Title

Discovery of several thousand highly diverse circular DNA viruses

Authors

Michael J. Tisza[1], Diana V. Pastrana[1], Nicole L. Welch[1], Brittany Stewart[1], Alberto Peretti[1], Gabriel J. Starrett[1], Yuk-Ying S. Pang[1], Arvind Varsani[2,3], Siddharth R. Krishnamurthy[4], Patricia A. Pesavento[5], David H. McDermott[6], Philip M. Murphy[6], Jessica L. Whited [7,8,9], Bess Miller [7,8], Jason M. Brenchley[10], Stephan P. Rosshart[11], Barbara Rehermann[11], John Doorbar[12], Blake A. Ta'ala[13], Olga Pletnikova[14], Juan Troncoso[14], Susan M. Resnick[15], Anca M. Segall[16], Christopher B. Buck[1]*

Affiliations

[1]Lab of Cellular Oncology, NCI, NIH, Bethesda, MD 20892-4263
[2]The Biodesign Center of Fundamental and Applied Microbiomics, School of Life Sciences, Center for Evolution and Medicine, Arizona State University, 1001 S. McAllister Ave, Tempe, AZ 85287-5001, USA
[3]Structural Biology Research Unit, Department of Clinical Laboratory Sciences, University of Cape Town, Rondebosch, 7701, Cape Town, South Africa.
[4]Metaorganism Immunity Section, Laboratory of Immune System Biology, NIAID, NIH, Bethesda, MD 20892-4263
[5]Department of Pathology, Microbiology, and Immunology, University of California, Davis, CA 95616, USA.
[6]Molecular Signaling Section, Laboratory of Molecular Immunology, NIAID, Bethesda, Maryland, USA.
[7]Department of Orthopedic Surgery, Harvard Medical School, The Harvard Stem Cell Institute, Brigham and Women's Hospital, 60 Fenwood Road, Boston, MA, 02115, USA.
[8]Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, MA, 02142, USA.
[9]Department of Stem Cell and Regenerative Biology, Harvard University, 7 Divinity Avenue, Cambridge, MA, 02138, USA.
[10]Barrier Immunity Section, Lab of Viral Diseases, NIAID, NIH, MD 20892
[11]Immunology Section, Liver Diseases Branch, NIDDK, NIH, Bethesda, MD 20892-1800
[12]Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge, CB2 1QP, UK.
[13]Mililani Mauka Elementary, Mililani, HI, 96789, USA.
[14]Department of Pathology (Neuropathology), Johns Hopkins University School of Medicine, Baltimore, MD, USA.
[15]Laboratory of Behavioral Neuroscience, National Institute on Aging, Baltimore, MD 21224, USA.
[16]Viral Information Institute & Department of Biology, San Diego State University, San Diego, CA, 92182-4614, USA.
*Corresponding author: buckc@mail.nih.gov

Abstract

Virologists have posited the existence of millions of distinct viral species, but fewer than 9000 viral species are catalogued in GenBank's RefSeq database. We selectively enriched for and amplified the genomes of circular DNA viruses in over 70 animal samples, ranging from cultured soil nematodes to human tissue specimens. Over 2500 complete circular genomes, each representing a new viral taxon, were assembled, thoroughly annotated, and deposited in GenBank. The new genomes belong to dozens of established and emerging viral families. Some of the genomes have unexpected gene compositions that appear to be the result of recombination between ssDNA viruses 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 and cell culture assays to validate that

some predicted capsid proteins formed virus-like particles. These data further the understanding of viral sequence diversity and allow for more comprehensive analysis of the virosphere.

Main

There has been a rush to utilize 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[1], 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[2,3]. Some of this dark matter undoubtedly consists of viral sequences, which have remained poorly characterized due to their enormous diversity[4]. Recent efforts have shown that our understanding of viral diversity, even of viruses known to directly infect humans, has been woefully incomplete[5]. To increase the power of future studies seeking to find additional associations between microbes and host animal phenotypes, representative sequences 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 RNAs) that can readily be identified through sequence analysis, viruses, as a whole, do not have any universally conserved sequence components[6,7]. 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[8]. Also, many bacteriophages of the order Caudovirales, such as the families Siphoviridae, Podoviridae, and Myoviridae, are extensively characterized due to their and their hosts' culturability and their detectability using viral plaque assays[9-11]. In contrast, many small DNA viruses are not easily cultured[12], use diverse genome replication strategies, and typically lack polymerase genes[13]. 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 particularly sparse in reference databases.

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. 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. In this study, we use physical and topological features common to many DNA viruses to uncover a diverse array of complete circular viral genomes in various human and non-human animal samples. A high-throughput bioinformatics pipeline was developed to classify and annotate over 2500 candidate viral genomes.

Results

<u>Virion enrichment and sequencing</u>

We have previously developed methods for discovery of viruses (primarily polyomaviruses and papillomaviruses) from skin swabs and complex tissue specimens[14]. The approach begins with detergent-mediated cell disruption and nuclease digestion of non-encapsidated nucleic acid. This is followed by ultracentrifugation of virions through Optiprep gradients. The purified virions are then subjected to protein digestion, and the encapsidated DNA is amplified by random-primed RCA using bacteriophage phi29 DNA polymerase, which selectively amplifies circular DNA templates and conducts second strand synthesis on ssDNA. The amplified DNA is then deep-sequenced, and reads are *de novo* assembled into contigs with SPAdes[15]. This approach was applied to a large set of human and animal samples (Table S1). The samples sources were varied, but emphasis was placed on using a variety of tissues from vertebrates and model organisms.

97 For the current study, we developed a bioinformatics pipeline, Cenote-Taker (a portmanteau of
98 *cenote*, a naturally occurring circular water pool, and *note-taker*), to identify and annotate *de novo*
99 assembled contigs with terminal direct repeats consistent with circular DNA molecules. In this pipeline,
100 putative circular sequences were queried against GenBank's nucleotide database using BLASTN to
101 remove circles with extensive nucleotide identity (>90% across any 500 bp window) to known
102 sequences. Sequences with >90% identity to previously reported viral sequences represented less than
103 1.5% of circular contigs. Open reading frames (ORFs) from remaining unidentified circular DNA
104 sequences >240 nucleotides (nt) in length were translated and used for RPS-BLAST queries of GenBank's
105 Conserved Domain Database. ORFs that did not yield E values better than 1×10^{-4} in RPS-BLAST were
106 subjected to BLASTP searches of viral sequences in GenBank's nr database[16-18]. ORFs that were not
107 confidently identified in BLAST searches were subjected to searches using HHBlits[19]. The results were
108 used to annotate and name each sequence in a human-readable genome map as well as a format
109 suitable for submission to GenBank. After checking the Cenote-Taker output of each genome, minor
110 revisions were made and files were submitted to GenBank (BioProjects PRJNA393166 and
111 PRJNA396064).
112
113 Characterization of DNA viruses in animal metagenomes
114 Of the novel circular sequences detected in the survey, 1844 encode genes with similarity to proteins of
115 ssDNA viruses and 55 encode genes with similarity to dsDNA viral proteins (Fig. 1A). The large majority
116 of genomes from this study are highly divergent from previously reported sequences (Fig. S1). The most
117 common taxon was the viral family *Microviridae*, a class of small bacteriophages, with 670 complete
118 genomes. This represents a dramatic expansion beyond the 459 non-redundant microvirus genomes
119 previously listed in GenBank (of which 44 were listed in the RefSeq database) (Fig. 1B). We discovered
120 562 genomes that had similarity to Circular Rep-Encoding Single-Stranded (CRESS) viruses that have yet
121 to be assigned into specific families, i.e. unclassified CRESS viruses (compared to the 85 listed in RefSeq).
122 Three hundred genomes displayed recognizable similarity to members of the *Circoviridae*, a class of
123 animal-tropic CRESS viruses, adding to the 427 non-redundant circovirus genomes previously listed in
124 GenBank (167 genomes were listed in the RefSeq database). Other genomes that were uncovered
125 represent *Anelloviridae* (n=170), *Inoviridae* (n=70), *Genomoviridae* (n=58), *Siphoviridae* (n=18),
126 unclassified phage (n=14), unclassified virus (n=6), *Geminiviridae* (n=8), *Podoviridae* (n=10), *Myoviridae*
127 (n=7), *Papillomaviridae* (n=4), and *CrAssphage* (n=2) (Fig. 1B).
128 It is difficult to assign a host to most of the viruses from this study due to their divergence from
129 known viral sequences. However, three viruses had exact matches to CRISPR spacers in bacterial
130 genomes (Siphoviridae sp. isolate ctcj11:Shewanella sp. W3-18-1, Inoviridae sp. isolate ctce6:
131 Shewanella baltica OS195, Microviridae sp. isolate ctbe523:Paludibacter propionicigenes WB4) and one
132 virus had an exact match to the CRISPR spacer of an archaeon (Caudovirales sp. isolate
133 cthg227:Methanobrevibacter sp. AbM4), implying that these organisms are infected by these viruses.
134 In addition to circular genomes with recognizable similarity to known viruses, 609 circular
135 contigs appeared to represent elements that lacked discernable similarity to known viruses (Fig. 1A).
136
137 The vast majority of the *de novo* assembled circular genomes were <10 kb in length (Fig S2). This
138 reflects the tendency of ssDNA genomes to be small (the largest reported is ~10 kb)[7]. However, this
139 observation may also be due to the greater difficulty in *de novo* assembling complete gapless contigs for
140 larger genomes, as large genomes require more reads for uninterrupted coverage and are more likely to
141 contain difficult-to-assemble regions. Further, RCA, which we used to amplify viral genomes, may
142 preferentially amplify smaller circular genomes. Despite these technical obstacles, our detection of a
143 new tailed bacteriophage with a 419kb genome (Myoviridae sp. isolate ctbc4, GenBank Accession:

144 MH622943), along with 45 other >10 kb circular sequences (Fig S2), indicates that the methods used for
145 the current work can detect large viral genomes.
146 There has been a recent renewal of interest in the hypothesis that viruses may be etiologically
147 associated with degenerative brain diseases, such as Alzheimer's disease[20,21]. Samples of brain tissue
148 from individuals who died of Alzheimer's disease (n=6) and other forms of dementia (n=6) were
149 subjected to virion enrichment and deep sequencing. Although complete or partial genomes of known
150 papillomaviruses, Merkel cell polyomavirus, and/or anelloviruses were observed in some samples
151 (Supplementary File S1), no novel complete viral genomes were recovered (Table S1). No recognizably
152 virus-like sequences were observed in a follow-up RNAseq analysis of the samples. It is likely that the
153 observed viral DNA sequences represent viral particles from blood vessels in these tissues or from
154 environmental sources. It is important to note that the methods used in the present study focused on
155 virion enrichment and thus would not detect latent or past viral infections.
156
157 Assignment of recognizable sequences to clusters
158 Sequence similarity network analyses were used to display related groups of genes from viral genomes
159 (Fig. 2, Fig S3). Protein sequences from RefSeq (ssDNA viruses) and from viral genomes discovered here
160 were aligned pairwise (each sequence was compared to all other sequences) using EFI-EST[22]. The output
161 was visualized with Cytoscape[23], with each node representing an amino acid sequence. Nodes
162 (sequences) with significant amino acid identity (E value $<1\times10^{-5}$) were connected with lines. This
163 method generated dozens of clusters of related sequences, one of which is displayed in Fig. 2. Note that
164 not all sequences in a given cluster necessarily have a significant similarity to all other sequences in that
165 cluster. Rather, many sequences have a second- or third-degree relatedness to one another. The nodes
166 added by this study populate and provide greater connectivity in the networks for many viral protein
167 families.
168
169 Large CRESS viruses
170 Although no single family of viruses accounts for the majority of genomes in this study, these results
171 expand the knowledge of the vast diversity of CRESS viruses, which appear to be ubiquitous among
172 eukaryotes[24-27]. The CRESS virus supergroup, which is primarily defined by a characteristic superfamily 3
173 DNA helicase (Rep) protein, encompasses the recognized families *Circoviridae*, *Geminiviridae*, and many
174 others. Characterized CRESS viruses have small icosahedral virions (<25 nm in diameter) with a simple
175 T=1 geometry[28]. This capsid architecture likely limits genome size, as nearly all previously reported
176 CRESS genomes/genome segments are under 3.5 kb. Exceptions to this size rule are bacilladnaviruses,
177 which have 4.5 - 6 kb genomes[29] and cruciviruses, which have 3.5 - 5.5 kb genomes[30]. Interestingly, the
178 genomes of these larger CRESS viruses encode capsid genes that appear to have been acquired
179 horizontally from RNA viruses[31]. In our dataset, eight CRESS-like circular genomes exceed 6 kb in length
180 (Fig. 3C-J), making them the largest CRESS genomes ever identified. Further, this study's large CRESS
181 genomes are apparently attributable to several independent acquisitions of capsid genes from other
182 taxa and/or capsid gene duplication events.
183 Notably, a large CRESS genome (CRESS virus isolate ctdh33, associated with rhabditid
184 nematodes that were serially cultured from a soil sample) encoded three separate genes with predicted
185 homology to the capsid gene of satellite tobacco necrosis virus (STNV, an RNA virus satellite) (Fig 3C).
186 The STNV capsid protein, in turn, exhibits similarity to geminivirus capsid proteins[13]. The three predicted
187 STNV capsid homologs in the nematode virus are highly divergent from one another, with only 28-30%
188 amino acid similarity. A possible explanation for this observation is that the capsid gene array is the
189 result of gene duplication events. Similar STNV capsid-like genes have previously been observed as a
190 single copy in unclassified <3 kb CRESS viruses[32] and several additional examples of such genes were

