De novo virus inference and host prediction from metagenome using CRISPR spacers

Ryota Sugimoto 1), Luca Nishimura 1,2), Phuong Nguyen Thanh 1,2), Jumpei Ito 3), Nicholas F. Parrish 4), Hiroshi Mori 5), Ken Kurokawa 5), Hirofumi Nakaoka 6) and Ituro Inoue* 1)

1) Human Genetics Laboratory, National Institute of Genetics, Research Organization of Information and Systems, 1111 Yata, Mishima, Shizuoka, Japan

2) The Graduate University for Advanced Studies, SOKENDAI, 1111 Yata, Mishima, Shizuoka, Japan

3) Division of Systems Virology, Department of Infectious Disease Control, International Research Center for Infectious Diseases, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo, Japan

4) Genome Immunobiology RIKEN Hakubi Research Team, Center for Integrative Medical Sciences, RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, Japan

5) Genome Evolution Laboratory, National Institute of Genetics, Research Organization of Information and Systems, 1111 Yata, Mishima, Shizuoka, Japan

6) Department of Cancer Genome Research, Sasaki Institute, 2-2 Kandasurugadai, Chiyoda-ku, Tokyo, Japan

Abstract

Viruses are the most numerous biological entity, existing in all environments and infecting all cellular organisms. Compared with cellular life, the evolution and origin of viruses are poorly understood; viruses are enormously diverse and most lack sequence similarity to cellular genes. To uncover viral sequences without relying on either reference viral sequences from databases or marker genes known to characterize specific viral taxa, we developed an analysis pipeline for virus inference based on clustered regularly interspaced short palindromic repeats (CRISPR). CRISPR is a prokaryotic nucleic acid restriction system that stores memory of previous exposure. Our protocol can infer viral sequences targeted by CRISPR and predict their hosts using unassembled short-read metagenomic sequencing data. Analysing human gut metagenomic data, we extracted 11,391 terminally redundant CRISPR-targeted sequences which are likely complete circular genomes of viruses or plasmids. The sequences include 257 complete crAssphage family genomes, 11 genomes larger than 200 kilobases, 766 genomes of *Microviridae* species, 114 genomes of *Inoviridae* species and many entirely novel genomes of unknown taxa. We predicted the host(s) of approximately 70% of discovered genomes by linking protospacers to taxonomically assigned CRISPR direct repeats. These results support that our protocol is efficient for de novo inference of viral genomes and host prediction. In addition, we investigated the origin of the diversity-generating retroelement (DGR) locus of the crAssphage family. Phylogenetic analysis and gene locus comparisons indicate that DGR is orthologous in human gut crAssphages and shares a common ancestor with baboon-derived crAssphage; however, the locus has likely been lost in multiple lineages recently.

41 **Introduction**

42 Viruses are not included among the domains of life, but they are the most numerous biological
43 entity, presumably containing the most genetic material on Earth. The number of viruses infecting
44 microbial populations is estimated at 10^31[1]. Viruses exist in all environments and infect all
45 cellular organisms. In the historical experiments of Hershey and Chase[2], a virus was used to first
46 show that DNA is the genetic material, and the first sequenced complete genome was viral[3,4];
47 despite those historical efforts to obtain genetic information, the evolution and origins of viruses
48 are poorly understood, reasons for which are related to the nature of viruses. Firstly, viral genomes
49 are extremely diverse in size, ranging from several thousand to a few million bases, and in genomic
50 content, encode various non-overlapping sets of genes. In fact, there are no genes universally
51 present among all viruses, indicating that a monophyletic explanation of viral origin is
52 implausible[5,6]. Secondly, most viral genes show little to no homology to cellular genes. Despite
53 their parasitic nature, viral genomes contain sparse sequences from which to infer shared lineage
54 or evolutional history with cellular life. Considering these confounding biological features, we
55 were motivated to uncover viral genomic sequences without relying on either marker genes or
56 known sequences.

57 Metagenomics[7] became a popular method to sequence viral genomes. Metagenomics
58 sequences all genetic materials extracted from a given organism or tissue. The advantage of a
59 metagenomic approach is, with sufficient depth of sequencing, presumably all genetic information
60 of viruses in the sample can be obtained. However, because metagenomic approaches obtain a
61 mixture of cellular and viral genomic information, it is necessary to subsequently distinguish viral
62 from cellular sequences.

63 Several methods have been developed to detect viral sequences from metagenomic data,
64 after assembling sequences into those derived from presumably contiguous DNA molecules called
65 contigs. One virus detection software, VirSorter[8] detects regions within contigs enriched with viral
66 'hallmark genes', viral-like genes, uncharacterized genes, and shortened genes. To detect viral
67 proteins, this program uses a custom-made protein database including both reference viral proteins
68 and predicted proteins from virome samples. Another virus detection method, VirFinder[9] is a k-
69 mer based machine learning program using a logistic regression model trained by RefSeq virus
70 and prokaryote genomes. It takes k-mer features of a given contig as input and outputs the
71 probability that the contig is a virus. Although these methods have been successfully used to
72 uncover viral diversity from metagenomics[10], both rely on similarity to known viral genomes.
73 Therefore, as reference-based approaches, they cannot uncover viral lineages totally different from
74 known viral genomes, if such lineages exist. Thus, we sought to develop a non-reference-based
75 analysis pipeline to detect viral sequences from metagenomes; however, we were tasked with the
76 question 'What information could be used for this purpose'? The underlying biology of the
77 clustered regularly interspaced short palindromic repeats (CRISPR) system, a prokaryotic form of
78 adaptive immunological memory[11], provides a potential source. After infection by viruses or
79 horizontal transfer of plasmids, some archaeal and bacterial cells incorporate fragments of 'non-
80 self' genetic materials in specialized genomic loci between CRISPR direct repeats (DRs). The
81 incorporated sequences, called 'spacers', are identical in sequence to part of the previously
82 infecting mobile genetic element. Thus, the genetic information encoded in CRISPR spacers can
83 be inferred as likely viral, and distinguishable to the genetic material of the organism encoding
84 CRISPR which is most often cellular, but potentially also viral[12,13].

