts and bolts of Cenote-Taker2 were presented. This pipeline, already being used by many labs worldwide, utilizes a clear and precise method to find familiar and divergent viruses in any type of shotgun sequencing data or genome assembly. Using highly sensitive models that represent structurally or experimentally defined virus hallmark genes utilizes the power of computation without divorcing a discovered sequence from human understanding. For example, if a sequence has a gene that Cenote-Taker2 determines is a virus tail tube protein, a user can easily confirm this with a quick CDD search or HHpred search of that protein. My own back-and-forth with the computer on these types of puzzles has led to a highly refined database of Hidden Markov Models of virus hallmark genes for virus discovery. On the other hand, virion structural and replication genes might sometimes be co-opted by other mobile genetic elements, and this can lead to occasional false positive virus calls. For example, some bacterial

conjugative transposons have genes with similarity to phage tail proteins[182], and bacterial gene transfer agents have major capsid protein-like genes[183]. Moreover, the hallmark gene strategy, while effective, is certainly not the only viable strategy, as laid out in Chapters 2 and 3.

Chapter 5, *The Human Virome: Over Eighty Thousand Distinct Viruses and Specific Associations with Chronic Diseases*, brings the lessons learned from other the virus discovery projects into human relevance. Completion of the Cenote-Taker2 pipeline empowered me to mine 6000 deep sequencing datasets. Although the large number of virus taxa (83,681) discovered in this work might seem impressive, it is probably still far from a complete catalog. For example, researchers have confidently assembled draft genomes for over 24,000 species of bacteria from human metagenomes[22], and each bacterial species likely hosts many distinct viruses, some of which may belong to virus families that have not yet been recognized.

As discussed in Chapter 5, a single "snapshot" of a microbial community using DNA sequencing is not necessarily the best method of finding important taxa. Longitudinal studies would likely increase resolution, as well as RNA sequencing. It would also certainly be ideal to have prospective studies that periodically collected samples from a large group of individuals without a disease, then compared how microbial communities change for those who acquired the disease. Invoking the infamous “hit-and-run” phenomenon, a particular virus could infect and decimate a bacterial population early in a disease, leading to some of the changes responsible for a disease. However, after killing all of its potential hosts would not be available for detection in an advanced disease state.

Even when strong associations are found, substantial additional work will be required to determine whether any particular virus or viruses contribute to a disease. Mice have been used

to research the effects of microbial communities on various health outcomes, but, like any model system, there are a variety of ways that mice do not recapitulate human physiology, and microbial communities implanted into mice undergo a variety of changes once implanted[184]. That said, particularly attractive phage candidates would have predicted metabolic genes that were known to be upregulated in a particular disease state. This would allow an experimentalist to focus on a single gene and one or a few metabolites that could be more tractably studied both in cell culture and in a mouse model.

Bibliography

1 Krupovic, M., Dolja, V. V. & Koonin, E. V. Origin of viruses: primordial replicators recruiting capsids from hosts. *Nat Rev Microbiol* **17**, 449-458, doi:10.1038/s41579-019-0205-6 (2019).

2 Woolhouse, M., Scott, F., Hudson, Z., Howey, R. & Chase-Topping, M. Human viruses: discovery and emergence. *Philos Trans R Soc Lond B Biol Sci* **367**, 2864-2871, doi:10.1098/rstb.2011.0354 (2012).

3 Mushegian, A. R. Are There 10(31) Virus Particles on Earth, or More, or Fewer? *J Bacteriol* **202**, doi:10.1128/JB.00052-20 (2020).

4 Koonin, E. V. *et al.* Global Organization and Proposed Megataxonomy of the Virus World. *Microbiol Mol Biol Rev* **84**, doi:10.1128/MMBR.00061-19 (2020).

5 Lustig, A. & Levine, A. J. One hundred years of virology. *J Virol* **66**, 4629-4631 (1992).

6 Fiers, W. *et al.* Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260**, 500-507, doi:10.1038/260500a0 (1976).

7 Simmonds, P. *et al.* Consensus statement: Virus taxonomy in the age of metagenomics. *Nat Rev Microbiol* **15**, 161-168, doi:10.1038/nrmicro.2016.177 (2017).

8 Breitbart, M. *et al.* Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* **99**, 14250-14255, doi:10.1073/pnas.202488399 (2002).

9 Bronfenbrenner, J. J. & Korb, C. Studies on the Bacteriophage of D'herelle : Iii. Some of the Factors Determining the Number and Size of Plaques of Bacterial Lysis on Agar. *J Exp Med* **42**, 483-497, doi:10.1084/jem.42.4.483 (1925).

10 Dayaram, A. *et al.* Diverse small circular DNA viruses circulating amongst estuarine molluscs. *Infect Genet Evol* **31**, 284-295, doi:10.1016/j.meegid.2015.02.010 (2015).

11 Nielsen, H. B. *et al.* Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat Biotechnol* **32**, 822-828, doi:10.1038/nbt.2939 (2014).

12 Seguritan, V. *et al.* Artificial neural networks trained to detect viral and phage structural proteins. *PLoS Comput Biol* **8**, e1002657, doi:10.1371/journal.pcbi.1002657 (2012).

13 Shkoporov, A. N. & Hill, C. Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome. *Cell Host Microbe* **25**, 195-209, doi:10.1016/j.chom.2019.01.017 (2019).

14 Dutilh, B. E. *et al.* A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat Commun* **5**, 4498, doi:10.1038/ncomms5498 (2014).

15 Enault, F. *et al.* Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *ISME J* **11**, 237-247, doi:10.1038/ismej.2016.90 (2017).

16 Krupovic, M. & Koonin, E. V. Multiple origins of viral capsid proteins from cellular ancestors. *Proc Natl Acad Sci U S A* **114**, E2401-E2410, doi:10.1073/pnas.1621061114 (2017).

17 Zimmermann, L. *et al.* A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J Mol Biol* **430**, 2237-2243, doi:10.1016/j.jmb.2017.12.007 (2018).

18 Kazlauskas, D., Varsani, A., Koonin, E. V. & Krupovic, M. Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. *Nat Commun* **10**, 3425, doi:10.1038/s41467-019-11433-0 (2019).