191 detected in the present study. Network analysis of new sequences suggests a possible model in which
192 the STNV capsid gene originated from an ancestral CRESS DNA virus (Fig S4).
193 To further characterize the soil nematode virus, the three candidate capsid proteins were
194 overexpressed in combination in human-derived 293TT cells[33]. Particles apparently composed of all
195 three capsid proteins were able to sediment in Optiprep gradients. Further, analysis by negative stain
196 electron microscopy (EM) revealed ~50 nm particles (Fig S5A). Interestingly these particles were found
197 to contain nonspecific nuclease-resistant RNA (Fig S5B). This supports the concept that CRESS virus
198 capsid proteins might theoretically be re-targeted to packing RNA virus genomes or vice versa.
199 CRESS genomes ctba10, ctcc19, ctbj26, ctcd34, and ctbd1037 (ranging from 3.5 - 6.2 kb in
200 length) also each encode two divergent capsid gene homologs (Fig. 3D, Fig S6). Single genomes encoding
201 multiple capsid genes with related but distinct amino acid sequences have been observed in RNA
202 viruses[34], but we believe that this is the first time it has been reported in ssDNA viruses.
203 Two related large CRESS viruses (ctdb796 and ctce741) encode capsid proteins similar to those
204 of bacilladnaviruses (Fig 3G, I). Interestingly, the Rep genes of the two viruses do not show detectable
205 similarity to known bacilladnavirus Reps and are instead similar to the Reps of certain unclassified CRESS
206 viruses, suggesting that CRESS ctdb796 and CRESS ctce741 are representatives of a new hybrid CRESS
207 virus family. Two other CRESS virus genomes (ctca5 and ctgh4) encode capsid genes that show amino
208 acid similarity to distinct groups of icosahedral T=3 ssRNA virus capsids[35] (tombus- and tombus-like
209 viruses) (Fig 3A,F, Fig. S7). Further, a 6.6 kb CRESS virus (ctbd466) (Fig 3H) was found to encode a gene
210 with similarity to the capsid region of the polyprotein of a newly described ssRNA virus, ciliovirus[35,36].
211 Protein fold predictor Phyre2[37] showed a top hit (58% confidence) for the capsid protein of a norovirus
212 (ssRNA virus with T=3 icosahedral capsid). Taken together, these findings suggest multiple instances of
213 chimerization between ssDNA and ssRNA viruses.
214 Two CRESS genomes (ctbb30 and ctbc27) from separate Rhesus macaque stool samples
215 combine CRESS Rep genes with several genes specific to *Inoviridae* (a family of circular ssDNA
216 bacteriophages), including inovirus-like capsid genes, which encode proteins that form a filamentous
217 virion (Fig 3B, J). Other inovirus-like genes encoded in these genomes include homologs of zonular
218 occludens toxin (ZOT, a packaging ATPase) and RstB (a DNA-binding protein required for host genome
219 integration)[38] (Fig. 3). Curiously, TBLASTX searches using CRESS ctbb30 and ctbc27 sequences yielded
220 large segments of similarity to various bacterial chromosomes (e.g., GenBank accession numbers
221 AP012044 and AP018536), presumably representing integrated prophages. This suggests that CRESS
222 ctbb30 and ctbc27 represent a previously undescribed prokaryote-infecting branch of the CRESS virus
223 supergroup.
224
225 Network analysis of genetic "dark matter"
226 We defined potential viral "dark matter" in the survey as circular contigs with no hits with E values
227 $<1\times10^{-5}$ in BLASTX searches of a database of viral and plasmid proteins. To categorize the predicted
228 proteins of the 609 dark matter elements, we used pairwise comparison with EFI-EST. A great majority
229 of translated gene sequences could be categorized into dark matter protein clusters (DMPCs) containing
230 four or more members (Fig. 4A). Further, groups of related dark matter elements (i.e. dark matter
231 genome groups (DMGGs)), much like viral families, could be delineated by the presence of a conserved,
232 group-specific marker gene. For example, DMPC1 can be thought of as the marker gene for DMGG1.
233 Certain DMPCs tend to co-occur on the same DMGG. For instance, DMPC7 and DMPC17 ORFs are always
234 observed in genomes with a DMPC1 ORF (i.e., DMGG1) (Fig. 4B). This *pro tempore* categorization
235 method is useful for visualizing the data, but we stress that is not necessarily taxonomically definitive.
236 Powerful new bioinformatic methods for detecting distant homology rely on the observation
237 that protein structures are typically more conserved than primary amino acid sequences. HHpred, for
238 example, compares the predicted fold of a query sequence to a library of known protein structures[39].

239 Alignments were prepared for each major DMPC to identify conserved residues and each alignment was
240 used for HHPRED queries. This approach identified that ten DMPCs are likely viral capsid proteins and 11
241 are rolling circle replicases (Fig. 4A).
242 While most of the circular dark matter in the survey could be characterized using these
243 methods, dark matter contigs represent a small remaining fraction in some samples (Fig. S9).
244
245 Functional analysis of candidate capsid proteins by expression in cell culture
246 In contrast to viral genes with conserved enzymatic functions, sequences of the capsid genes are
247 often poorly conserved, even within a given viral family[40]. Moreover, it appears that capsid proteins
248 have arisen repeatedly through capture and modification of different host cell proteins[31]. This makes it
249 challenging to detect highly divergent capsid proteins using alignment-based approaches. We therefore
250 turned to an alignment-independent approach known as iVireons, an artificial neural network trained by
251 comparing alignment-independent variables between a large set of known viral structural proteins and
252 known non-structural proteins[41] (https://vdm.sdsu.edu/ivireons/). As an example of the approach,
253 iVireons scores for DMPCs associated with DMGG1 are shown in Fig. 4C. Other sets of iVireons scores
254 can be seen in Fig. S8.
255 Of the 17 DMGGs for which HHPRED did not identify capsid genes, iVireons predicted that ten
256 contain at least one DMPC predicted to encode some type of virion structural protein (median score of
257 cluster >0.70). This allowed us to generate the testable hypothesis that some of these predicted
258 structural proteins would form virus-like particles (VLPs) if expressed in cell culture.
259 A subset of predicted capsid genes were expressed in *E. coli* and subjected to size exclusion
260 chromatography. Electron microscopic analysis of the proteins showed distinct, albeit irregular, VLPs for
261 three of these predicted capsid genes, whereas a negative control protein did not form particles (Fig. 5).
262 The results suggest that, in multiple cases, iVireons correctly predicted the identity of viral capsid
263 proteins.
264
265 **Discussion**
266 Massive parallel DNA sequencing surveys characterizing microbial communities typically yield a
267 significant fraction of reads that cannot be mapped to known genes. The present study sought to
268 provide the research community with an expanded catalog of viruses with circular DNA genomes
269 associated with humans and animals, as well as a means to characterize future datasets. We hope that
270 the availability of this expanded viral sequence catalog will facilitate future investigation into
271 associations between viral communities and disease states.
272 At present, GenBank's RefSeq database includes complete sequences for 8112 viral genomes,
273 most of which fit into 131 families recognized by the International Committee on Taxonomy of Viruses
274 (ICTV)[42]. The current study, which focused on circular DNA viruses with detergent-resistant capsids,
275 found 2508 new complete circular genomes which contribute new information and understanding to a
276 broad range of established and emerging taxa.
277 The predominance of small (<10kb) genomes with remote similarity to ssDNA viruses in our
278 dataset probably reflects the underappreciated ubiquity of small ssDNA viruses in nature. These viruses
279 are often overlooked in studies that only characterize sequences that are closely related to reference
280 genomes. In addition, ssDNA is not detected by some current DNA sequencing technologies unless
281 second-strand synthesis (such as the RCA approach used in the current study) is conducted.
282 While many of the viruses discovered in this study appear to be derived from prokaryotic
283 commensals, it is important to note that bacteriophages can contribute to human and animal diseases
284 by transducing toxins, antimicrobial resistance proteins, or genes that alter the physiology of their
285 bacterial hosts[43]. Furthermore, interaction between animal immune systems and bacteriophages
286 appears to be extensive[44].

287 Over 100 distinct human anellovirus sequences were found in human blood. Anelloviruses have
288 yet to be causally associated with any human disease, but this study indicates that we are likely still just
289 scratching the surface of the sequence diversity of human anelloviruses. It will be important to fully
290 catalog this family of viruses to address the field's general assumption that they are harmless.
291 Several of the CRESS viruses detected in this study are larger than any other CRESS DNA virus
292 genomes that have been described previously. In some cases, the larger size of these genomes may have
293 been enabled by a process involving capsid gene duplication events. Further, CRESS virus acquisition of
294 T=3 capsids from ssRNA *Nodaviridae* and *Tombusviridae* families has been previously suggested as the
295 origin of bacilladnaviruses and cruciviruses, respectively. We present evidence of additional
296 recombination events between CRESS viruses and ssRNA viruses, ssRNA satellites, and ssDNA
297 bacteriophages. These data suggest that CRESS viruses are at the center of a tangled evolutionary
298 history of viruses in which genomes change not just via gradual point mutations but also through larger
299 scale recombination and chimerization events.
300 It is likely that some dark matter sequences detected in this study share a common ancestor
301 with known viruses but are too divergent to retain discernable sequence similarity. In some cases, the
302 dark matter circles may represent a more divergent segment of a virus with a multipartite genome.
303 Alternatively, some of these sequences likely represent entirely new viral lineages that have not
304 previously been recognized.
305 This study suggests that, whether they are in our communities, in our homes, on our dinner
306 tables, or in us, there are likely hordes of undiscovered viruses still waiting to be seen.
307
308 **Methods**
309
310 Sample collection and sequencing
311
312 De-identified human swabs and tissue specimens were collected under the approval of various
313 Institutional Review Boards (Supplemental Table 1). Animal tissue samples were collected under the
314 guidance of various Animal Care and Use Committees.
315
316 Nematodes were cultured out of soil samples collected in Bethesda, Maryland, USA on OP50-Seeded
317 NGM-lite plates (C. elegans kit, Carolina Biological Supply).
318
319 Viral particles were concentrated by subjecting nuclease-digested detergent-treated lysate to
320 ultracentrifugation over an Optiprep step gradient, as previously described
321 https://ccrod.cancer.gov/confluence/display/LCOTF/Virome[14]. DNA from individually collected fractions
322 of the gradient was amplified by RCA using phi29 polymerase (TempliPhi, GE Health Sciences). While we
323 expected most viral particles to travel to the middle of the gradient based on previous experiments, RCA
324 was conducted on individual fractions spanning the gradient, in an attempt to detect viruses with
325 different biophysical properties. Pooled fractions were sequenced with Illumina technology on either
326 MiSeq or NextSeq500 sequencers. Contigs were assembled using SPAdes with the 'plasmid' setting.
327 Circularity was confirmed by assessing assembly graphs using Bandage[45].
328
329 Analysis of brain samples
330 Brain samples were initially analyzed by Optiprep gradient purification, RCA amplification, and deep
331 sequencing, as described above. JC polyomavirus, which has previously been reported in brain
332 samples[46], can display high buoyancy in Optiprep gradients[47]. Fractions from near the top of the
333 Optiprep gradient were subjected to an alternative method of virion enrichment using microcentrifuge
334 columns (Pierce) packed with 2 ml of Sepharose 4B Bead suspension exchanged into PBS. Fractions were