85 CRISPR spacers have been used to detect viral genomes[14–18] and predict viral hosts[19,20].
86 They have previously been extracted from assembled bacterial genomes to assess CRISPR 'dark
87 matter', finding that 80–90% of identifiable material matches known viral genomes[17,18]. In the
88 current study, we extended this conceptual approach to enormous amount of unassembled short-
89 read metagenomic data. CRISPR repeats are relatively easily identifiable, particularly compared
90 to unknown viral sequences. This trait enables us to search immense metagenomic datasets for
91 reads comprised in part as CRISPR DR sequences, and in other parts, unknown sequences inferred
92 as CRISPR spacers. By analysing these reads and the contigs assembled from them, we
93 successfully extracted 11,391 terminally redundant (TR) CRISPR-targeted sequences ranging
94 from 894 to 292,414 bases. These sequences are expected to be complete or near-complete circular
95 genomes which can be linked to their predicted hosts. The sequences we inferred de novo in this
96 manner included 257 complete genomes of the crAssphage family[21,22], 11 genomes larger than
97 200 kilobases (kb), 766 genomes of *Microviridae* species, 114 genomes of *Inoviridae* species and
98 many novel, taxonomically unplaced genomes. Furthermore, by comparing the spacer-associated
99 metagenomic CRISPR DR sequences to CRISPRs from reference bacterial and archaeal genomes,
100 we predicted CRISPR targeting hosts for 69.7% of the discovered genomes.
101
102 **Results and Discussion**
103 **Extraction of CRISPR-targeted sequences.** For this study, we analysed human gut metagenomes
104 as they serve as an 'ecosystem' with the most abundant metagenomic data available to our
105 knowledge. We downloaded 11,817 human gut metagenome datasets equivalent to 50.7 Tb from
106 the European Nucleotide Archive FTP server. FASTQ files were pre-processed and assembled to
107 180,068,349 contigs comprising 767.7 Gb of data (Supplementary Table 1). We discovered 11,223

108 unique CRISPR DRs from the assembled contigs which were used to extract CRISPR spacers from
109 raw reads, resulting in 1,969,721 unique CRISPR spacers (Supplementary data). These spacers
110 were then used as query to reveal candidate protospacers (i.e. contigs containing the spacer
111 sequence not within a CRISPR locus). Spacers were mapped to CRISPR masked contigs with a
112 93% minimum sequence identity threshold. To increase specificity, we checked that the 5' and 3'
113 adjacent sequences of spacer-mapped positions were not similar either to each other or the spacer-
114 associated DR. A total of 164,590,387 candidate protospacer loci, attributed to 1,114,947 unique
115 spacers (56.6% of all unique spacers) were revealed (Supplementary data), showing a substantially
116 higher discovery rate compared with that of a previous study (~7%[17]) which used National Center
117 for Biotechnology Information (NCBI) nucleotide sequences for protospacer discovery. Spacers
118 were clustered based on protospacer co-occurrence and used to extract contigs targeted by more
119 than 30% of members of a spacer cluster. Finally, 764,883 gapless CRISPR-targeted sequences
120 (15.9 Gb) were extracted (Supplementary Fig. 1 and 2). Of these, 11,391 unique sequences were
121 identified as TR[23], which are expected if originating as complete or near-complete circular mobile
122 genetic elements. The size of CRISPR-targeted TR sequences ranged from 894 to 292,414 bases
123 (Supplementary table 2 and Supplementary data).

124 We then investigated protein coding genes encoded in CRISPR-targeted sequences. To
125 maximize recovery of novel genes, 3,099,178 protein coding genes were predicted from CRISPR-
126 targeted sequences, including both TR (400,566 genes) and non-TR sequences (2,698,612 genes).
127 Protein sequences were clustered based on a 30% sequence identity threshold, resulting in 83,796
128 clusters with more than two members and comprising 2,996,113 genes (96.7%). Each cluster was
129 aligned and used to build a hidden Markov model (HMM), then used as query against UniRef50
130 and RefSeq viral proteins. We found that 12,283 HMMs (14.7%), comprising 891,121 genes

131 (28.8%), were similar to 60,296 different RefSeq viral protein sequences (E-value < 10^-10). A
132 total of 6,853 TR sequences (60%) encoded at least one of these genes, suggesting that a substantial
133 portion of TR sequences are viral genomes.
134
135 **Length-based classification of CRISPR-targeted TR sequences.** Evaluation of TR sequence
136 length showed a multi-modal distribution with a distinct trough at 20 kb (Fig. 1a). For reference,
137 we refer to the 8,837 TR sequences shorter than 20 kb as 'small' and the 2,554 TR sequences
138 longer than 20 kb as 'large'. This simple classification was previously used to infer capsid
139 morphology[24,25]. Among large TR sequences, 1,509 sequences (59%) encode a set of tail, portal
140 and terminase large-subunit genes (score > 30), whereas only 50 small TR sequences (0.6%)
141 encoded this set of genes (Supplementary Fig. 3), indicating that the majority of large TR
142 sequences are tailed phages, also known as *Caudovirales*. In addition, 766 small TR sequences,
143 ranging in size from 4,574 to 12,988 bases (average 5,766.14 bases), encoded both *Microviridae*
144 major capsid and replication initiator homologues[26]. 114 small TR sequences, ranging in size from
145 4,783 to 18,988 bases (average 6,715.6 bases), encoded both *Inoviridae* major coat and replication
146 initiator homologues. We consider that this portion of small TR sequences are likely viruses with
147 a non-tailed morphology[27,28].
148
149 **Predicted CRISPR targeting hosts of TR sequences.** As our approach uses CRISPR spacers to
150 extract CRISPR-targeted sequences, we hypothesised that the relationship between virus and
151 targeted host can be resolved for the majority of TR sequences. CRISPR DR sequences were
152 searched on RefSeq genomes and taxonomically assigned. Based on counts of protospacers
153 associated with taxonomically assigned DRs, 7,937 TR sequences (69.7%) were resolved to a