19 Hooper, L. V., Littman, D. R. & Macpherson, A. J. Interactions between the microbiota and the immune system. *Science* **336**, 1268-1273, doi:10.1126/science.1223490 (2012).

20 Karlsson, F., Tremaroli, V., Nielsen, J. & Backhed, F. Assessing the human gut microbiota in metabolic diseases. *Diabetes* **62**, 3341-3349, doi:10.2337/db13-0844 (2013).

21 Lynch, S. V. & Pedersen, O. The Human Intestinal Microbiome in Health and Disease. *N Engl J Med* **375**, 2369-2379, doi:10.1056/NEJMra1600266 (2016).
22 Nayfach, S., Shi, Z. J., Seshadri, R., Pollard, K. S. & Kyrpides, N. C. New insights from uncultivated genomes of the global human gut microbiome. *Nature* **568**, 505-510, doi:10.1038/s41586-019-1058-x (2019).
23 Hsu, B. B. *et al.* Dynamic Modulation of the Gut Microbiota and Metabolome by Bacteriophages in a Mouse Model. *Cell Host Microbe* **25**, 803-814 e805, doi:10.1016/j.chom.2019.05.001 (2019).
24 Wagner, P. L. & Waldor, M. K. Bacteriophage control of bacterial virulence. *Infect Immun* **70**, 3985-3993, doi:10.1128/iai.70.8.3985-3993.2002 (2002).
25 Sarowska, J. *et al.* Virulence factors, prevalence and potential transmission of extraintestinal pathogenic Escherichia coli isolated from different sources: recent reports. *Gut Pathog* **11**, 10, doi:10.1186/s13099-019-0290-0 (2019).
26 Norman, J. M. *et al.* Disease-specific alterations in the enteric virome in inflammatory bowel disease. *Cell* **160**, 447-460, doi:10.1016/j.cell.2015.01.002 (2015).
27 Reyes, A. *et al.* Gut DNA viromes of Malawian twins discordant for severe acute malnutrition. *Proc Natl Acad Sci U S A* **112**, 11941-11946, doi:10.1073/pnas.1514285112 (2015).
28 Clooney, A. G. *et al.* Whole-Virome Analysis Sheds Light on Viral Dark Matter in Inflammatory Bowel Disease. *Cell Host Microbe* **26**, 764-778 e765, doi:10.1016/j.chom.2019.10.009 (2019).
29 Beller, L. & Matthijnssens, J. What is (not) known about the dynamics of the human gut virome in health and disease. *Curr Opin Virol* **37**, 52-57, doi:10.1016/j.coviro.2019.05.013 (2019).
30 Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61-66, doi:10.1038/nature23889 (2017).
31 Krishnamurthy, S. R. & Wang, D. Origins and challenges of viral dark matter. *Virus Res* **239**, 136-142, doi:10.1016/j.virusres.2017.02.002 (2017).
32 Oh, J. *et al.* Biogeography and individuality shape function in the human skin metagenome. *Nature* **514**, 59-64, doi:10.1038/nature13786 (2014).
33 Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536**, 425-430, doi:10.1038/nature19094 (2016).
34 Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nat Microbiol* **3**, 870-880, doi:10.1038/s41564-018-0190-y (2018).
35 Pastrana, D. V. *et al.* Metagenomic Discovery of 83 New Human Papillomavirus Types in Patients with Immunodeficiency. *mSphere* **3**, doi:10.1128/mSphereDirect.00645-18 (2018).
36 Turnbaugh, P. J. *et al.* The human microbiome project. *Nature* **449**, 804-810, doi:10.1038/nature06244 (2007).
37 Gilbert, J. A. *et al.* Meeting report: the terabase metagenomics workshop and the vision of an Earth microbiome project. *Stand Genomic Sci* **3**, 243-248, doi:10.4056/sigs.1433550 (2010).
38 O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745, doi:10.1093/nar/gkv1189 (2016).
39 Brister, J. R., Ako-Adjei, D., Bao, Y. & Blinkova, O. NCBI viral genomes resource. *Nucleic Acids Res* **43**, D571-577, doi:10.1093/nar/gku1207 (2015).
40 Sullivan, M. B. Viromes, not gene markers, for studying double-stranded DNA virus communities. *J Virol* **89**, 2459-2461, doi:10.1128/JVI.03289-14 (2015).
41 Rohwer, F. & Edwards, R. The Phage Proteomic Tree: a genome-based taxonomy for phage. *J Bacteriol* **184**, 4529-4535, doi:10.1128/jb.184.16.4529-4535.2002 (2002).
42 Shi, M. *et al.* Redefining the invertebrate RNA virosphere. *Nature*, doi:10.1038/nature20167 (2016).