335 clarified at 5000 x g for 1 minute, and 200 µl of clarified extract was loaded onto the gel bed. The
336 column was spun at 735 x g and the eluate was digested with proteinase K, ethanol-precipitated, and
337 subjected to RCA. No additional viral sequences were detected by this method.
338
339 The brain samples were also subjected to confirmatory analysis by RNA sequencing. RNA was extracted
340 with Qiagen Lipid Tissue RNeasy Mini Kit and subjected to human ribosomal RNA depletion with Thermo
341 RiboMinus. The library was prepared with NEBNext Ultra™ II Directional RNA Library Prep Kit for Illumina
342 and subjected to massive parallel sequencing on the Illumina HiSeq platform.
343
344 Cenote-Taker, Virus Discovery and Annotation Pipeline
345 Cenote-Taker, a bioinformatics pipeline written for this project, was used for collection and detailed
346 annotation of each circular sequence. The flow of the program can be described as follows:
347 (1) Removes contigs (assembled with SPAdes) smaller than 1000 nts
348 (2) Predicts which contigs are circular based on overlapping ends
349 (3) Determines whether circular contig has any ORFs of 80 AA or larger or else discards sequence
350 (4) Uses BLASTN against GenBank 'nt' to disregard any circular sequences that are >90% identical to
351 known sequences across a >500 bp window
352 (5) Uses Circlator [48] to rotate circular contigs so that a non-intragenic start codon of one of the ORFs will
353 be the wrap point
354 (6) Uses BLASTX against a custom virus + plasmid database (derived from GenBank 'nr' and RefSeq) to
355 attempt to assign the circular sequence to a known taxon
356 (7) Translates each ORF of 80 AA or larger
357 (8) Uses RPS-BLAST to predict function of each ORF by aligning to known NCBI Conserved Domains
358 (9) Generates a tbl file of RPS-BLAST results
359 (10) Takes ORFs without RPS-BLAST hits and queries the GenBank 'nr viral' database with BLASTP
360 (11) Generates a tbl file of BLASTP results
361 (12) Takes ORFs without any BLASTP hits and queries HHblits (databases: uniprot20, pdb70, scop70,
362 pfam_31, NCBI_CD)
363 (13) Generates a tbl file of HHblits results
364 (14) Complies with a GenBank request to remove annotations for ORFs-within-ORFs that do not contain
365 conserved sequences
366 (15) Combines all tbl files into a master tbl file
367 (16) Generates a name for each virus based on taxonomic results and nature of sample
368 (17) Generates properly formatted fsa and tbl files in a separate directory
369 (18) Uses tbl2asn to make gbf, val, and sqn files
370
371 A 'beta version' of Cenote-Taker is available at the GitHub website:
372 (https://github.com/mtisza1/Cenote-Taker). This GitHub project is intended to demonstrate the steps
373 used to analyze the described sequences. Work is underway to make the program portable for general
374 use in other high-performance computing environments. Interested parties may contact the authors
375 about Cenote-Taker.
376
377 This work utilized the computational resources of the NIH HPC Biowulf cluster. (http://hpc.nih.gov).
378 Genome maps were drawn using MacVector 16.

379
380 Anelloviruses
381 Analysis of linear contigs in the survey found many instances of recognizable viral sequences.
382 One noteworthy example were anelloviruses, where many contigs terminated near the GC-rich stem-
383 loop structure that is thought to serve as the origin of replication. This segment of the anellovirus
384 genome is presumably incompatible with the short read deep sequencing technologies used in this
385 study. Nearly complete anellovirus genomes, defined as having a complete ORF1 gene and at least 10-
386 fold coverage, were deposited in GenBank (Table S1).
387
388 Expressing Potential Viral Structural Proteins in human 293TT cells
389 Codon-modified expression constructs encoding the three predicted STNV-like capsid proteins from the
390 nematode virus (CRESS virus isolate ctdh33) were designed according to a previously reported "as
391 different as possible" algorithm
392 https://ccrod.cancer.gov/confluence/display/LCOTF/Pseudovirus+Production. 293TT cells were
393 transfected with combinations of STNV capsid-like protein expression constructs for roughly 48 hours.
394 Cells were lysed in a small volume of PBS with 0.5% Triton X-100 or Brij-58 and Benzonase DNase/RNase
395 (Sigma). After several hours of maturation at neutral pH, the lysate was clarified at 5000 x g for 10 min.
396 The clarified lysate was loaded onto a 27-33-39% Optiprep gradient in PBS with 0.8 M NaCl. Gradient
397 fractions were collected by bottom puncture of the tube and screened by PicoGreen nucleic acid stain
398 (Invitrogen), BCA, and SDS-PAGE analysis. Electron microscopic analysis was then performed.
399
400 Expressing Potential Viral Structural Proteins in *E. coli*
401 Several genes that were identified by iVireons as being potential viral structural proteins were cloned
402 into plasmids with a T7 polymerase-responsive promoter. Plasmids were transfected into T7 Express
403 lysY/I^q *E. coli*, which express T7 polymerase under the induction of IPTG. Bacteria were grown at 37ºC in
404 LB broth until OD600 = 0.5. Flasks were cooled to room temperature, IPTG was added to 1 mM, and
405 cultures were shaken at room temperature for approximately 16 hours. Cells were then pelleted for
406 immediate processing.
407 Total protein was extracted with a BPER (Pierce) and nuclease solution. Then, virion-sized particles were
408 enriched from the clarified lysate using size exclusion chromatography with 2% agarose beads
409 https://ccrod.cancer.gov/confluence/display/LCOTF/GelFiltration. Fractions were analyzed using
410 Coomassie-stained SDS-PAGE gels for presence of a unique band corresponding to the expressed
411 protein. Fractions of interest were analyzed using negative stain electron microscopy.
412
413 Electron Microscopy
414 Five µl samples were adsorbed onto a carbon-deposited copper grid for one minute. Sample was then
415 washed 5 times on water droplets then stained with 0.5% uranyl acetate for 1 second. The negatively
416 stained samples were examined on a FEI Tecnai T12 transmission electron microscope.
417
418 Sequencing
419 Illumina sequencing was conducted at the CCR Genomics Core at the National Cancer Institute,
420 NIH, Bethesda, MD 20892.
421
422 **Acknowledgements**
423

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

References

1 Lloyd-Price, J. *et al.* Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature* **550**, 61-66, doi:10.1038/nature23889 (2017).

2 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).

3 Oh, J. *et al.* Biogeography and individuality shape function in the human skin metagenome. *Nature* **514**, 59-64, doi:10.1038/nature13786 (2014).

4 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).

5 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).

6 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).

7 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).

8 Shi, M. *et al.* Redefining the invertebrate RNA virosphere. *Nature*, doi:10.1038/nature20167 (2016).

9 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).

10 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).

11 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).

12 Bedell, M. A. *et al.* Amplification of human papillomavirus genomes in vitro is dependent on epithelial differentiation. *J Virol* **65**, 2254-2260 (1991).

13 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).

14 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).

15 Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* **19**, 455-477, doi:10.1089/cmb.2012.0021 (2012).

16 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).

17 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).

18 Marchler-Bauer, A. *et al.* CDD: NCBI's conserved domain database. *Nucleic Acids Res* **43**, D222-226, doi:10.1093/nar/gku1221 (2015).

19 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).

20 Itzhaki, R. F. *et al.* Microbes and Alzheimer's Disease. *J Alzheimers Dis* **51**, 979-984, doi:10.3233/JAD-160152 (2016).
21 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).
22 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).
23 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).
24 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).
25 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).
26 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).
27 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).
28 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).
29 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).
30 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).
31 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).
32 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).
33 Buck, C. B., Pastrana, D. V., Lowy, D. R. & Schiller, J. T. Efficient intracellular assembly of papillomaviral vectors. *J Virol* **78**, 751-757 (2004).
34 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).
35 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).
36 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).
37 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).
38 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).

525 39 Zimmermann, L. *et al.* A Completely Reimplemented MPI Bioinformatics Toolkit with a
526 New HHpred Server at its Core. *J Mol Biol* **430**, 2237-2243,
527 doi:10.1016/j.jmb.2017.12.007 (2018).
528 40 Buck, C. B. *et al.* The Ancient Evolutionary History of Polyomaviruses. *PLoS Pathog* **12**,
529 e1005574, doi:10.1371/journal.ppat.1005574 (2016).
530 41 Seguritan, V. *et al.* Artificial neural networks trained to detect viral and phage structural
531 proteins. *PLoS Comput Biol* **8**, e1002657, doi:10.1371/journal.pcbi.1002657 (2012).
532 42 King, A. M. Q. *et al.* Changes to taxonomy and the International Code of Virus
533 Classification and Nomenclature ratified by the International Committee on Taxonomy of
534 Viruses (2018). *Arch Virol*, doi:10.1007/s00705-018-3847-1 (2018).
535 43 Waldor, M. K. & Mekalanos, J. J. Lysogenic conversion by a filamentous phage
536 encoding cholera toxin. *Science* **272**, 1910-1914 (1996).
537 44 Hodyra-Stefaniak, K. *et al.* Mammalian Host-Versus-Phage immune response
538 determines phage fate in vivo. *Sci Rep* **5**, 14802, doi:10.1038/srep14802 (2015).
539 45 Wick, R. R., Schultz, M. B., Zobel, J. & Holt, K. E. Bandage: interactive visualization of
540 de novo genome assemblies. *Bioinformatics* **31**, 3350-3352,
541 doi:10.1093/bioinformatics/btv383 (2015).
542 46 Chalkias, S. *et al.* ViroFind: A novel target-enrichment deep-sequencing platform reveals
543 a complex JC virus population in the brain of PML patients. *PLoS One* **13**, e0186945,
544 doi:10.1371/journal.pone.0186945 (2018).
545 47 Geoghegan, E. M. *et al.* Infectious Entry and Neutralization of Pathogenic JC
546 Polyomaviruses. *Cell Rep* **21**, 1169-1179, doi:10.1016/j.celrep.2017.10.027 (2017).
547 48 Hunt, M. *et al.* Circlator: automated circularization of genome assemblies using long
548 sequencing reads. *Genome Biol* **16**, 294, doi:10.1186/s13059-015-0849-0 (2015).
549

Figure 1: 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. Note that genomes in this study were assigned taxonomy based on at least one region with a BLASTX hit with an E value $<1\times10^{-5}$, suggesting commonality with a known viral family. Some may ultimately be characterized as representatives of novel families