154 targeting host at the phylum level (Fig. 1b and Supplementary Fig. 4), and 6,083 TR sequences
155 (53.4%) were resolved to a targeting host at the order level (Supplementary Fig. 5 and
156 Supplementary table 2). The most frequent host was *Firmicutes*, followed by *Bacteroidetes* and
157 *Actinobacteria*. Notably, these are the most common bacteria in human intestine[29]. In addition,
158 1,418 TR sequences (12.5%) were predicted to be targeted by multiple phyla. Although some of
159 these TR sequences were targeted by multiple monoderm or diderm phyla exclusively, there was
160 an exceptional host ambiguity between *Firmicutes* and *Verrucomicrobia* which crosses the
161 monoderm-diderm boundary (Fig. 1c). We did not use CRISPR DR sequences shared between
162 cross-phyla species for targeting host prediction; targeting host ambiguity is unlikely because of
163 ambiguous taxonomic assignment of DR sequences. To determine whether the targeting host
164 predicted by this method corresponds with reported permissive hosts, we calculated GC content[30]
165 of the TR sequences (Fig. 1d). Among the 2,050 high-GC (GC% > 55%) TR sequences, 1,109
166 (54%) and 249 (12.1%) were predicted to be targeted by *Actinobacteria* and *Firmicutes*
167 respectively, and the targeting host was undetermined for 222 TR sequences (10.8%). The large
168 fraction of high-GC TR sequences predicted to be targeted by *Actinobacteria* likely indicate
169 genomic adaption of parasitic genetic elements which infect and routinely become targeted by
170 CRISPR systems of high-GC gram-positive bacteria.

171 Host range within a viral lineage was further assessed by phylogenetic analysis of TR
172 sequences determined to represent *Microviridae* species. The predicted targeting hosts of putative
173 *Microviridae* species from our study were *Bacteroides*, *Firmicutes* and *Proteobacteria*, indicating
174 a wide host range of *Microviridae* species. Molecular phylogeny of the *Microviridae* major capsid
175 protein clearly separates sequences based on their targeting host (Fig. 2). Interestingly, most
176 known *E. coli*-infecting viral species (such as phiX174) and other *Proteobacteria*-infecting species

177 (such as phi MH2K[31]) used as reference in this phylogeny split to two clades, and the clade
178 containing the latter was within a clade of TR sequences targeted by *Firmicutes*. The nested
179 phylogenetic structure of capsids encoded by TR sequences predicted to represent *Microviridae*
180 species may indicate that host-switching, a critically important topic in virus evolution[32,33], has
181 occurred within the viral lineage.

182 These results suggest that approximately 70% of discovered sequences are targeted by
183 specific host phyla. 12.5% of TR sequences were targeted by multiple phyla; however, we are
184 uncertain whether these elements actually infect multiple hosts at present, recently host-switched,
185 or some genomes become CRISPR-targeted as a result of abortive infections[34]. Importantly,
186 expanding this approach to detect viral genomes from different environments may be useful to
187 infer the evolutionary history and genetic factors involved in host-switching.

188

189 **Comparison of CRISPR-targeted TR sequences to available viral and plasmid sequences.** TR
190 sequences identified in this study were compared to virus and plasmid genome databases,
191 including RefSeq[35] virus, RefSeq plasmid, IMG/VR[36,37] and GVD[10] (Fig. 3). IMG/VR is currently
192 the largest database of uncultivated viral genomes. GVD is a database of viral genomes discovered
193 from human gut metagenome datasets using VirSorter and VirFinder. Both of these databases rely
194 on protein homology, which complements to our approach, to discover the viral genomes.
195 Therefore, we consider these databases are appropriate to validate our results.

196 Among the 2,554 large TR sequences, we found that 1,726 TR sequences (67.6%) were
197 represented in RefSeq virus, IMG/VR or GVD (Fig. 3a), using a threshold of 85% sequence
198 identity with at least 75% aligned fraction of the query sequence to a unique subject sequence.
199 These sequences included 257 crAssphage genomes, ranging in size from 92,182 to 100,327 bases

200 (average 96,984.8 bases). Members of *Bacteroidales* were predicted as targeting hosts of
201 crAssphages in this study, findings which are consistent with those hosts reported to propagate
202 crAssphage in previous studies[21,22]. We also found that 154 large TR sequences (6%) were listed
203 in RefSeq plasmid, of which 7 were also listed in the IMG/VR database. Based on our investigation,
204 we determined that there is adequate classification between plasmids and viruses for large genomes.
205 Notably, we discovered 11 TR sequences larger than 200 kb; 7 of which correspond to recently
206 reported 'huge phage' genomes[13] and 4 of these very large TR sequences were novel. Finally, we
207 found that 681 large TR sequences (26.7%) were novel. In summary, we conclude that many of
208 the large TR sequences are already represented in virus or plasmid databases, with the exception
209 of those greater than 200 kb which were only recently reported.

210 In contrast to the large TR sequences, most of the small TR sequences were not represented
211 in the databases searched (Fig. 3b). Among the 8,837 small TR sequences, 491 small TR sequences
212 (5.6%) were listed in IMG/VR, whereas 2,573 small TR sequences (29.1%) were listed in GVD.
213 Only 256 small TR sequences (2.9%) were listed in both IMG/VR and GVD. Finally, 5,518 small
214 TR sequences (62.4%) were novel, indicating that the majority of small viral or plasmid sequences
215 targeted by CRISPR are still unexplored, even in the intensively studied human gut metagenome.
216 However, IMG/VR filters out genomes shorter than 5 kb to minimize the rate of false positive
217 predictions[37], which likely explains the substantially lower represented sequences in IMG/VR for
218 small TR sequences. Moreover, among the 942 RefSeq plasmid-listed sequences, 444 small TR
219 sequences were listed in RefSeq virus, IMG/VR and/or GVD, suggesting a possible
220 misclassification between virus and plasmid within these databases. We investigated database
221 representation of CRISPR-targeted TR sequences predicted to represent *Microviridae* and
222 *Inoviridae* species. Among the 766 putative *Microviridae* genomes from this study, we found that

223 639 genomes (83.4%) were represented in at least one virus database, and none were listed among
224 RefSeq plasmids. In contrast to *Microviridae*, among the 114 putative *Inoviridae* genomes from
225 this study, only 2 were listed in the virus databases and 11 were listed in the RefSeq plasmid
226 database. Next, we compared our putative *Inoviridae* genomes with recently reported *Inoviridae*
227 genomes discovered using a machine learning approach[38], and found that 43 genomes from our
228 study were highly similar to the genomes from the previous study, supporting our prediction that
229 these sequences are indeed *Inoviridae* genomes.