43 Pope, W. H. *et al.* Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. *Elife* **4**, e06416, doi:10.7554/eLife.06416 (2015).
44 Grose, J. H. & Casjens, S. R. Understanding the enormous diversity of bacteriophages: the tailed phages that infect the bacterial family Enterobacteriaceae. *Virology* **468-470**, 421-443, doi:10.1016/j.virol.2014.08.024 (2014).
45 Grose, J. H., Jensen, G. L., Burnett, S. H. & Breakwell, D. P. Genomic comparison of 93 Bacillus phages reveals 12 clusters, 14 singletons and remarkable diversity. *BMC Genomics* **15**, 855, doi:10.1186/1471-2164-15-855 (2014).
46 Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from microbial genomic data. *PeerJ* **3**, e985, doi:10.7717/peerj.985 (2015).
47 Gregory, A. C. *et al.* Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* **177**, 1109-1123 e1114, doi:10.1016/j.cell.2019.03.040 (2019).
48 Roux, S. *et al.* Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nat Microbiol*, doi:10.1038/s41564-019-0510-x (2019).
49 Bedell, M. A. *et al.* Amplification of human papillomavirus genomes in vitro is dependent on epithelial differentiation. *J Virol* **65**, 2254-2260 (1991).
50 Koonin, E. V., Dolja, V. V. & Krupovic, M. Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology* **479-480**, 2-25, doi:10.1016/j.virol.2015.02.039 (2015).
51 Blinkova, O. *et al.* Novel circular DNA viruses in stool samples of wild-living chimpanzees. *The Journal of general virology* **91**, 74-86, doi:10.1099/vir.0.015446-0 (2010).
52 Dayaram, A. *et al.* Diverse circular replication-associated protein encoding viruses circulating in invertebrates within a lake ecosystem. *Infect Genet Evol* **39**, 304-316, doi:10.1016/j.meegid.2016.02.011 (2016).
53 Labonte, J. M. & Suttle, C. A. Previously unknown and highly divergent ssDNA viruses populate the oceans. *ISME J* **7**, 2169-2177, doi:10.1038/ismej.2013.110 (2013).
54 Rosario, K. *et al.* Virus discovery in all three major lineages of terrestrial arthropods highlights the diversity of single-stranded DNA viruses associated with invertebrates. *PeerJ* **6**, e5761, doi:10.7717/peerj.5761 (2018).
55 Victoria, J. G. *et al.* Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *J Virol* **83**, 4642-4651, doi:10.1128/JVI.02301-08 (2009).
56 Kim, K. H. *et al.* Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Appl Environ Microbiol* **74**, 5975-5985, doi:10.1128/AEM.01275-08 (2008).
57 Peretti, A., FitzGerald, P. C., Bliskovsky, V., Buck, C. B. & Pastrana, D. V. Hamburger polyomaviruses. *The Journal of general virology* **96**, 833-839, doi:10.1099/vir.0.000033 (2015).
58 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410, doi:10.1016/S0022-2836(05)80360-2 (1990).
59 Marchler-Bauer, A. & Bryant, S. H. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res* **32**, W327-331, doi:10.1093/nar/gkh454 (2004).
60 Marchler-Bauer, A. *et al.* CDD: NCBI's conserved domain database. *Nucleic Acids Res* **43**, D222-226, doi:10.1093/nar/gku1221 (2015).
61 Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* **9**, 173-175, doi:10.1038/nmeth.1818 (2011).
62 El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Res* **47**, D427-D432, doi:10.1093/nar/gky995 (2019).
63 UniProt, C. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **47**, D506-D515, doi:10.1093/nar/gky1049 (2019).

64 Chandonia, J. M., Fox, N. K. & Brenner, S. E. SCOPe: classification of large macromolecular structures in the structural classification of proteins-extended database. *Nucleic Acids Res* **47**, D475-D481, doi:10.1093/nar/gky1134 (2019).
65 Burley, S. K. *et al.* Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. *Methods Mol Biol* **1607**, 627-641, doi:10.1007/978-1-4939-7000-1_26 (2017).
66 Roux, S. *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat Biotechnol* **37**, 29-37, doi:10.1038/nbt.4306 (2019).
67 Zolfo, M. P., F.; Asnicar, F.; Manghi, P.; Tett, A.; Bushman, F. D.; Segata, N. Detecting contamination in viromes using ViromeQC. *Nat Biotechnol* **37**, 1408–1412(2019), doi:https://doi.org/10.1038/s41587-019-0334-5 (2019).
68 Zhao, L., Rosario, K., Breitbart, M. & Duffy, S. Eukaryotic Circular Rep-Encoding Single-Stranded DNA (CRESS DNA) Viruses: Ubiquitous Viruses With Small Genomes and a Diverse Host Range. *Adv Virus Res* **103**, 71-133, doi:10.1016/bs.aivir.2018.10.001 (2019).
69 Itzhaki, R. F. *et al.* Microbes and Alzheimer's Disease. *J Alzheimers Dis* **51**, 979-984, doi:10.3233/JAD-160152 (2016).
70 Eimer, W. A. *et al.* Alzheimer's Disease-Associated beta-Amyloid Is Rapidly Seeded by Herpesviridae to Protect against Brain Infection. *Neuron* **99**, 56-63 e53, doi:10.1016/j.neuron.2018.06.030 (2018).
71 Coras, R. *et al.* No evidence for human papillomavirus infection in focal cortical dysplasia IIb. *Ann Neurol* **77**, 312-319, doi:10.1002/ana.24328 (2015).
72 Chen, J. *et al.* Detection of human papillomavirus in human focal cortical dysplasia type IIB. *Ann Neurol* **72**, 881-892, doi:10.1002/ana.23795 (2012).
73 Asplund, M. *et al.* Contaminating viral sequences in high-throughput sequencing viromics: a linkage study of 700 sequencing libraries. *Clin Microbiol Infect*, doi:10.1016/j.cmi.2019.04.028 (2019).
74 Meier, A. & Soding, J. Automatic Prediction of Protein 3D Structures by Probabilistic Multi-template Homology Modeling. *PLoS Comput Biol* **11**, e1004343, doi:10.1371/journal.pcbi.1004343 (2015).
75 Huang, Y. J., Mao, B., Aramini, J. M. & Montelione, G. T. Assessment of template-based protein structure predictions in CASP10. *Proteins* **82 Suppl 2**, 43-56, doi:10.1002/prot.24488 (2014).
76 Gerlt, J. A. *et al.* Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim Biophys Acta* **1854**, 1019-1037, doi:10.1016/j.bbapap.2015.04.015 (2015).
77 Su, G., Morris, J. H., Demchak, B. & Bader, G. D. Biological network exploration with Cytoscape 3. *Curr Protoc Bioinformatics* **47**, 8.13.11-24, doi:10.1002/0471250953.bi0813s47 (2014).
78 Iranzo, J., Krupovic, M. & Koonin, E. V. A network perspective on the virus world. *Commun Integr Biol* **10**, e1296614, doi:10.1080/19420889.2017.1296614 (2017).
79 Bolduc, B. *et al.* vConTACT: an iVirus tool to classify double-stranded DNA viruses that infect Archaea and Bacteria. *PeerJ* **5**, e3243, doi:10.7717/peerj.3243 (2017).
80 Lima-Mendez, G., Van Helden, J., Toussaint, A. & Leplae, R. Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol Biol Evol* **25**, 762-777, doi:10.1093/molbev/msn023 (2008).
81 Lefeuvre, P. *et al.* Evolution and ecology of plant viruses. *Nat Rev Microbiol*, doi:10.1038/s41579-019-0232-3 (2019).
82 Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nat Biotechnol* **37**, 632-639, doi:10.1038/s41587-019-0100-8 (2019).