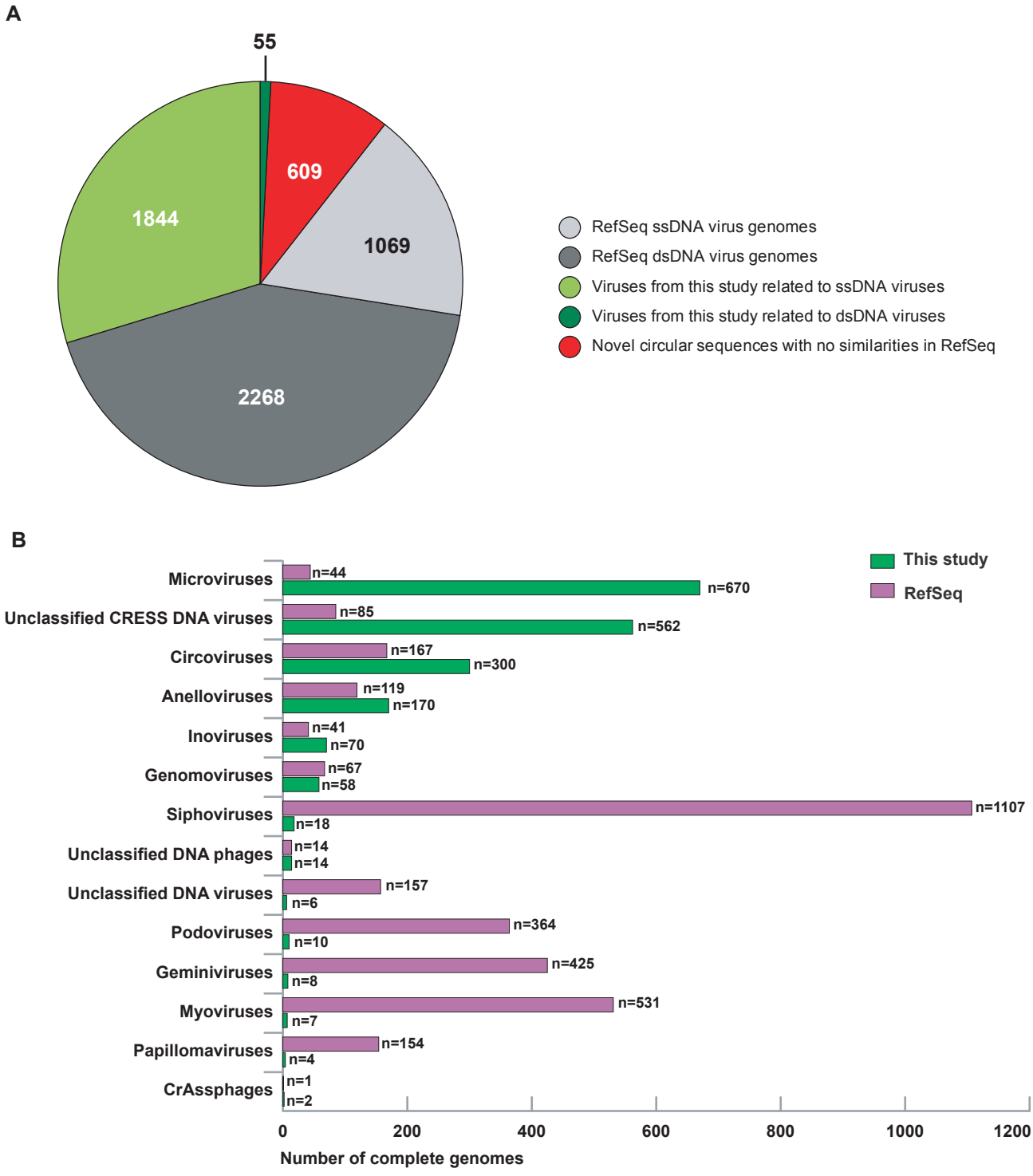
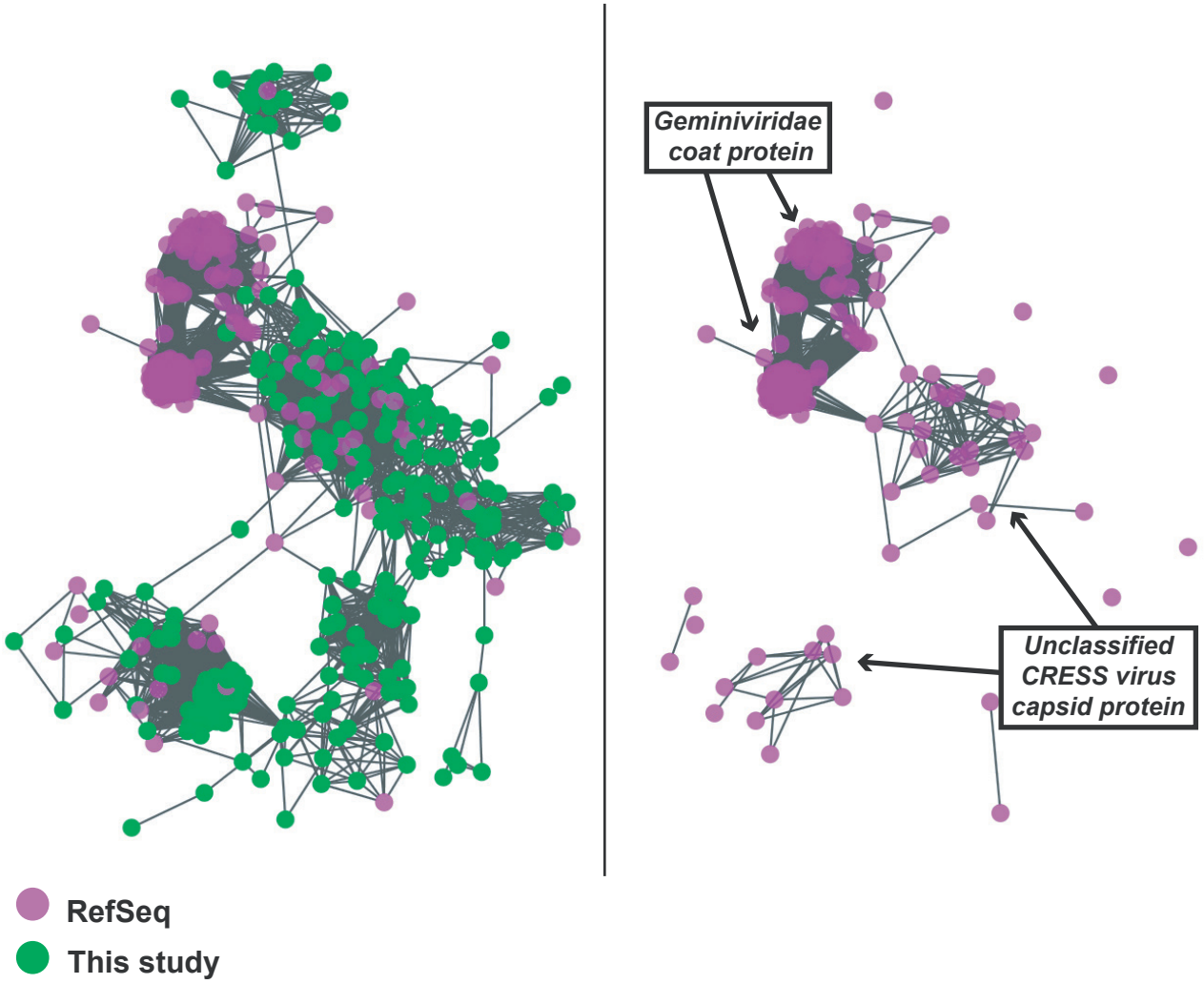


Figure 2: Network analysis of a putative category of CRESS capsid proteins

EFI-EST softwar[22] was used to perform "all against all" alignments of all RefSeq ssDNA virus proteins and all predicted proteins identified in this study. The figure shows one example of a cluster that encompasses a class of known CRESS virus capsid proteins. Additional clusters encompassing other known viral proteins are shown in Fig. S3.



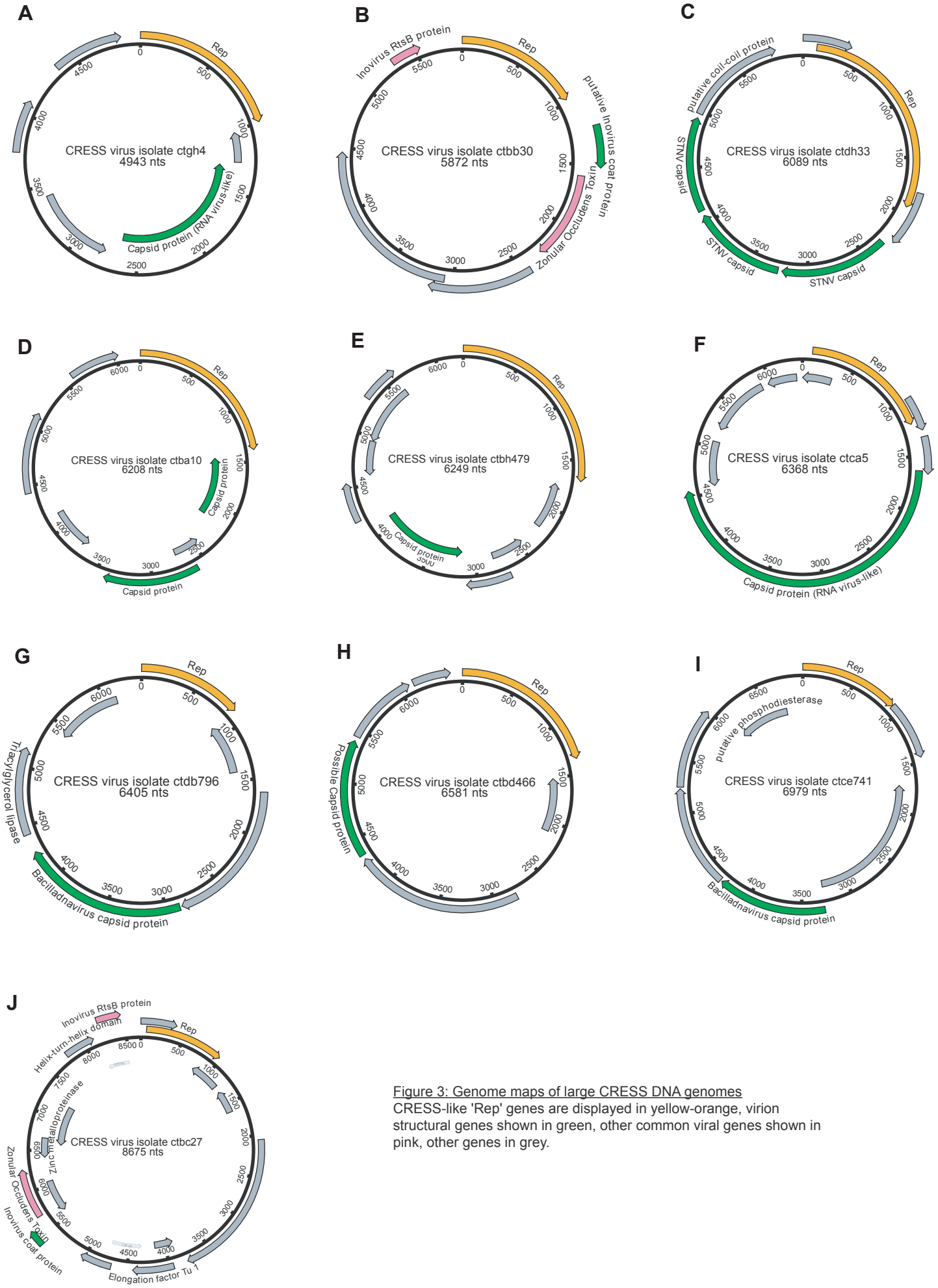


Figure 3: Genome maps of large CRESS DNA genomes
CRESS-like 'Rep' genes are displayed in yellow-orange, virion structural genes shown in green, other common viral genes shown in pink, other genes in grey.