230 To determine whether our novel TR sequences are viruses, we searched sequences for
231 capsid genes using a high-sensitivity homology detection method. We built a capsid HMM
232 database using reference capsid proteins as baits which were then used to scan TR sequences. We
233 identified 1,582 small TR sequences (17.9%) encoded capsid genes (score > 30), suggesting that
234 at least this portion of TR sequences are viral genomes. However, capsid genes are unlikely to
235 share a single common ancestor[39], and it is unlikely that the reference sequences cover the entire
236 sequence diversity of capsids. In addition, some viruses do not encode capsids[40,41]. Therefore,
237 although this is a conservative estimate of the truly viral fraction of the taxonomically
238 unclassifiable sequences reported here, we are unable to strongly conclude whether the remaining
239 small TR sequences are viruses or not at this point. With DNA synthesis technology and predicted
240 host information from this study, one might be able to experimentally investigate these potentially
241 viral genomes.

242

243 **Gene content based hierarchical clustering of CRISPR-targeted TR sequences.** To scrutinize
244 CRISPR-targeted TR sequences in a genomic context, we hierarchically clustered TR sequences
245 based on gene content. For this analysis, we selected the top 1,000 genes recurrently observed

246 from large and small TR sequences. The clustering results for large TR sequences (Fig. 4a) showed
247 that the majority of genomes, with a variety of gene contents, have already been listed in virus
248 databases. This finding further supports that these databases contain at least representatives similar,
249 at broad taxonomic levels, to the large viruses present in human gut. Specific gene contents were
250 observed for crAssphages, which are exclusively targeted by *Bacteroidetes*. In addition, RefSeq
251 plasmid hit sequences formed an exclusive cluster with conjugation-related genes. As the
252 conjugation proteins included those involved in pili formation and intercellular DNA transfer
253 proteins, these sequences are likely F plasmids. TR sequences predicted to be targeted by
254 monoderm and diderm hosts clustered separately, indicating there is little gene flow between them.
255 Except for the likely F plasmid sequences, most of these sequences encode capsids. Combined
256 with the fact that most of the large TR sequences encoded tailed phage essential genes, this result
257 further supports the conclusion that the majority of the large TR sequences are tailed phages.

258 In contrast to the large TR sequences, we found that the small TR sequences remained
259 largely enigmatic when clustered by gene content. Few clusters were represented in IMG/VR, and
260 although GVD covers a relatively broader range, representatives are still missing or sparse for
261 some clusters. Several clusters were also listed in both plasmid and virus databases. As many
262 clusters did not encode detectable capsid genes, both clusters representing *Microviridae* evince
263 capsid genes. Another pattern shown by this analysis was that high-GC content TR sequences were
264 frequently observed with *Actinobacteria* as the predicted targeting host.

265

266 **Diversity-generating retroelements of crAssphage lineage.** Recently, one crAssphage genome
267 encoding a diversity-generating retroelement (DGR) was reported[42]. DGRs are retroelements
268 found in prokaryotes and viruses[43]. They minimally consist of an error-prone reverse transcriptase

269 and a non-coding region called a template repeat. The transcript of a template repeat is reverse
270 transcribed, and this complementary DNA (cDNA) product includes mutations from adenine to a
271 random nucleotide (A-to-N). The cDNA is incorporated to a specific position of the target gene
272 called the variable repeat; thus, this element effectively alters the coding sequence of the target
273 gene which may benefit from rapid adaptation, such as expansion to new hosts and
274 environments[44,45]. Among the 257 putative crAssphage genomes identified in our study, 98
275 genomes encoded DGRs including a reverse transcriptase gene with an associated non-coding
276 region recognizable as a putative template repeat showing A-to-N (adenine-to-any base) mutations
277 to the tail-collar coding gene. Using these abundant putative crAssphage sequences, we further
278 investigated this locus to uncover the origin of this retroelement.

279 The position of the DGR locus was consistent among DGR-encoding crAssphages
280 (Supplementary Fig. 6), suggesting that DGRs in crAssphages observed so far are orthologous and
281 likely originate from a common ancestral DGR insertion. The reverse transcriptase gene and
282 template repeat within this module showed high sequence conservation, whereas the variable
283 repeat within the adjacent gene that encodes tail-collar fibre showed high diversity, indicating that
284 this DGR was recently generating diversity (Fig. 5a). Notably, however, the tail-collar fibre of
285 putative crAssphage sequences without DGR were similarly diverse. We used the protein sequence
286 of DGR reverse transcriptase to query non-human-derived crAssphage-like genomes[46], and only
287 one crAssphage-like genome, which was derived by sequencing baboon faeces, encoded this
288 element (Fig. 5a). Pair-wise alignment of template and variable repeats within the baboon
289 crAssphage-like genome showed A-to-N substitutions that are likely facilitated by the DGR (Fig.
290 5b), which indicates the DGR in the baboon crAssphage-like genome is also functional. We then
291 constructed a Bayesian phylogeny of crAssphage genomes (Fig. 5c). We found that the distribution