83 Kraberger, S. *et al.* Characterisation of a diverse range of circular replication-associated protein encoding DNA viruses recovered from a sewage treatment oxidation pond. *Infect Genet Evol* **31**, 73-86, doi:10.1016/j.meegid.2015.01.001 (2015).
84 Krupovic, M., Ravantti, J. J. & Bamford, D. H. Geminiviruses: a tale of a plasmid becoming a virus. *BMC Evol Biol* **9**, 112, doi:10.1186/1471-2148-9-112 (2009).
85 Hipp, K., Grimm, C., Jeske, H. & Bottcher, B. Near-Atomic Resolution Structure of a Plant Geminivirus Determined by Electron Cryomicroscopy. *Structure* **25**, 1303-1309 e1303, doi:10.1016/j.str.2017.06.013 (2017).
86 Bottcher, B., Unseld, S., Ceulemans, H., Russell, R. B. & Jeske, H. Geminate structures of African cassava mosaic virus. *J Virol* **78**, 6758-6765, doi:10.1128/JVI.78.13.6758-6765.2004 (2004).
87 Zhang, W. *et al.* Structure of the Maize streak virus geminate particle. *Virology* **279**, 471-477, doi:10.1006/viro.2000.0739 (2001).
88 Fontenele, R. S. *et al.* Single Stranded DNA Viruses Associated with Capybara Faeces Sampled in Brazil. *Viruses* **11**, doi:10.3390/v11080710 (2019).
89 Kraberger, S., Schmidlin, K., Fontenele, R. S., Walters, M. & Varsani, A. Unravelling the Single-Stranded DNA Virome of the New Zealand Blackfly. *Viruses* **11**, doi:10.3390/v11060532 (2019).
90 Krupovic, M., Ghabrial, S. A., Jiang, D. & Varsani, A. Genomoviridae: a new family of widespread single-stranded DNA viruses. *Arch Virol* **161**, 2633-2643, doi:10.1007/s00705-016-2943-3 (2016).
91 Zerbini, F. M. *et al.* ICTV Virus Taxonomy Profile: Geminiviridae. *The Journal of general virology* **98**, 131-133, doi:10.1099/jgv.0.000738 (2017).
92 Rosario, K. *et al.* Revisiting the taxonomy of the family Circoviridae: establishment of the genus Cyclovirus and removal of the genus Gyrovirus. *Arch Virol* **162**, 1447-1463, doi:10.1007/s00705-017-3247-y (2017).
93 Varsani, A. & Krupovic, M. Smacoviridae: a new family of animal-associated single-stranded DNA viruses. *Arch Virol* **163**, 2005-2015, doi:10.1007/s00705-018-3820-z (2018).
94 Díez-Villaseñor, C. & Rodriguez-Valera, F. CRISPR analysis suggests that small circular single-stranded DNA smacoviruses infect Archaea instead of humans. *Nature Communications* **10**, 294, doi:10.1038/s41467-018-08167-w (2019).
95 Khayat, R. *et al.* The 2.3-angstrom structure of porcine circovirus 2. *J Virol* **85**, 7856-7862, doi:10.1128/JVI.00737-11 (2011).
96 Tomaru, Y. *et al.* Isolation and characterization of a single-stranded DNA virus infecting Chaetoceros lorenzianus Grunow. *Appl Environ Microbiol* **77**, 5285-5293, doi:10.1128/aem.00202-11 (2011).
97 Quaiser, A., Krupovic, M., Dufresne, A., Francez, A. J. & Roux, S. Diversity and comparative genomics of chimeric viruses in Sphagnum-dominated peatlands. *Virus Evol* **2**, vew025, doi:10.1093/ve/vew025 (2016).
98 Kazlauskas, D. *et al.* Evolutionary history of ssDNA bacilladnaviruses features horizontal acquisition of the capsid gene from ssRNA nodaviruses. *Virology* **504**, 114-121, doi:10.1016/j.virol.2017.02.001 (2017).
99 Agranovsky, A. A., Lesemann, D. E., Maiss, E., Hull, R. & Atabekov, J. G. "Rattlesnake" structure of a filamentous plant RNA virus built of two capsid proteins. *Proc Natl Acad Sci U S A* **92**, 2470-2473 (1995).
100 Schulz, F. *et al.* Giant viruses with an expanded complement of translation system components. *Science* **356**, 82-85, doi:10.1126/science.aal4657 (2017).
101 Makino, D. L., Larson, S. B. & McPherson, A. The crystallographic structure of Panicum Mosaic Virus (PMV). *J Struct Biol* **181**, 37-52, doi:10.1016/j.jsb.2012.10.012 (2013).
102 Greninger, A. L. & DeRisi, J. L. Draft Genome Sequences of Ciliovirus and Brinovirus from San Francisco Wastewater. *Genome Announc* **3**, doi:10.1128/genomeA.00651-15 (2015).