Figure 4: 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 typified by Inoviridae. *DNA-Binding* = DNA-binding domain. *PLA2* = phospholipase A2. *FtsL* = FtsL 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 Fig. 5). (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 Fig. S8.

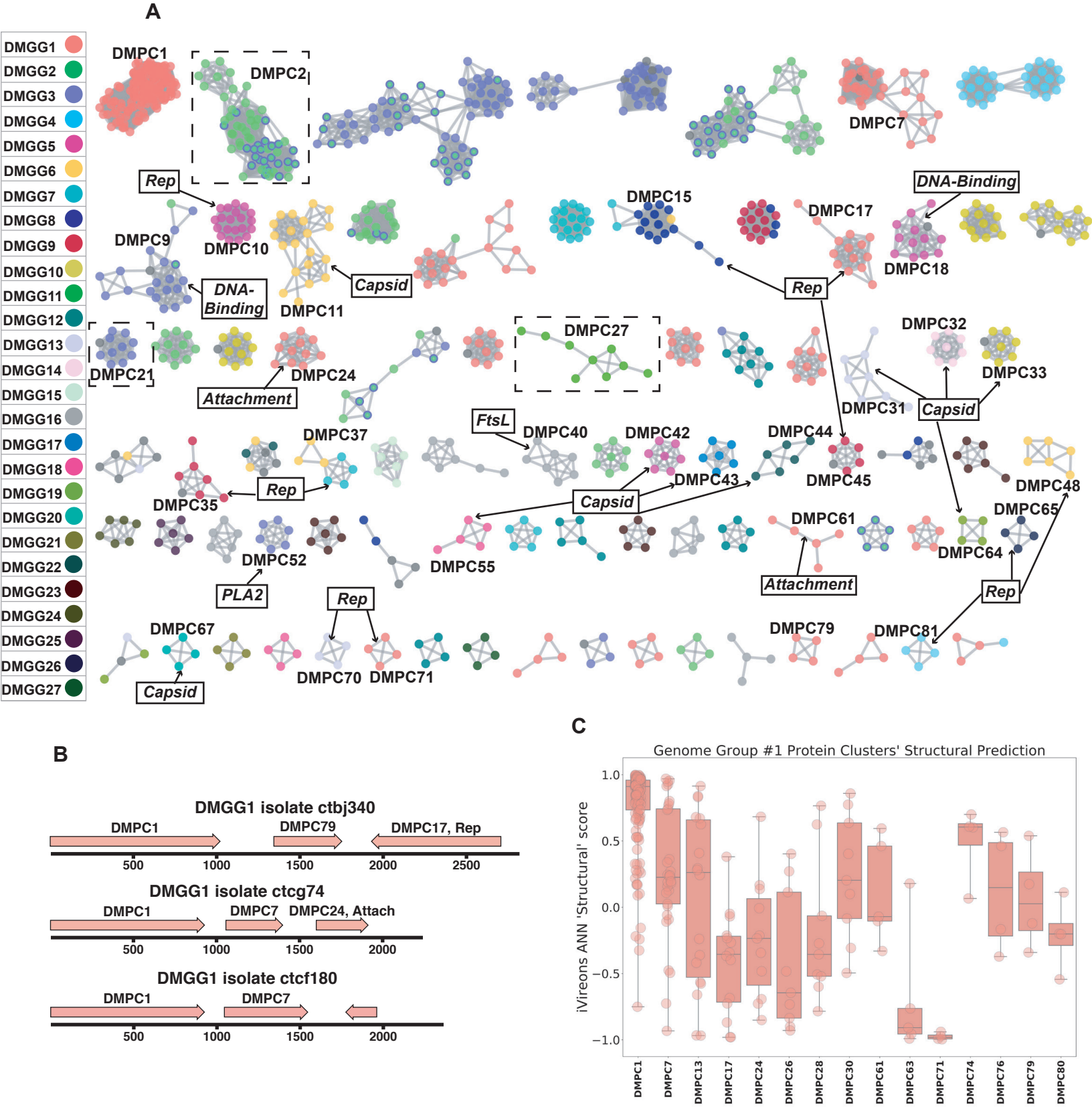
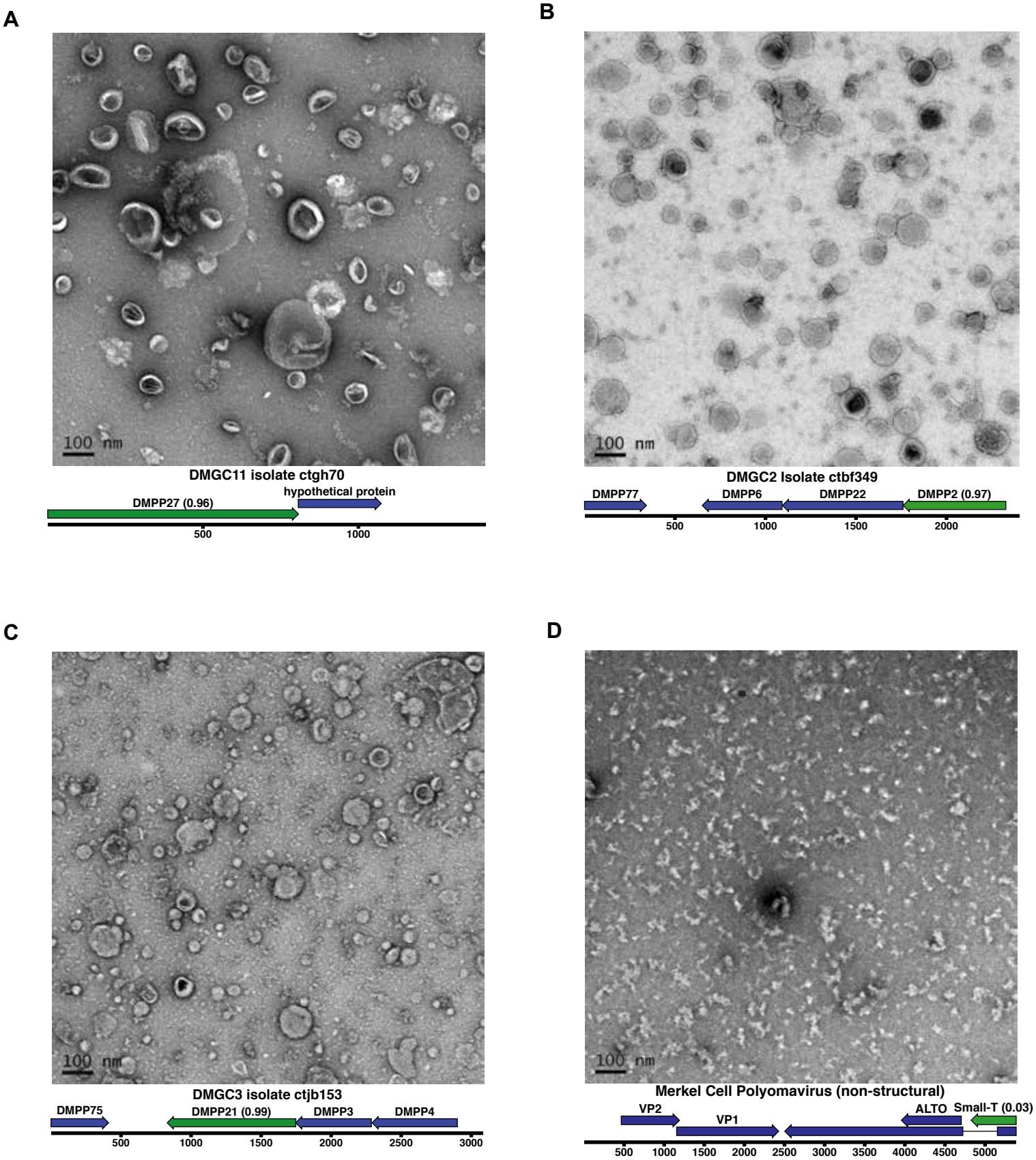


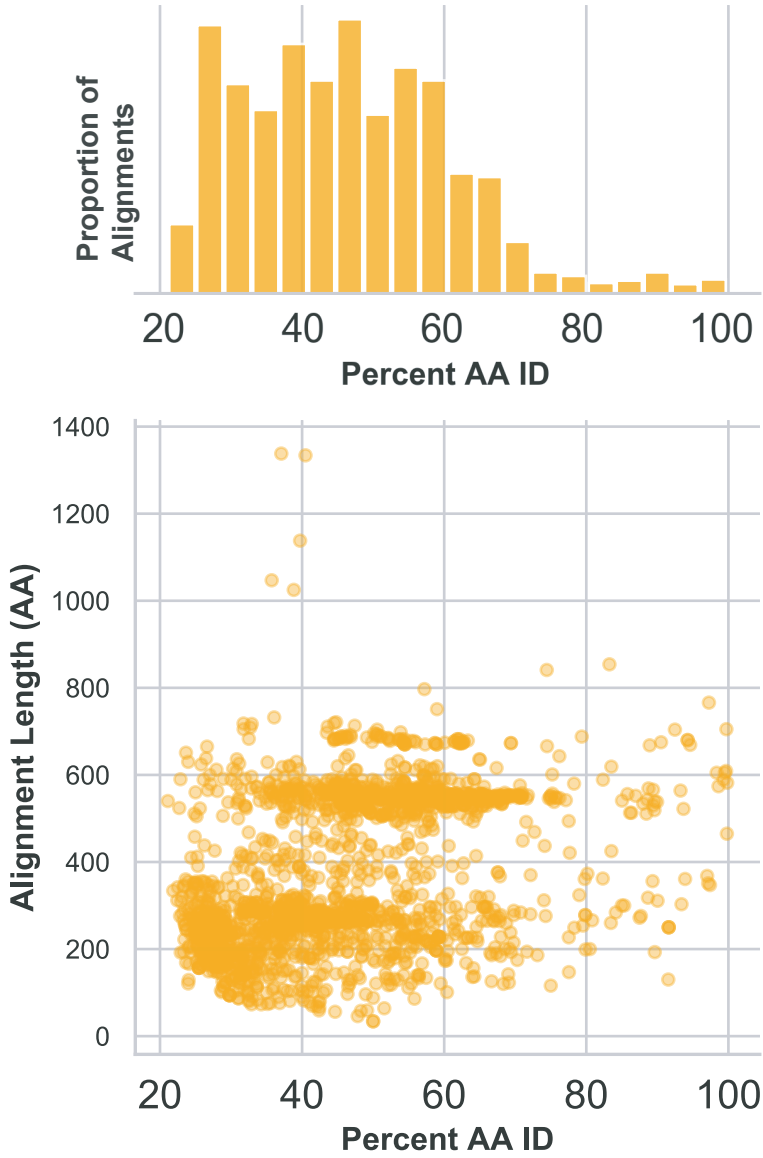
Figure 5: Expression of putative capsid proteins

Images taken by negative stain electron microscopy. Genome maps are linearized for display purposes. Expressed genes are colored green. Other genes are colored blue. 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 (a viral non-structural protein) is shown as a negative control.