292 of reverse transcriptase-encoding genomes (shown in red text in Fig. 5c) did not form a cluster, a
293 finding which is most consistent with multiple losses of this retroelement, and potentially
294 explaining the high diversity of the tail collar fibre even in assembled sequences that do not
295 currently contain a DGR. A deeper-rooted molecular phylogeny with distant crAssphages-like
296 genomes would be required to uncover more details about the origin of this retroelement.
297
298 **Remnant CRISPR spacers and overall contribution of CRISPR-targeted sequences to**
299 **identified spacers.** Viruses and other mobile genetic elements can escape from CRISPR targeting
300 by acquiring mutations to protospacer loci[47,48]. While the corresponding spacers are thus no longer
301 effective, they can remain in the host genome. To investigate these potential 'remnant' CRISPR
302 spacers, we mapped all unique CRISPR spacers to TR sequences as well as scrambled sequences
303 with various sequence identity thresholds (Fig. 6). The scrambled sequences were used to monitor
304 false positive matches arising by chance (see Methods). Based on the observation of incremental
305 false positive matches of spacers to scrambled sequences, 84% is a sequence identity threshold at
306 which some putative remnant spacers can be mapped, but very few false-positive matches. At an
307 84% sequence identity threshold, 269,808 spacers and 126,616 spacers were mapped to large and
308 small TR sequences, respectively. Taken together, 20.13% of all unique spacers (396,424 spacers)
309 were mapped to TR sequences. Compared to the stringent identity threshold initially applied (93%),
310 91.91% more spacers were mapped to small TR sequences with relaxed threshold. By comparison,
311 only 42.68% more spacers were mapped to large TR sequences with relaxed threshold. These
312 results suggest that a substantial fraction of CRISPR spacers imperfectly matches to protospacers
313 within circular mobile genetic elements in human gut. This potentially reflects an 'escape mutation'

314 phenomenon. By percentage, small TR sequences can be explained in this manner than large TR
315 sequences.

316 While we focused TR sequences in this study because of high confidence of the genomic
317 completeness, spacers could be derived from incompletely reconstructed or non-circular genomes.
318 To estimate the contribution of discovered CRISPR-targeted sequences to identified CRISPR
319 spacers, all unique spacers were mapped to all representative CRISPR-targeted sequences with
320 84% sequence identity threshold. Under these conditions, 971,224 spacers (49.31% of all unique
321 spacers) were mapped.

322 Overall, spacer mapping with looser criterion suggesting that at least about one-fifth of
323 discovered CRISPR spacers originated from TR sequences or their recognizable evolutionary
324 predecessors, whereas about half of CRISPR spacers originated from our discovered CRISPR-
325 targeted sequences including both TR and non-TR.

326

327 **Conclusion**

328 In the current study, we demonstrated that CRISPR spacers can be used as a tool to detect viral
329 genomes and other mobile genetic elements from metagenome sequences. Using spacers to
330 confidently infer which sequences are targeted by CRISPR, we substantially expanded the
331 diversity of mobile genetic elements identifiable from the human gut metagenome, which has
332 already been a topic of intense investigation for virus discovery. Comparing the sequences
333 predicted by this approach against viral databases showed that our protocol effectively detects viral
334 genomes without requiring similarity to any known viral sequence. Although the majority of large
335 (>20 kb) genomes were predicted as tailed phages with high confidence based on sequence
336 homology, we found that the majority of small (<20 kb) genomes remain unclassified because of

337 a lack of similar genomes in annotated databases. We showed that the source of nearly half of the
338 CRISPR spacers encoded by residents of the human gut remains unknown, suggesting that
339 additional protospacer reservoirs, whether extinct or simply unsampled, are still to be characterized.
340 Applying this conceptual advance to additional metagenomic datasets will increase the breadth of
341 the lens through which we can study the diversity of Earth's virome.
342

343 **Materials and Methods**

344 **Materials.** Sequencing data were selected based on NCBI metadata. Filtering parameters for the
345 query were as follows: layout = PAIRED, platform = ILLUMINA, selection = RANDOM, strategy
346 = WGS, source = METAGENOME, NCBI Taxonomy = 408170 (human gut metagenome) and
347 minimum library size = 1 Gb. If a sample contained multiple runs, we selected the run with the
348 most bases for simplicity of the analysis pipeline and to avoid possible bias to protospacer counts
349 from nearly identical metagenomes.
350

351 **Database versions and download dates.** RefSeq Release 98 was downloaded on 10 January 2020
352 and IMG/VR Release Jan. 2018 was downloaded on 21 October 2019. GVD was downloaded on
353 11 March 2020 and UniRef50 was downloaded on 16 December 2019.
354

355 **Metagenome assembly.** All downloaded paired FASTQ files were pre-processed based on
356 guidance in BBTools[49] (version 38.73). Adapters, phi X and human sequences were removed using
357 BBDuk and BBMap. Sequencing errors were corrected using Tadpole. Each pre-processed pair of
358 FASTQ files was assembled using SPAdes[50] (version 3.12) with -meta option. Contigs smaller
359 than 1 kb were discarded.

360

361 **Detection of CRISPR and spacer extraction.** Assembled contigs were scanned with
362 CRISPRDetect[51] (version 2.2) to extract CRISPR DRs which were deduplicated using CD-HIT-
363 EST[52] (version 4.7) and used to mask the raw reads using BBDuk. We extracted CRISPR spacers
364 from the raw reads to maximize spacer capture from the library. Sequences between the masked
365 regions within the raw reads were considered CRISPR spacers which were extracted by a simple
366 Python program (available in our source code repository) and then deduplicated.

367

368 **Detection of protospacer loci.** All DRs were mapped to contigs using BBMap with a 93%
369 minimum sequence identity. DR mapped positions and their flanking 60 bases were masked as
370 CRISPR loci. Next, identified spacers were mapped to all CRISPR masked contigs with a 93%
371 minimum sequence identity. Rather than excluding all contigs with CRISPR loci, we exclusively
372 masked CRISPR loci to identify viruses that encode CRISPR systems which are themselves
373 targeted by other CRISPR systems. To increase specificity, we aligned the 5' and 3' adjacent
374 regions of spacer-mapped positions. These adjacent sequences were also aligned to the DR
375 sequence associated with the mapped spacer. We discarded those loci in which any alignment
376 score divided by length was higher than 0.5, using the alignment parameters: match = 1, mismatch
377 = -1, gap = -1 and gap extension = -1. The remaining positions were considered authentic
378 protospacer loci.