103 Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature protocols* **10**, 845-858, doi:10.1038/nprot.2015.053 (2015).
104 Falero, A. *et al.* DNA binding proteins of the filamentous phages CTXphi and VGJphi of Vibrio cholerae. *J Bacteriol* **191**, 5873-5876, doi:10.1128/JB.01206-08 (2009).
105 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**, 772-780, doi:10.1093/molbev/mst010 (2013).
106 Buck, C. B. *et al.* The Ancient Evolutionary History of Polyomaviruses. *PLoS Pathog* **12**, e1005574, doi:10.1371/journal.ppat.1005574 (2016).
107 King, A. M. Q. *et al.* Changes to taxonomy and the International Code of Virus Classification and Nomenclature ratified by the International Committee on Taxonomy of Viruses (2018). *Arch Virol*, doi:10.1007/s00705-018-3847-1 (2018).
108 Paez-Espino, D. *et al.* IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Res* **47**, D678-D686, doi:10.1093/nar/gky1127 (2019).
109 Waldor, M. K. & Mekalanos, J. J. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* **272**, 1910-1914 (1996).
110 Hodyra-Stefaniak, K. *et al.* Mammalian Host-Versus-Phage immune response determines phage fate in vivo. *Sci Rep* **5**, 14802, doi:10.1038/srep14802 (2015).
111 Steel, O. *et al.* Circular replication-associated protein encoding DNA viruses identified in the faecal matter of various animals in New Zealand. *Infect Genet Evol* **43**, 151-164, doi:10.1016/j.meegid.2016.05.008 (2016).
112 Roux, S. *et al.* Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses. *Nat Commun* **4**, 2700, doi:10.1038/ncomms3700 (2013).
113 Krupovic, M. *et al.* Multiple layers of chimerism in a single-stranded DNA virus discovered by deep sequencing. *Genome Biol Evol* **7**, 993-1001, doi:10.1093/gbe/evv034 (2015).
114 Kauffman, K. M. *et al.* A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. *Nature* **554**, 118-122, doi:10.1038/nature25474 (2018).
115 Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* **31**, 3350-3352, doi:10.1093/bioinformatics/btv383 (2015).
116 Chalkias, S. *et al.* ViroFind: A novel target-enrichment deep-sequencing platform reveals a complex JC virus population in the brain of PML patients. *PLoS One* **13**, e0186945, doi:10.1371/journal.pone.0186945 (2018).
117 Geoghegan, E. M. *et al.* Infectious Entry and Neutralization of Pathogenic JC Polyomaviruses. *Cell Rep* **21**, 1169-1179, doi:10.1016/j.celrep.2017.10.027 (2017).
118 Hunt, M. *et al.* Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol* **16**, 294, doi:10.1186/s13059-015-0849-0 (2015).
119 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150-3152, doi:10.1093/bioinformatics/bts565 (2012).
120 Pei, J. & Grishin, N. V. PROMALS3D: multiple protein sequence alignment enhanced with evolutionary and three-dimensional structural information. *Methods Mol Biol* **1079**, 263-271, doi:10.1007/978-1-62703-646-7_17 (2014).
121 Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**, 268-274, doi:10.1093/molbev/msu300 (2015).
122 Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* **47**, W256-W259, doi:10.1093/nar/gkz239 (2019).

123 Buck, C. B., Pastrana, D. V., Lowy, D. R. & Schiller, J. T. Efficient intracellular assembly of papillomaviral vectors. *J Virol* **78**, 751-757 (2004).
124 Gilbert, J. A. *et al.* Current understanding of the human microbiome. *Nat Med* **24**, 392-400, doi:10.1038/nm.4517 (2018).
125 Shkoporov, A. N. *et al.* The Human Gut Virome Is Highly Diverse, Stable, and Individual Specific. *Cell Host Microbe* **26**, 527-541 e525, doi:10.1016/j.chom.2019.09.009 (2019).
126 Qin, J. *et al.* A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55-60, doi:10.1038/nature11450 (2012).
127 Zhang, X. *et al.* The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat Med* **21**, 895-905, doi:10.1038/nm.3914 (2015).
128 Feng, Q. *et al.* Gut microbiome development along the colorectal adenoma-carcinoma sequence. *Nat Commun* **6**, 6528, doi:10.1038/ncomms7528 (2015).
129 Karlsson, F. H. *et al.* Gut metagenome in European women with normal, impaired and diabetic glucose control. *Nature* **498**, 99-103, doi:10.1038/nature12198 (2013).
130 Jie, Z. *et al.* The gut microbiome in atherosclerotic cardiovascular disease. *Nat Commun* **8**, 845, doi:10.1038/s41467-017-00900-1 (2017).
131 Dubinkina, V. B. *et al.* Links of gut microbiota composition with alcohol dependence syndrome and alcoholic liver disease. *Microbiome* **5**, 141, doi:10.1186/s40168-017-0359-2 (2017).
132 Lwoff, A. Lysogeny. *Bacteriol Rev* **17**, 269-337 (1953).
133 Gamage, S. D., Patton, A. K., Hanson, J. F. & Weiss, A. A. Diversity and host range of Shiga toxin-encoding phage. *Infect Immun* **72**, 7131-7139, doi:10.1128/IAI.72.12.7131-7139.2004 (2004).
134 Schuch, R. & Fischetti, V. A. Detailed genomic analysis of the Wbeta and gamma phages infecting Bacillus anthracis: implications for evolution of environmental fitness and antibiotic resistance. *J Bacteriol* **188**, 3037-3051, doi:10.1128/JB.188.8.3037-3051.2006 (2006).
135 Fridman, S. *et al.* A myovirus encoding both photosystem I and II proteins enhances cyclic electron flow in infected Prochlorococcus cells. *Nat Microbiol* **2**, 1350-1357, doi:10.1038/s41564-017-0002-9 (2017).
136 Howard-Varona, C. *et al.* Phage-specific metabolic reprogramming of virocells. *ISME J* **14**, 881-895, doi:10.1038/s41396-019-0580-z (2020).
137 Al-Shayeb, B. *et al.* Clades of huge phages from across Earth's ecosystems. *Nature* **578**, 425-431, doi:10.1038/s41586-020-2007-4 (2020).
138 Gilbert, S. F. A holobiont birth narrative: the epigenetic transmission of the human microbiome. *Front Genet* **5**, 282, doi:10.3389/fgene.2014.00282 (2014).
139 Yang, C. *et al.* Fecal IgA Levels Are Determined by Strain-Level Differences in Bacteroides ovatus and Are Modifiable by Gut Microbiota Manipulation. *Cell Host Microbe* **27**, 467-475 e466, doi:10.1016/j.chom.2020.01.016 (2020).
140 Gil-Cruz, C. *et al.* Microbiota-derived peptide mimics drive lethal inflammatory cardiomyopathy. *Science* **366**, 881-886, doi:10.1126/science.aav3487 (2019).
141 Wang, K. *et al.* Parabacteroides distasonis Alleviates Obesity and Metabolic Dysfunctions via Production of Succinate and Secondary Bile Acids. *Cell Rep* **26**, 222-235 e225, doi:10.1016/j.celrep.2018.12.028 (2019).
142 Shkoporov, A. N. *et al.* PhiCrAss001 represents the most abundant bacteriophage family in the human gut and infects Bacteroides intestinalis. *Nat Commun* **9**, 4781, doi:10.1038/s41467-018-07225-7 (2018).
143 Martinez-Rubio, R. *et al.* Phage-inducible islands in the Gram-positive cocci. *ISME J* **11**, 1029-1042, doi:10.1038/ismej.2016.163 (2017).