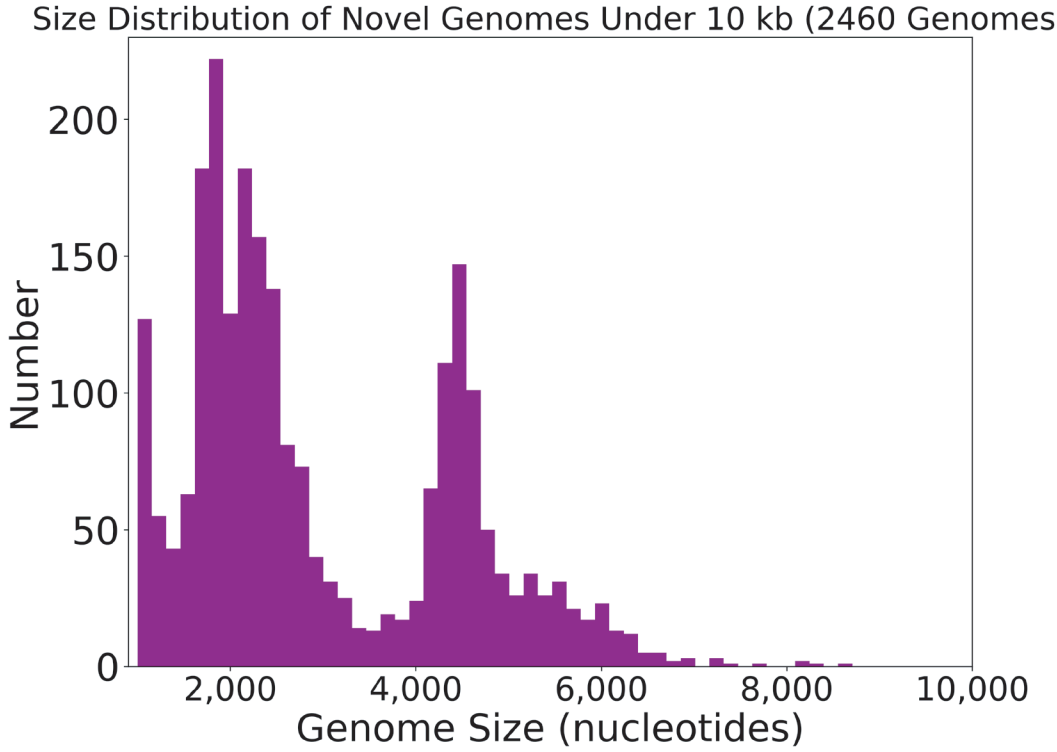
Supplemental Figure 1: Divergence of proteins encoded by circular contigs

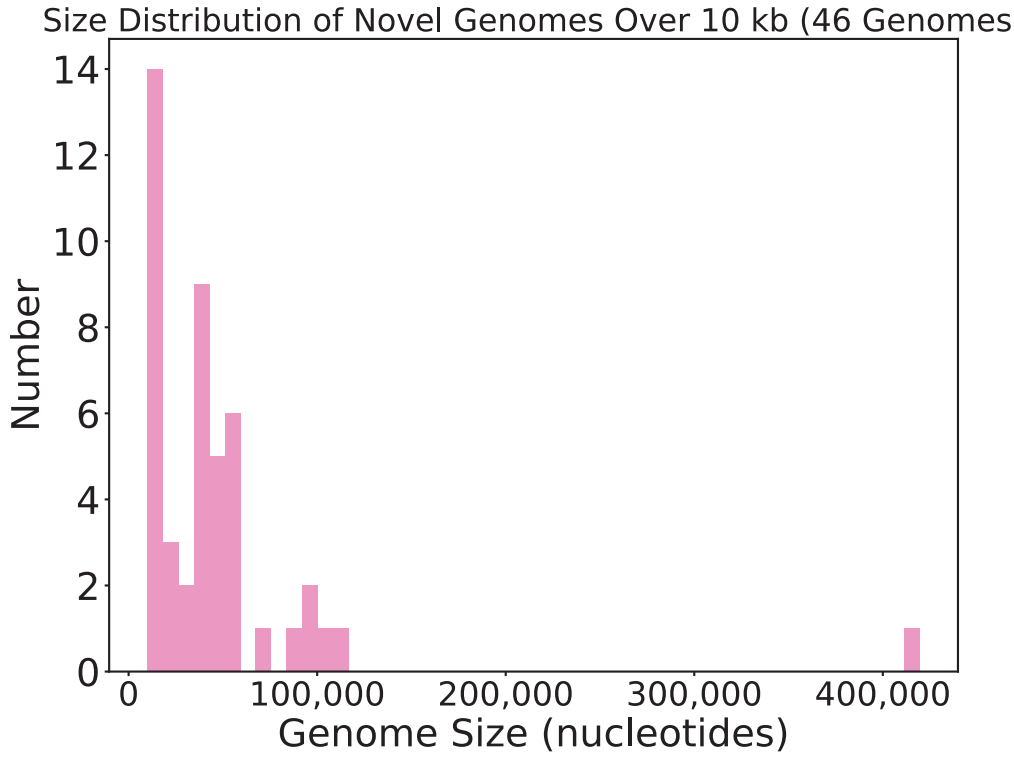
BLASTX summary of each circular DNA molecule recovered from virus enriched samples. Sequences were queried against a database of viral and plasmid sequences. Only hits with E values < 1×10^{-5} were plotted. Here, BLASTX only reports the most significant stretch of amino acid sequence from each circular contig, and, therefore, other regions of each contig can be assumed to be equally or less conserved.



Supplemental Figure 2: Size distribution of circular DNA sequences from this study

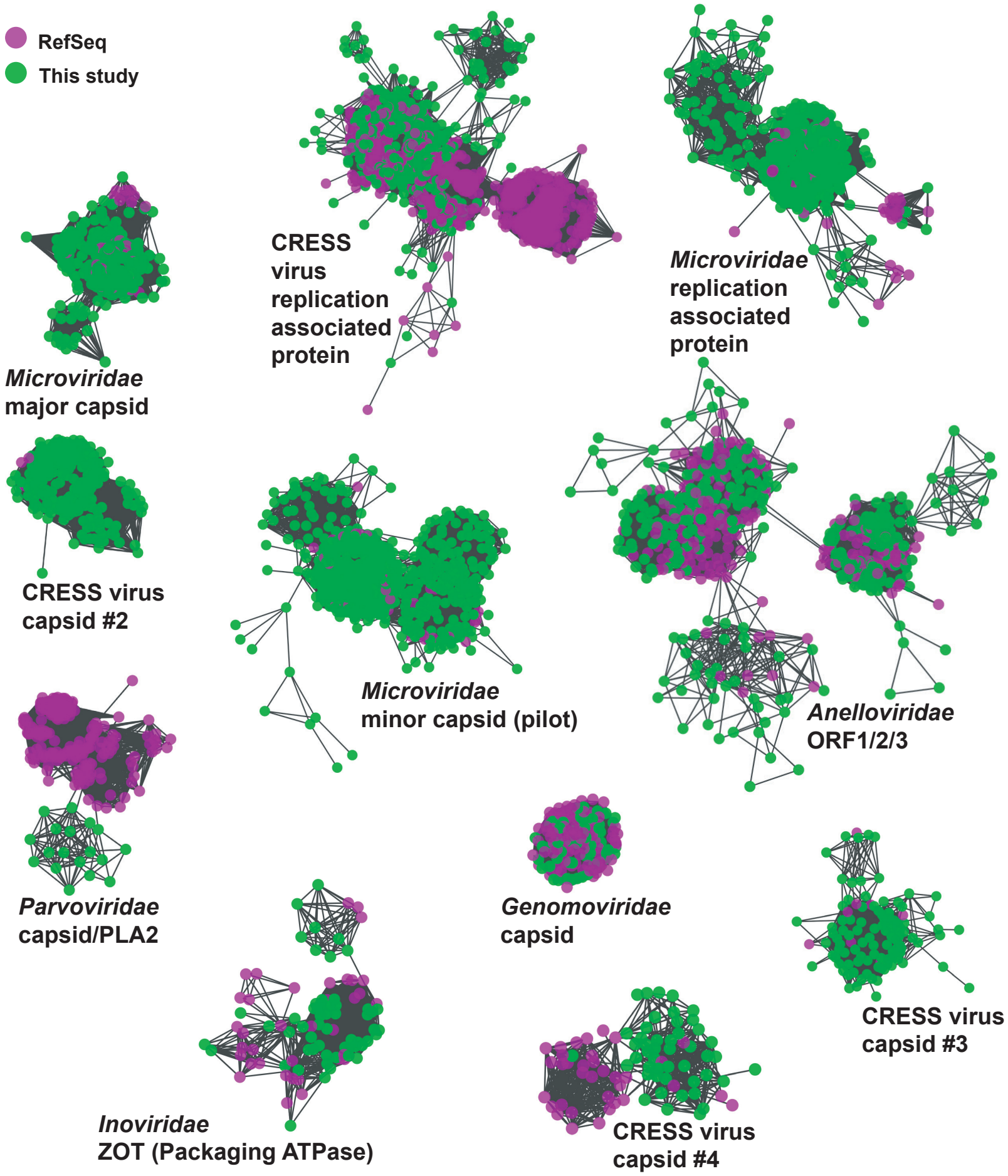
Length, in nucleotides, of circular DNA sequences representing putative viral genomes from this study.





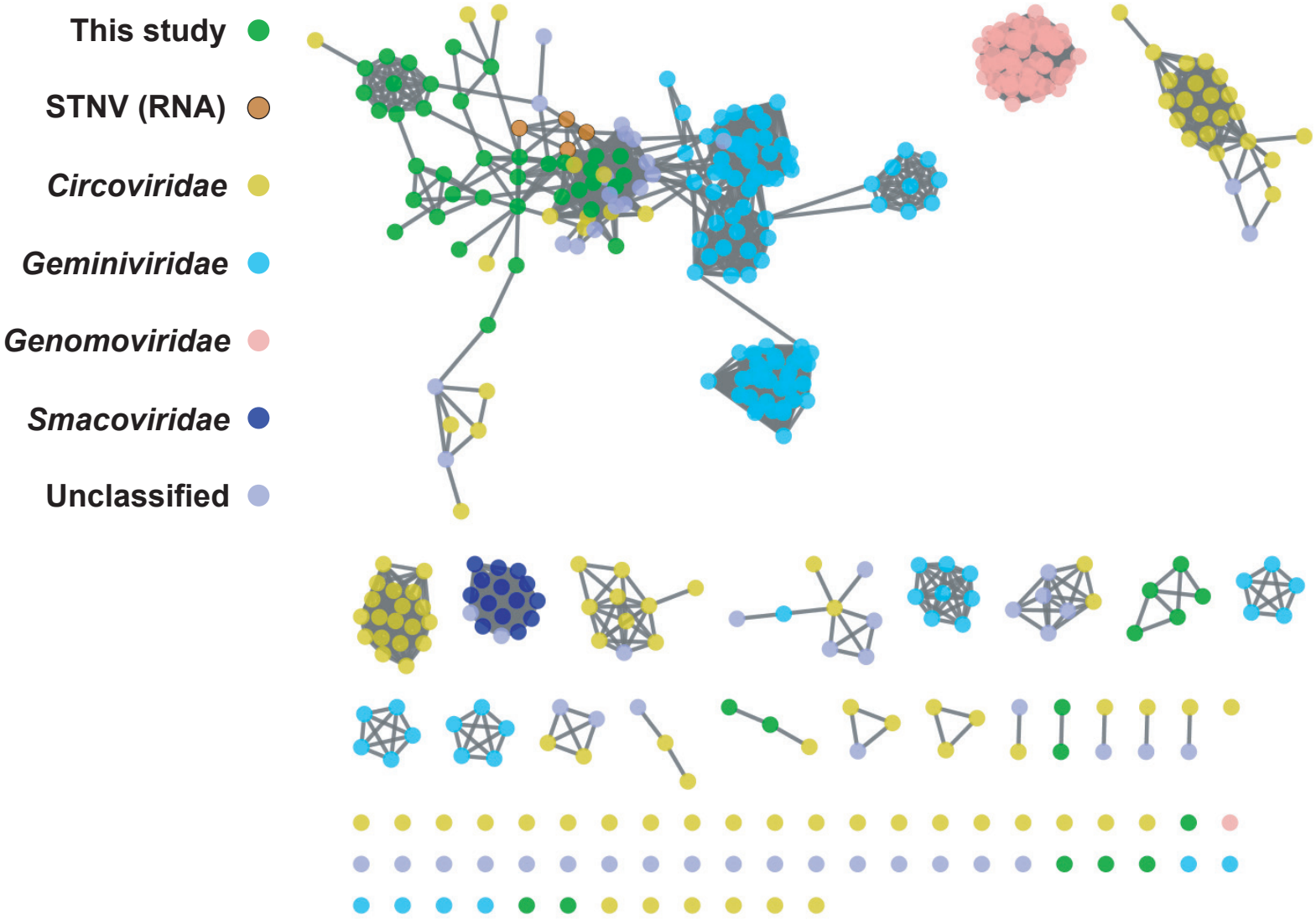
Supplementary Figure 3: Network analysis of other categories of viral proteins

The figure depicts the degree to which selected groups of novel proteins reported in the current study are related to known viral protein categories in the RefSeq database. See Figure 2 for additional information.