379

380 **Co-occurrence-based spacer clustering.** We clustered spacers in two steps (Supplementary Fig.
381 2). First, we clustered protospacer loci located within 50 kb of another protospacer locus. We then
382 further clustered spacers based on the co-occurrence of protospacers represented as a graph. In this

383 graph, protospacers are nodes and the edges represent co-occurrence of connected protospacers.
384 The weights of edges were observed counts of co-occurrence of the connected protospacers
385 defined in the previous clustering. Graph communities were detected using Markov Clustering
386 algorithm[53] (options: -I 4 -pi 0.4; version 14-137). Clusters with a size lower than 10 and a global
387 clustering coefficient lower than 0.5 were discarded. Finally, 12,749 clusters comprising 591,189
388 spacers were derived.
389

390 **Extraction of CRISPR-targeted sequences.** Contiguous regions of contigs targeted by more than
391 30% of members of a spacer cluster were marked as a bed file using BEDTools[54]. To join the
392 fragmented clusters, adjacent regions within 1 kb were concatenated. Marked regions were
393 extracted, and sequences containing assembly gap were discarded. Finally, both ends of each
394 extracted sequence were compared to identify TR sequences using a python program utilizing the
395 Biopython[55] package (available in our source code repository).
396

397 **Deduplication of CRISPR-targeted sequences.** TR sequences were clustered using PSI-CD-HIT
398 (options: -c 0.95 -aS 0.95 -aL 0.95 -G 1 -g 1 -prog blastn -circle 1). The remaining CRISPR-
399 targeted sequences were clustered twice using linclust[56] (options: --cluster-mode 2 --cov-mode 1
400 -c 0.9 --min-seq-id 0.95), then clustered again using PSI-CD-HIT (options: -c 0.9 -aS 0.95 -G 1 -
401 g 1 -prog blastn -circle 1).
402

403 **Gene prediction and annotation of CRISPR-targeted sequence.** Protein-coding genes were
404 predicted from both TR and non-TR sequences using Prodigal (version 2.6.3) with the -p meta

405 option. To recover truncated genes from TR sequences, each TR sequence was concatenated in
406 silico and unique predicted genes were selected. Predicted protein sequences with partial frags
407 were discarded. The remaining protein sequences were clustered based on a 30% sequence identity
408 threshold and aligned using mmseqs[57] (version e1a1c1226ef22ac3d0da8e8f71adb8fd2388a249).
409 HMMs were constructed from each aligned sequence using hmmbuild from HMMER[58] (version
410 3.2.1). Constructed HMMs were then used as query to search against UniRef50 and RefSeq viral
411 proteins (E value < 10e-10) using hmmsearch.
412

413 **Capsid HMM database construction.** We selected capsids and coat proteins from RefSeq viral
414 protein sequences. Sequences containing 'capsid' and 'coat' in the description were selected and
415 then manually inspected to remove sequences unlikely to actually be capsid, such as containing
416 the term 'associate' or phrase 'assembly chaperone'. A total of 5,069 sequences were clustered
417 based on a 30% sequence identity, resulting in 1,324 representative sequences. Each representative
418 sequence was used as bait and underwent five jackhmmer iterations on UniRef50 database. HMMs
419 from the final iteration were used to build the capsid HMM database.
420

421 **Targeting host prediction.** DRs were mapped to RefSeq bacteria and archaea genomes using
422 BBMap. A locus with more than three consecutive DR hits within 100 bases was considered an
423 authentic CRISPR locus associated with the mapped DR. DRs mapped to multiple taxa in a given
424 taxonomic level were not taxonomically assigned for that level. DRs assigned to taxa were used
425 to predict the targeting host. We counted the protospacers linked to taxonomically assigned DRs
426 within a TR sequence. If the count of a given taxon was ≥10 and higher than 90% exclusiveness,

427 we considered that the corresponding taxon is a targeting host of a given contig. Host prediction
428 was done for each taxonomic level: species, genus, family, order, class, phylum and domain.
429

430 **Gene content-based hierarchical clustering of TR sequences.** TR sequences were scanned by
431 HMMs derived from the clustering results of the predicted protein sequences using both TR and
432 non-TR sequences. The scanned result was represented by a binary matrix (score > 60). We
433 selected the top 1,000 genes recurrently observed within TR sequences. The matrix was
434 hierarchically clustered with ComplexHeatmap[59] (version 2.5.3), and then annotated with database
435 hits, host, GC content, capsids and predicted gene functions. Clustering was separately performed
436 for large and small TR sequences.
437

438 **Phylogenetic analysis of *Microviridae* major capsid proteins.** Representative *Microviridae*
439 major capsid protein sequences were selected by clustering all capsids based on an 85% sequence
440 identity threshold throughout the entire length of the protein. Representative and reference protein
441 sequences were aligned using MAFFT[60] (version 7.310) and then trimmed using trimAl[61] (version
442 1.4). Aligned sequences were used for Bayesian phylogenetic analysis using MrBayes[62] (version
443 3.2.7). A mixed substitution model with a uniform prior that converged to Blosum62 (posterior
444 probability = 1.000) was selected. All other priors were set to default. Two chains of Markov chain
445 Monte Carlo (MCMC) with identical priors run over ten million generations and sampled every
446 500 generations. The standard deviation of split frequencies approached close to zero (0.007837)
447 over the run.
448

449 **Phylogenetic analysis of crAssphage core genes.** 55 core crAssphage genes which were encoded
450 in all crAssphage genomes from our study and the reference crAssphage genome were selected.
451 Representative crAssphage genomes were selected by clustering RNA polymerase coding
452 sequences based on a 97.5% sequence identity threshold. Each gene was aligned using MAFFT
453 and then trimmed using trimAl. Aligned core genes were then concatenated and used for Bayesian
454 phylogenetic analysis using MrBayes. A GTR + I + Γ substitution model was selected. Two chains
455 of MCMC with identical priors were run over ten million generations and sampled every 500
456 generations. The standard deviation of split frequencies approached close to zero (0.000499) over
457 the run.
458
459 **Generation of scrambled sequences.** Scrambled sequences are random sequences with identical
460 lengths to the TR sequences. The sequences are generated based on sampled nucleotide
461 frequencies from the TR sequences using Biopython package (available in our source code
462 repository).
463
464 **Code availability.** The source codes used in this study are available at https://github.com/ryota-
465 sugimoto/virome_scripts
466
467 **Data availability.** Data set including discovered CRISPR spacers, direct repeats, protospacers, co-
468 occurrence-based spacer clustering result, predicted protein sequences, built HMMs, database
469 comparison results, phylogenetic analysis results and CRISPR-targeted TR sequences are
470 available in our google drive. (URL:

https://drive.google.com/drive/folders/1QwpG3TilpBFwmOD5_qi2NESWA_XC0_g0?usp=sharing MD5: d0f2ebe8dd20de4033a68735015943d4)