144 Kotloff, K. L. *et al.* Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *Lancet* **382**, 209-222, doi:10.1016/S0140-6736(13)60844-2 (2013).
145 Potter, S. C. *et al.* HMMER web server: 2018 update. *Nucleic Acids Res* **46**, W200-W204, doi:10.1093/nar/gky448 (2018).
146 Obbard, D. J., Shi, M., Roberts, K. E., Longdon, B. & Dennis, A. B. A new lineage of segmented RNA viruses infecting animals. *Virus Evol* **6**, vez061, doi:10.1093/ve/vez061 (2020).
147 Vanaja, S. K. *et al.* Bacterial Outer Membrane Vesicles Mediate Cytosolic Localization of LPS and Caspase-11 Activation. *Cell* **165**, 1106-1119, doi:10.1016/j.cell.2016.04.015 (2016).
148 Sears, C. L. *et al.* Association of enterotoxigenic Bacteroides fragilis infection with inflammatory diarrhea. *Clin Infect Dis* **47**, 797-803, doi:10.1086/591130 (2008).
149 Paez-Espino, D., Pavlopoulos, G. A., Ivanova, N. N. & Kyrpides, N. C. Nontargeted virus sequence discovery pipeline and virus clustering for metagenomic data. *Nature protocols* **12**, 1673-1682, doi:10.1038/nprot.2017.063 (2017).
150 Tisza, M. J. *et al.* Discovery of several thousand highly diverse circular DNA viruses. *Elife* **9**, doi:10.7554/eLife.51971 (2020).
151 Lowe, T. M. & Chan, P. P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res* **44**, W54-57, doi:10.1093/nar/gkw413 (2016).
152 McNair, K., Zhou, C., Dinsdale, E. A., Souza, B. & Edwards, R. A. PHANOTATE: a novel approach to gene identification in phage genomes. *Bioinformatics* **35**, 4537-4542, doi:10.1093/bioinformatics/btz265 (2019).
153 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119, doi:10.1186/1471-2105-11-119 (2010).
154 Marchler-Bauer, A. *et al.* CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res* **45**, D200-D203, doi:10.1093/nar/gkw1129 (2017).
155 Gruning, B. *et al.* Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat Methods* **15**, 475-476, doi:10.1038/s41592-018-0046-7 (2018).
156 Manrique, P. *et al.* Healthy human gut phageome. *Proc Natl Acad Sci U S A* **113**, 10400-10405, doi:10.1073/pnas.1601060113 (2016).
157 Breitbart, M. *et al.* Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* **185**, 6220-6223, doi:10.1128/jb.185.20.6220-6223.2003 (2003).
158 Minot, S. *et al.* The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res* **21**, 1616-1625, doi:10.1101/gr.122705.111 (2011).
159 Minot, S. *et al.* Rapid evolution of the human gut virome. *Proc Natl Acad Sci U S A* **110**, 12450-12455, doi:10.1073/pnas.1300833110 (2013).
160 Gandon, S., Buckling, A., Decaestecker, E. & Day, T. Host-parasite coevolution and patterns of adaptation across time and space. *J Evol Biol* **21**, 1861-1866, doi:10.1111/j.1420-9101.2008.01598.x (2008).
161 Silveira, C. B. & Rohwer, F. L. Piggyback-the-Winner in host-associated microbial communities. *NPJ Biofilms Microbiomes* **2**, 16010, doi:10.1038/npjbiofilms.2016.10 (2016).
162 Tarafder, A. K. *et al.* Phage liquid crystalline droplets form occlusive sheaths that encapsulate and protect infectious rod-shaped bacteria. *Proc Natl Acad Sci U S A* **117**, 4724-4731, doi:10.1073/pnas.1917726117 (2020).
163 Nakatsu, G. *et al.* Alterations in Enteric Virome Are Associated With Colorectal Cancer and Survival Outcomes. *Gastroenterology* **155**, 529-541 e525, doi:10.1053/j.gastro.2018.04.018 (2018).
164 Abu-Ali, G. S. *et al.* Metatranscriptome of human faecal microbial communities in a cohort of adult men. *Nat Microbiol* **3**, 356-366, doi:10.1038/s41564-017-0084-4 (2018).

165 Li, D. *et al.* MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3-11, doi:10.1016/j.ymeth.2016.02.020 (2016).
166 Guerin, E. *et al.* Biology and Taxonomy of crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host Microbe* **24**, 653-664 e656, doi:10.1016/j.chom.2018.10.002 (2018).
167 Yutin, N. *et al.* Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nat Microbiol* **3**, 38-46, doi:10.1038/s41564-017-0053-y (2018).
168 Ma, Y., You, X., Mai, G., Tokuyasu, T. & Liu, C. A human gut phage catalog correlates the gut phageome with type 2 diabetes. *Microbiome* **6**, 24, doi:10.1186/s40168-018-0410-y (2018).
169 Sutton, T. D. S. & Hill, C. Gut Bacteriophage: Current Understanding and Challenges. *Front Endocrinol (Lausanne)* **10**, 784, doi:10.3389/fendo.2019.00784 (2019).
170 Shkoporov, A. N. *et al.* Reproducible protocols for metagenomic analysis of human faecal phageomes. *Microbiome* **6**, 68, doi:10.1186/s40168-018-0446-z (2018).
171 Le Chatelier, E. *et al.* Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541-546, doi:10.1038/nature12506 (2013).
172 Loomba, R. *et al.* Gut Microbiome-Based Metagenomic Signature for Non-invasive Detection of Advanced Fibrosis in Human Nonalcoholic Fatty Liver Disease. *Cell Metab* **30**, 607, doi:10.1016/j.cmet.2019.08.002 (2019).
173 Heintz-Buschart, A. *et al.* Integrated multi-omics of the human gut microbiome in a case study of familial type 1 diabetes. *Nat Microbiol* **2**, 16180, doi:10.1038/nmicrobiol.2016.180 (2016).
174 Li, J. *et al.* Gut microbiota dysbiosis contributes to the development of hypertension. *Microbiome* **5**, 14, doi:10.1186/s40168-016-0222-x (2017).
175 Wen, C. *et al.* Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biol* **18**, 142, doi:10.1186/s13059-017-1271-6 (2017).
176 Bedarf, J. R. *et al.* Functional implications of microbial and viral gut metagenome changes in early stage L-DOPA-naive Parkinson's disease patients. *Genome Med* **9**, 39, doi:10.1186/s13073-017-0428-y (2017).
177 Qin, N. *et al.* Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59-64, doi:10.1038/nature13568 (2014).
178 Debelius, J. *et al.* Tiny microbes, enormous impacts: what matters in gut microbiome studies? *Genome Biology* **17**, doi:ARTN 217
10.1186/s13059-016-1086-x (2016).
179 Sawilowsky, S. S. New Effect Size Rules of Thumb. *J Mod Appl Stat Meth* **8**, 597-599, doi:DOI 10.22237/jmasm/1257035100 (2009).
180 Chen, S. F., Zhou, Y. Q., Chen, Y. R. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, 884-890, doi:10.1093/bioinformatics/bty560 (2018).
181 Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **Chapter 4**, Unit 4 10, doi:10.1002/0471250953.bi0410s05 (2004).
182 Ho, B. T., Dong, T. G. & Mekalanos, J. J. A view to a kill: the bacterial type VI secretion system. *Cell Host Microbe* **15**, 9-21, doi:10.1016/j.chom.2013.11.008 (2014).
183 McDaniel, L. D. *et al.* High frequency of horizontal gene transfer in the oceans. *Science* **330**, 50, doi:10.1126/science.1192243 (2010).
184 Arrieta, M. C., Walter, J. & Finlay, B. B. Human Microbiota-Associated Mice: A Model with Challenges. *Cell Host Microbe* **19**, 575-578, doi:10.1016/j.chom.2016.04.014 (2016).