Supplemental Figure 4: STNV capsid sequence is within the CRESS DNA capsid "realm"

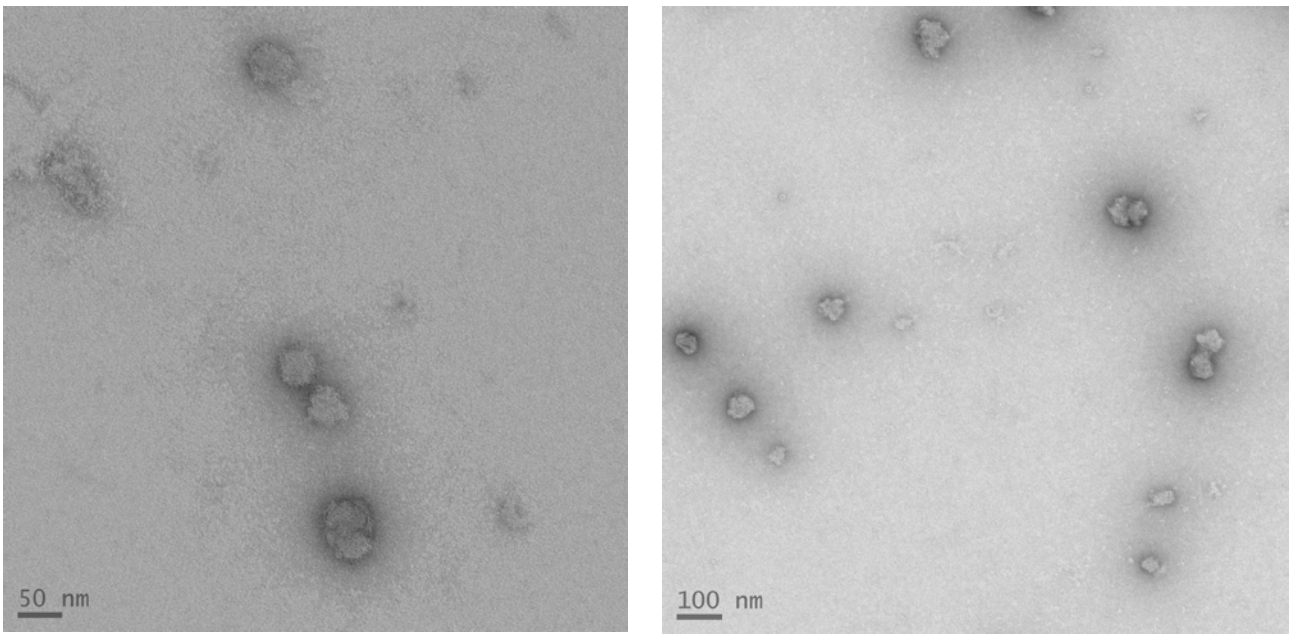
A sequence similarity network made using EFI-EST (E value<1×10^{-5}) shows capsid protein sequences of related CRESS DNA viruses in RefSeq, STNV and homologous RNA satellites from GenBank, as well as protein sequences from viral genomes discovered in the present study thought to contain the STNV fold (RPS-BLAST top hit = CDD Accession:cl27634; E value<1×10^{-5}).



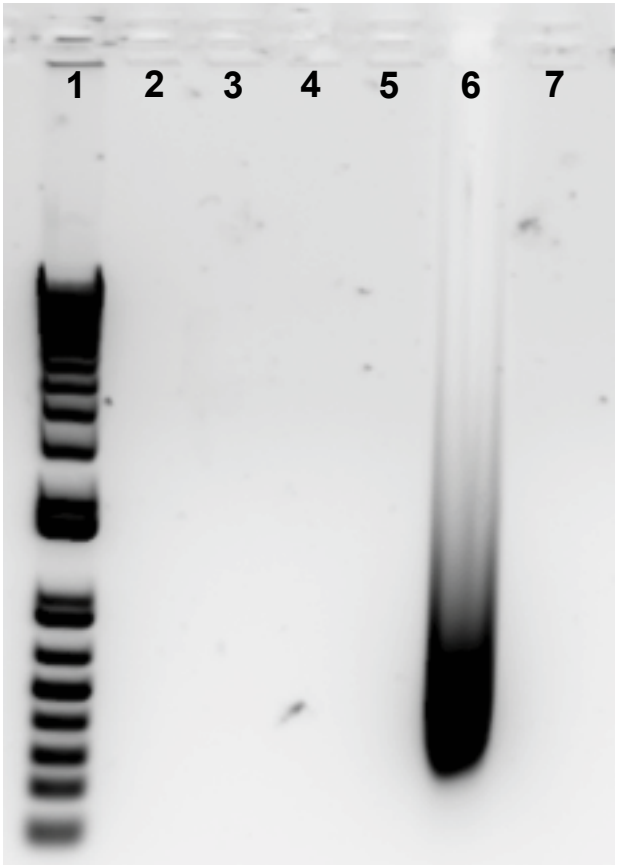
Supplemental Figure 5: Nematode-associated virus capsid expression

Putative capsid proteins from CRESS virus isolate ctdh33 associated with rhabditid nematodes (Fig. 3) were expressed in a human cell line. (A) Negative Stain EM images of virus-like particles resulting from the expression of all three genes predicted to encode for STNV-like capsid proteins. (B) Agarose gel electrophoretic analysis of nucleic acids extracted from ctdh33 VLPs shows that purified VLPs from fraction 4 of the Optiprep gradient contain RNA. Note that Benzonase digests both DNA and RNA, DnaseI only digests DNA and Rnase only digests RNA.

A



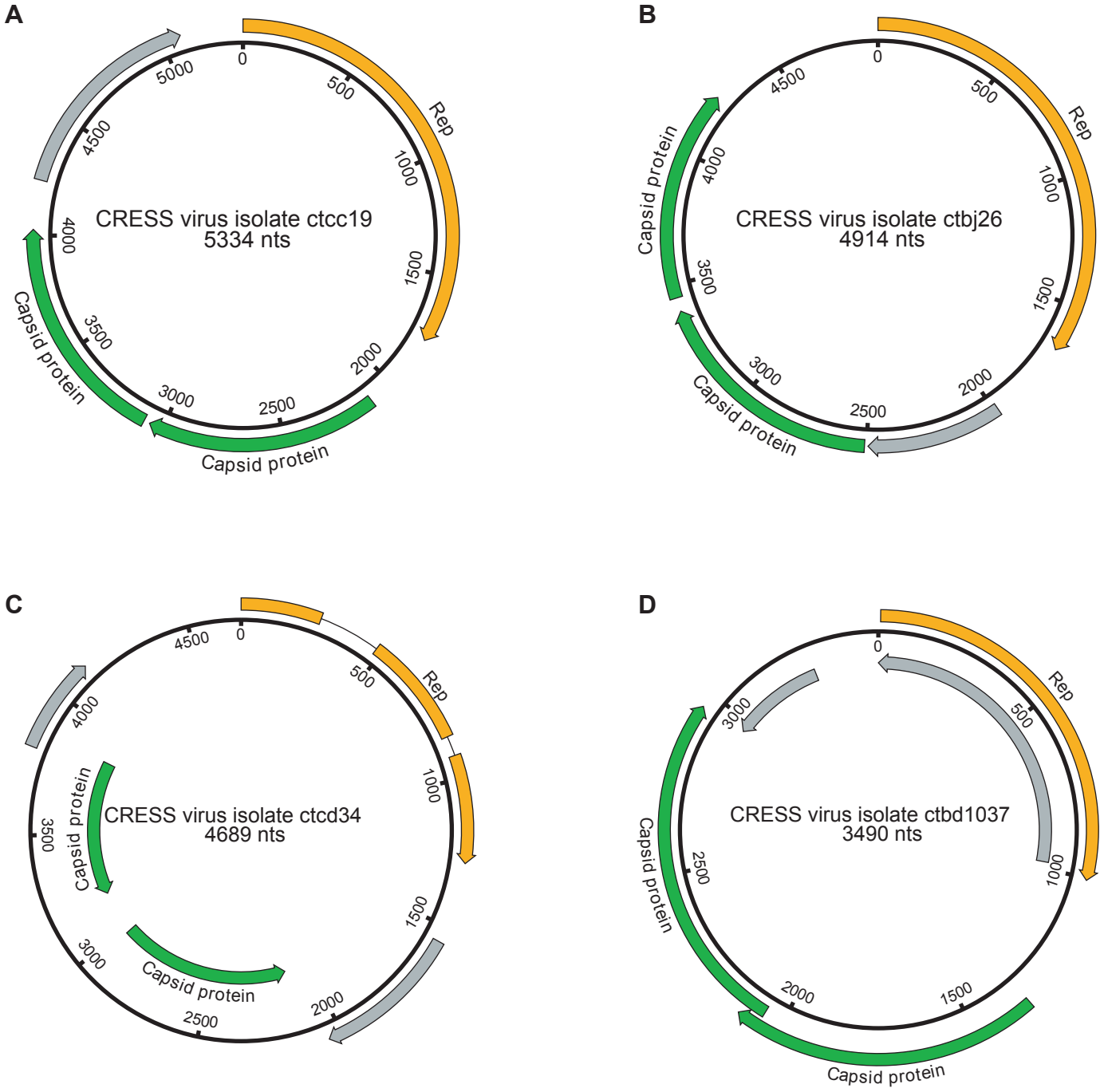
B



1 = DNA ladder
2 = Fraction 3 + Benzonase
3 = Fraction 3 + DNaseI
4 = Fraction 3 + RNase
5 = Fraction 4 + Benzonase
6 = Fraction 4 + DNaseI
7 = Fraction 4 + RNase