473 **References**

474 1. Wommack, K. E. & Colwell, R. R. Virioplankton: viruses in aquatic ecosystems. *Microbiol*
475 *Mol Biol Rev* **64**, 69-114 (2000).
476 2. HERSHEY, A. D. & CHASE, M. Independent functions of viral protein and nucleic acid in
477 growth of bacteriophage. *J Gen Physiol* **36**, 39-56 (1952).
478 3. Fiers, W. *et al*. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and
479 secondary structure of the replicase gene. *Nature* **260**, 500-507 (1976).
480 4. Sanger, F. *et al*. Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687-695
481 (1977).
482 5. Koonin, E. V., Senkevich, T. G. & Dolja, V. V. The ancient Virus World and evolution of
483 cells. *Biol Direct* **1**, 29 (2006).
484 6. Krupovic, M., Dolja, V. V. & Koonin, E. V. Origin of viruses: primordial replicators
485 recruiting capsids from hosts. *Nature Reviews Microbiology* **17**, 449-458 (2019).
486 7. Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular
487 biological access to the chemistry of unknown soil microbes: a new frontier for natural
488 products. *Chem Biol* **5**, R245-9 (1998).
489 8. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from
490 microbial genomic data. *PeerJ* **3**, e985 (2015).
491 9. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer based
492 tool for identifying viral sequences from assembled metagenomic data. *Microbiome* **5**, 69
493 (2017).
494 10. Gregory, A. C., Zablocki, O., Howell, A., Bolduc, B. & Sullivan, M. B. The human gut
495 virome database. *bioRxiv* doi: https://doi.org/10.1101/655910

496 11. Barrangou, R. *et al*. CRISPR provides acquired resistance against viruses in prokaryotes.
497 *Science* **315**, 1709-1712 (2007).
498 12. Seed, K. D., Lazinski, D. W., Calderwood, S. B. & Camilli, A. A bacteriophage encodes its
499 own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* **494**, 489-491
500 (2013).
501 13. Al-Shayeb, B. *et al*. Clades of huge phages from across Earth's ecosystems. *Nature* **578**, 425-
502 431 (2020).
503 14. Andersson, A. F. & Banfield, J. F. Virus population dynamics and acquired virus resistance
504 in natural microbial communities. *Science* **320**, 1047-1050 (2008).
505 15. Snyder, J. C., Bateson, M. M., Lavin, M. & Young, M. J. Use of Cellular CRISPR (Clusters
506 of Regularly Interspaced Short Palindromic Repeats) Spacer-Based Microarrays for
507 Detection of Viruses in Environmental Samples. *Applied and Environmental Microbiology*
508 **76**, 7251-7258 (2010).
509 16. Zhang, Q., Rho, M., Tang, H., Doak, T. G. & Ye, Y. CRISPR-Cas systems target a diverse
510 collection of invasive mobile genetic elements in human microbiomes. *Genome Biol* **14**, R40
511 (2013).
512 17. Shmakov, S. A. *et al*. The CRISPR spacer space is dominated by sequences from species-
513 specific mobilomes. *MBio* **8**, e01397-17 (2017).
514 18. Shmakov, S. A., Wolf, Y. I., Savitskaya, E., Severinov, K. V. & Koonin, E. V. Mapping
515 CRISPR spaceromes reveals vast host-specific viromes of prokaryotes. *Commun Biol* **3**, 321
516 (2020).
517 19. Paez-Espino, D. *et al*. Uncovering Earth's virome. *Nature* **536**, 425-430 (2016).

518 20. Stern, A., Mick, E., Tirosh, I., Sagy, O. & Sorek, R. CRISPR targeting reveals a reservoir of
519 common phages associated with the human gut microbiome. *Genome Res* **22**, 1985-1994
520 (2012).
521 21. Dutilh, B. E. *et al*. A highly abundant bacteriophage discovered in the unknown sequences of
522 human faecal metagenomes. *Nat Commun* **5**, 4498 (2014).
523 22. Yutin, N. *et al*. Discovery of an expansive bacteriophage family that includes the most
524 abundant viruses from the human gut. *Nat Microbiol* **3**, 38-46 (2018).
525 23. STREISINGER, G., EDGAR, R. S. & DENHARDT, G. H. CHROMOSOME STRUCTURE
526 IN PHAGE T4. I. CIRCULARITY OF THE LINKAGE MAP. *Proc Natl Acad Sci U S A* **51**,
527 775-779 (1964).
528 24. Cui, J., Schlub, T. E. & Holmes, E. C. An allometric relationship between the genome length
529 and virion volume of viruses. *J Virol* **88**, 6403-6410 (2014).
530 25. Hua, J. *et al*. Capsids and Genomes of Jumbo-Sized Bacteriophages Reveal the Evolutionary
531 Reach of the HK97 Fold. *mBio* **8**, (2017).
532 26. Roux, S., Krupovic, M., Poulet, A., Debroas, D. & Enault, F. Evolution and diversity of the
533 Microviridae viral family through a collection of 81 new complete genomes assembled from
534 virome reads. *PloS one* **7**, e40418 (2012).
535 27. Ackermann, H. W. Tailed bacteriophages: the order caudovirales. *Adv Virus Res* **51**, 135-201
536 (1998).
537 28. Ackermann, H. W. Phage classification and characterization. *Methods Mol Biol* **501**, 127-140
538 (2009).
539 29. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity,
540 stability and resilience of the human gut microbiota. *Nature* **489**, 220-230 (2012).