Intended to be blank

Curriculum Vitae

Michael Joseph Tisza, Ph.D. Candidate

Phone: (224) 305-4403
Email: michael.tisza@gmail.com
Github: mtisza1

Johns Hopkins University
Cellular, Molecular and Developmental Biology Program
3400 N. Charles St.
Baltimore, MD 21218

National Institutes of Health, National Cancer Institute
Lab of Cellular Oncology, Tumor Virus Molecular Biology Section
9000 Rockville Pike
Bldg 37, Room 4118
Bethesda, MD, 20892-4263

Education

Ph.D. Candidate, Johns Hopkins University, Baltimore, MD August 2014 – August 2020
NIH Graduate Partnership Program Fellow
Mentor: Christopher Buck, Ph.D.

B.A., Augustana College, Rock Island, IL 2008 – 2012
Biology and English Literature
Magna Cum Laude, Cumulative GPA: 3.76

Research Experience

Ph.D. Candidate, National Cancer Institute, Bethesda, MD 2015 - present
Discovering viruses in genetic dark matter
Advisor: Chris B. Buck, Ph.D.
In the last decades, the importance of the microbiome in human health and diseases has been made clear. However, the microbiome constituents and their effects have primarily been described in regard to bacterial species, while the viral, archaeal and eukaryotic components are often ignored. Partly due to this lack of emphasis, the extremely diverse and numerous "virome" is mostly unknown and unannotated, while its effects on humans are unexplored. I have published work looking at the circular DNA virome of humans and animals, contributing over 2,500 complete, high quality virus reference genomes, many establishing new families and some "dark matter" sequences being so divergent that they lack primary amino acid similarity to any previously known sequence. Confirmatory wet bench experiments confirmed that several dark matter sequences encoded viral capsid proteins. Beyond this, I developed Cenote-Taker, a bioinformatics package to identify and annotate viral sequences in any metagenome. Applying Cenote-Taker to thousands of human metagenomes has identified and thoroughly annotated over 86,000 unique complete or near-complete viral genomes, the large majority of which were previously undescribed. Work to determine any disease or health association of these viruses is ongoing.

Research Assistant, University of Texas Health Science Center, Houston, TX 2012 – 2014
Targeting plasma membrane properties in breast cancer
Advisor: Jeffrey T. Chang, Ph.D.

Metastatic disease in breast cancer is responsible for ~90% of mortality, and metastasis can be facilitated by a latent cellular developmental program called the Epithelial to Mesenchymal Transition (EMT). However, no therapies currently target EMT. I performed *in vitro* and *in vivo* experiments to discover how signaling pathways required for EMT are controlled by biophysical properties of the plasma membrane and can be derailed with drug therapies.

Academic Honors and Funding

NIH Graduate Student Research Award, 2017
Awarded for outstanding project as presented at the Graduate Student Symposium

NIH IRTA Pre-Doctoral Fellowship Recipient, 2014 - present
Funding for 6 years of Ph.D. training and opportunity to conduct dissertation work at the National Institutes of Health

Academic All-American in NCAA Cross Country/Track in 2012
National honor given by CoSIDA to select group of NCAA athletes who excel in the classroom and their sport

Member of Phi Beta Kappa National Academic Honor Society
2012 Inductee

2012 Knut Erickson Scholar Athlete Award
4-year Varsity Letter Winner at Augustana College with highest GPA

Two-time CCIW Jack Schwartz Award for excellent Academics and Athletics in 2011
Awarded to one athlete displaying exemplary scholarship and athletic success at Augustana College

Finalist for NCAA Post-Graduate Scholarship Award, Fall 2011
Award given to NCAA athlete pursuing graduate education

Invited Talks

Tisza MJ. Cenote-Taker2 Democratizes Virus Discovery and Annotation from Metagenomic Data. The Global Virome in Health and Disease Keystone Meeting, Tahoe City, CA. March 3, 2020

Tisza MJ. A New Class of Virus-Like Elements Is Highly Prevalent in the Human Gut. American Society for Virology Annual Meeting, College Park, MD, July 13, 2018

Tisza MJ. Shining a Spotlight on DNA Dark Matter in Animal Viromes. American Society for Virology Annual Meeting, Madison, WI, June 28, 2017

Publications

Upcoming

1. **Tisza MJ**, Miller B, Belford AK, Buck CB. Tens of thousands of unique viruses from human metagenomes and their disease associations. *In preparation.*