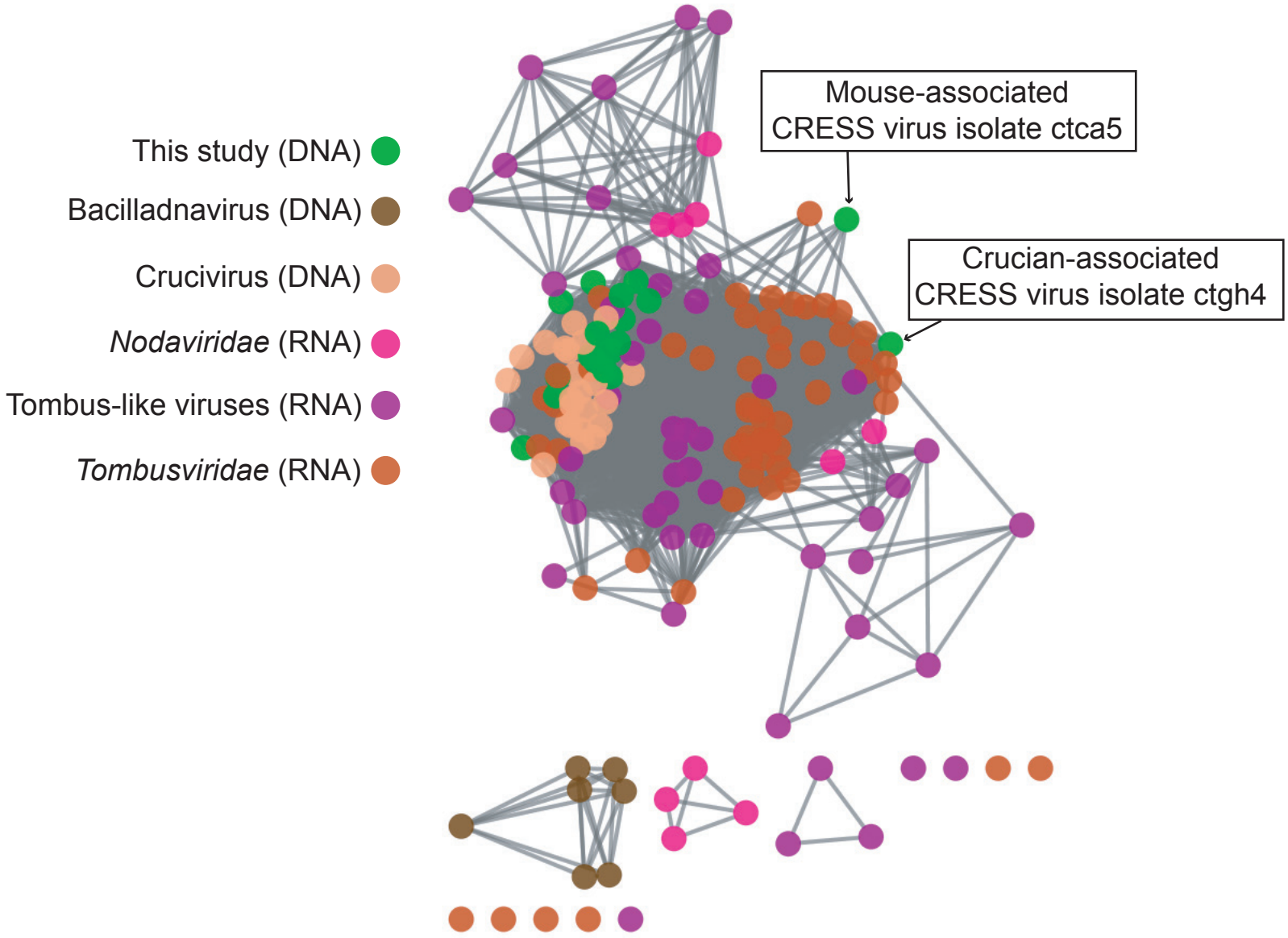
Supplemental Figure 6: Genome maps of additional CRESS genomes encoding multiple capsid genes

CRESS 'Rep' genes are shown in yellow-orange, predicted virion structural genes in green, other candidate genes in grey.



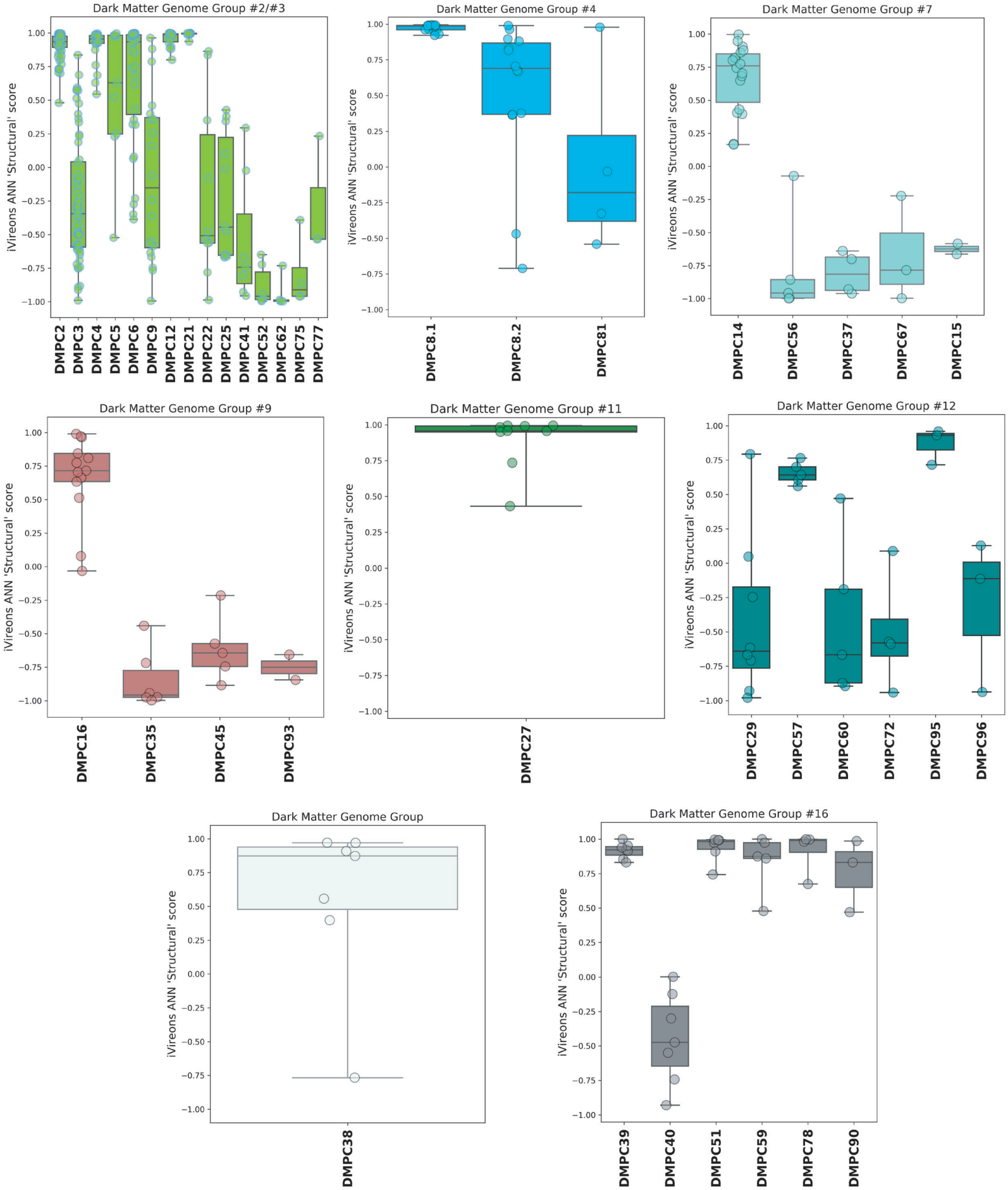
Supplemental Figure 7: Novel acquisitions of RNA virus capsids

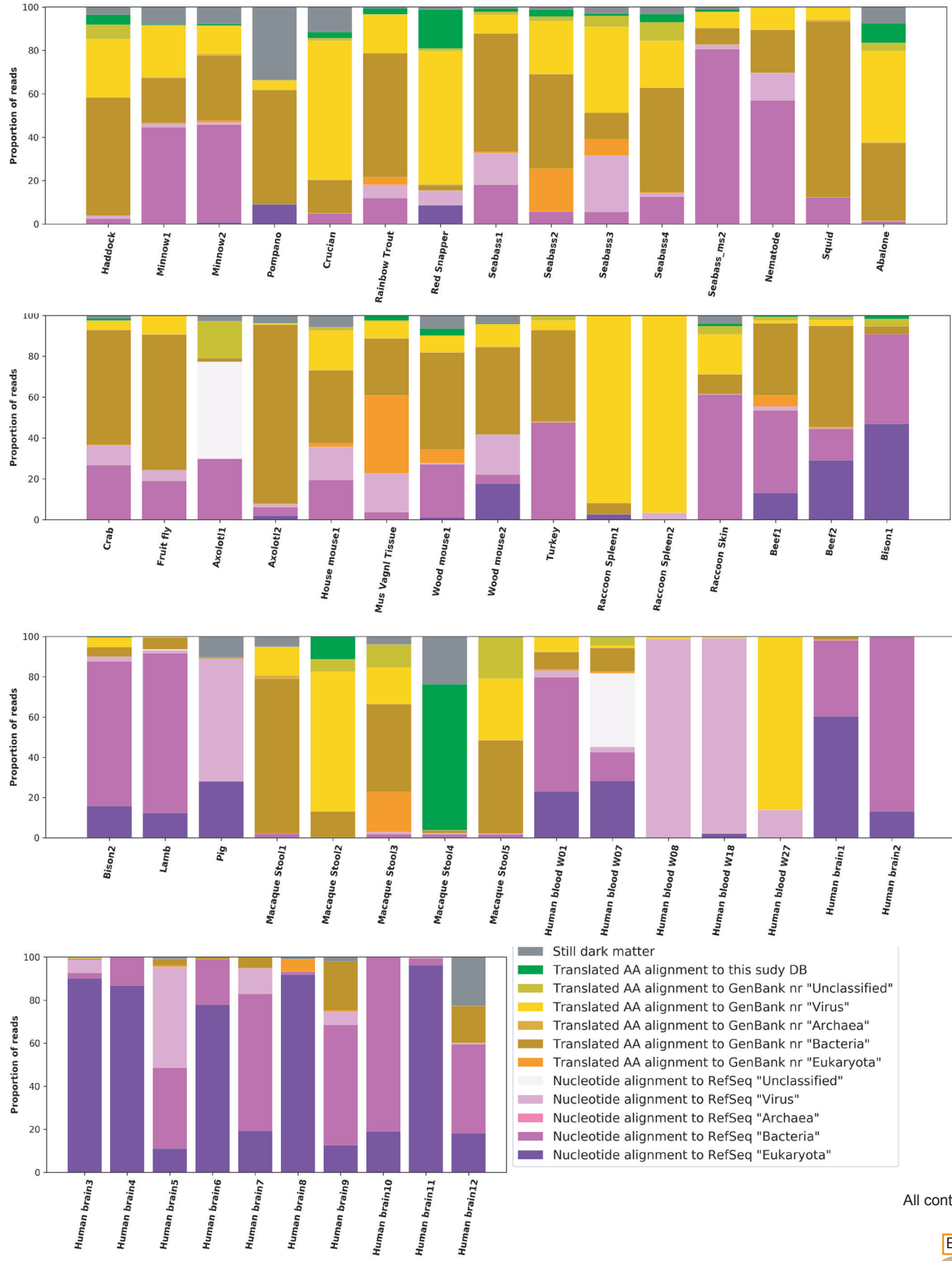
Sequence similarity network made in EFI-EST (E value <1x10^{-5}) showing capsid protein sequences of selected ssRNA viruses (Nodaviridae, Tombusviridae, tombus-like viruses) and ssDNA viruses (bacilladnavirus and crucivirus) together with protein sequences from viral DNA genomes observed in the present study with similarity to RNA virus coat protein (NCBI CDD: cl20209). CRESS virus isolate ctca5 capsid gene and CRESS virus isolate ctgh4 capsid gene had no detectible similarity to any known DNA virus sequences.

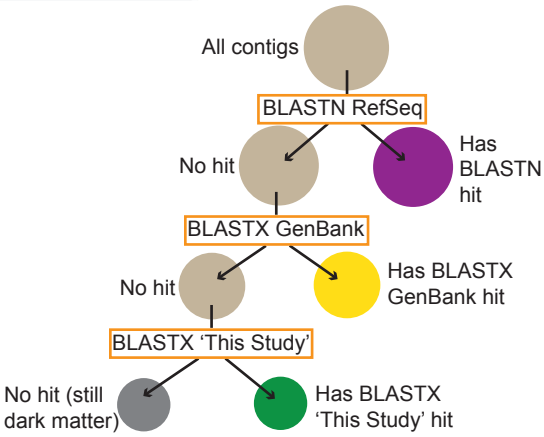


Supplementary Figure 8 (related to Figure 4C): iVireons scores of DMGGs with candidate viral structural gene(s)

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.







Supplemental Figure 9: Sample Characterizaition by iterative BLAST Searches

Contigs of over 1000 nts from each sample were subject to an iterative BLAST search chararcterization. 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.