2. **Tisza MJ**, Tanase L, Krishnamurthy S, Stine C, Pop M, Nasko D, Commichaux S, Hill C, Shkoporov A, Wang D, Buck CB, Almeida M. Bibiviruses are a new, unusual virus family common in the human gut. *In preparation.*

3. **Tisza MJ**, Belford AK, Bulduc B, Sullivan M, Starrett GJ, Buck CB. Cenote-Taker2 Democratizes Virus Discovery and Annotation from Metagenomic Data. *In preparation.*

4. Starrett GJ*, **Tisza MJ***, Welch NL, Belford AK, Peretti A, Pastrana DV, Buck CB. Adintoviruses: An Animal-Tropic Family of Midsize Eukaryotic Linear dsDNA (MELD) Viruses. *Under review.* Preprint: https://www.biorxiv.org/content/10.1101/697771v2
* Co-first authors

5. Welch NL, **Tisza MJ**, Starrett GJ, Belford AK, Pastrana DV, Pang YS, Schiller JT, An P, Cantalupo PG, Pipas JM, Koda S, Subramaniam K, Waltzek TB, Bian C, Shi Q, Ruan Z, Ng TF, Buck CB. Identification of Adomavirus Virion Proteins. *In preparation.* Preprint: https://www.biorxiv.org/content/10.1101/341131v1

Peer-reviewed

1. Uritsky G, **Tisza MJ**, Gelsinger DR, Munn A, Taylor J, DiRuggiero J. Cellular life from the three domains and viruses are transcriptionally active in a hypersaline desert community. *Environmental Microbiology*. 2020. doi: 10.1111/1462-2920.15023

2. Malki K, Rosario K, Sawaya NA, Székely AJ, **Tisza MJ**, Breitbart M. Prokaryotic and viral community composition of freshwater springs in Florida, USA. *mBio*. 2020; doi: 10.1128/mBio.00436-20.

3. **Tisza MJ**, Pastrana DV, Welch NL, Stewart B, Peretti A, Starrett GJ, Pang YS, Krishnamurthy SR, Pesavento PA, McDermott DH, Murphy PM, Whited JL, Miller B, Brenchley J, Rosshart SP, Rehermann B, Doorbar J, Ta'ala BA, Pletnikova O, Troncoso JC, Resnick SM, Bolduc B, Sullivan MB, Varsani A, Segall AM, Buck CB. Discovery of several thousand highly diverse circular DNA viruses. *eLife*. 2020; doi: https://doi.org/10.7554/eLife.51971 (**Top 3%** downloaded in bioRxiv microbiology preprints, all time)

4. Connor R, Brister R, Buchmann JP, Deboutte W, Edwards R, Martí-Carreras J, **Tisza MJ**, Zalunin V, Andrade-Martínez J, Cantu A, D'Amour M, Efremov A, Fleischmann L, Forero-Junco L, Garmaeva S, Giluso M, Glickman C, Henderson M, Kellman B, Kristensen D, Leubsdorf C, Levi K, Levi S, Pakala S, Peddu V, Ponsero A, Ribeiro E, Roy F, Rutter L, Saha S, Shakya M, Shean R, Miller M, Tully B, Turkington C, Youens-Clark K, Vanmechelen B, Busby B. NCBI's Virus Discovery Hackathon: Engaging Research Communities to Identify Cloud Infrastructure Requirements. *Genes*. 2019; doi: 10.3390/genes10090714

5. **Tisza MJ**, Yuan H, Schlegel R, Buck CB. Genomic Sequence of Canine Papillomavirus 19. *Genome Announcements*. 2016; doi: 10.1128/genomeA.01380-16

6. **Tisza MJ*,** Zhao W*, Fuentes JSR*, Prijic S, Chen X, Levental I, Chang JT. Motility and stem cell properties induced by the epithelial-mesenchymal transition require destabilization of lipid rafts. *Oncotarget*. 2016; doi:10.18632/oncotarget.9928
* Co-first authors

7. Buck CB, Van Doorslaer K, Peretti A, Geoghegan EM, **Tisza MJ**, An P, Katz JP, Pipas JM, McBride AA, Camus AC, McDermott AJ, Dill JA, Delwart E, Ng TFF, Farkas K, Austin C, Kraberger S, Davison W, Pastrana DV, Varsani A. The Ancient Evolutionary History of Polyomaviruses. *PLOS Pathogens*. 2016; 12 (4) e1005574

8. Zhao W., Prijic S., Urban B., **Tisza MJ.**, Sjol J., Zhi T., S., Sendurai M., and Chang J. Candidate Antimetastasis Drugs Suppress the Metastatic Capacity of Breast Cancer Cells by Reducing Membrane Fluidity. *Cancer research*. 2016; 76 (7), 2037-2049

9. Wierzchos J, DiRuggiero J, Vitek P, Artieda O, Souza-Egipsy V, **Tisza MJ**, Davila A, Vilchez C, Garbayo I, Ascaso C. Adaptation strategies of endolithic chlorophototrophs to survive the hyperarid and extreme solar radiation environment of the Atacama Desert. *Frontiers in Microbiology*. 2015; (6)

Teaching and Mentoring Experience

Lecturer/Teaching Assistant, National Institutes of Health, Bethesda, MD Fall 2015

Research Tools for Studying Disease

In this position, I delivered 2 two-hour lectures on molecular genetics and molecular biology theory to a class of 30 post-baccalaureate students. In these lectures, I used interactive technology, in-class quizzes, as well as traditional lecturing that incorporated modern pedagogy. I also served as a TA for the semester, creating and grading homework and tests.

Teaching Assistant, Johns Hopkins University, Baltimore, MD Spring 2015

Laboratory Section of Biology 102

In this position, I taught a weekly lab section of 25 undergraduates college-level molecular, physiological, and evolutionary biology concepts in a hands-on setting.

Team lead, NCBI Virus Hackathon 1 January 2019

Known Virus team

I designed a project to assess high-confidence viral information from petabytes of metagenomic sequencing information. Then I led a seven-person team to accomplish this

Team lead, NCBI Virus Hackathon 2 November 2019

Virus graph team

I led an eight-person team through designing software and visualizing important genetic differences between different strains of related viral genomes.