541 30. Almpanis, A., Swain, M., Gatherer, D. & McEwan, N. Correlation between bacterial G+C
542 content, genome size and the G+C content of associated plasmids and bacteriophages.
543 *Microb Genom* **4**, (2018).
544 31. Brentlinger, K. L. *et al*. Microviridae, a family divided: isolation, characterization, and
545 genome sequence of φMH2K, a bacteriophage of the obligate intracellular parasitic
546 bacterium Bdellovibrio bacteriovorus. *Journal of bacteriology* **184**, 1089-1094 (2002).
547 32. Sharp, P. M. & Hahn, B. H. Origins of HIV and the AIDS pandemic. *Cold Spring Harb*
548 *Perspect Med* **1**, a006841 (2011).
549 33. Zhou, P. *et al*. A pneumonia outbreak associated with a new coronavirus of probable bat
550 origin. *Nature* **579**, 270-273 (2020).
551 34. Watson, B. N. J. *et al*. Type IF CRISPR-Cas resistance against virulent phages results in
552 abortive infection and provides population-level immunity. *Nature communications* **10**, 1-8
553 (2019).
554 35. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a curated
555 non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**,
556 D61-5 (2007).
557 36. Paez-Espino, D. *et al*. IMG/VR: a database of cultured and uncultured DNA Viruses and
558 retroviruses. *Nucleic Acids Res* **45**, D457-D465 (2017).
559 37. Paez-Espino, D. *et al*. IMG/VR v.2.0: an integrated data management and analysis system for
560 cultivated and environmental viral genomes. *Nucleic Acids Res* **47**, D678-D686 (2019).
561 38. Roux, S. *et al*. Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's
562 biomes. *Nat Microbiol* **4**, 1895-1906 (2019).

563 39. Krupovic, M. & Koonin, E. V. Multiple origins of viral capsid proteins from cellular
564 ancestors. *Proc Natl Acad Sci U S A* **114**, E2401-E2410 (2017).
565 40. Lindsay, J. A., Ruzin, A., Ross, H. F., Kurepina, N. & Novick, R. P. The gene for toxic shock
566 toxin is carried by a family of mobile pathogenicity islands in Staphylococcus aureus.
567 *Molecular microbiology* **29**, 527-543 (1998).
568 41. Krupovic, M., Ravantti, J. J. & Bamford, D. H. Geminiviruses: a tale of a plasmid becoming
569 a virus. *BMC Evol Biol* **9**, 112 (2009).
570 42. Morozova, V. *et al*. First crAss-Like Phage Genome Encoding the Diversity-Generating
571 Retroelement (DGR). *Viruses* **12**, (2020).
572 43. Paul, B. G. *et al*. Retroelement-guided protein diversification abounds in vast lineages of
573 Bacteria and Archaea. *Nat Microbiol* **2**, 17045 (2017).
574 44. Liu, M. *et al*. Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage.
575 *Science* **295**, 2091-2094 (2002).
576 45. Guo, H., Arambula, D., Ghosh, P. & Miller, J. F. Diversity-generating Retroelements in
577 Phage and Bacterial Genomes. *Microbiol Spectr* **2**, (2014).
578 46. Edwards, R. A. *et al*. Global phylogeography and ancient evolution of the widespread human
579 gut virus crAssphage. *Nat Microbiol* **4**, 1727-1736 (2019).
580 47. Deveau, H. *et al*. Phage response to CRISPR-encoded resistance in Streptococcus
581 thermophilus. *Journal of bacteriology* **190**, 1390-1400 (2008).
582 48. Fineran, P. C. *et al*. Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc*
583 *Natl Acad Sci U S A* **111**, E1629-38 (2014).
584 49. Bushnell, B. BBTools software package. URL http://sourceforge.net/projects/bbmap (2014).

585 50. Bankevich, A. *et al*. SPAdes: A New Genome Assembly Algorithm and Its Applications to
586 Single-Cell Sequencing. *Journal of Computational Biology* **19**, 455-477 (2012).
587 51. Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C. & Brown, C. M. CRISPRDetect: a
588 flexible algorithm to define CRISPR arrays. *BMC genomics* **17**, 356 (2016).
589 52. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein
590 or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).
591 53. Van Dongen, S. M. Graph clustering by flow simulation.
592 https://dspace.library.uu.nl/bitstream/handle/1874/848/full.pdf (2000).
593 54. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic
594 features. *Bioinformatics* **26**, 841-842 (2010).
595 55. Cock, P. J. A. *et al*. Biopython: freely available Python tools for computational molecular
596 biology and bioinformatics. *Bioinformatics* **25**, 1422-1423 (2009).
597 56. Steinegger, M., Söding, J. Clustering huge protein sequence sets in linear time. *Nat*
598 *Commun* **9,** 2542 (2018).
599 57. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the
600 analysis of massive data sets. *Nature biotechnology* **35**, 1026-1028 (2017).
601 58. Eddy, S. R. & team, H. M. M. E. R. D. URL http://hmmer.org (2019).
602 59. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in
603 multidimensional genomic data. *Bioinformatics* **32**, 2847-2849 (2016).
604 60. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid multiple
605 sequence alignment based on fast Fourier transform. *Nucleic acids research* **30**, 3059-3066
606 (2002).

607 61. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated
608 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973
609 (2009).
610 62. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees.
611 *Bioinformatics* **17**, 754-755 (2001).

Acknowledgements

The majority of analysis has been done on the supercomputer systems in National Institute of Genetics (NIG) and Okinawa Institute of Science and Technology Graduate University. R.S. is supported by funds of Database Center for Life Science, Research Organization of Information and Systems, and MEXT KAKENHI Grant Number 20K21405. R.S. thanks to NIG Research Administrator room for scientific discussions.

Author contributions

R.S. designed the studies and performed most of the analyses. L.N. assisted in sequence alignment and phylogenetic analysis. P.N.T. assisted in metagenome assembly. J.I. gave scientific input in the data analyses and interpretation of the results. N.F.P. gave scientific input and helped write the manuscript. H.M., K.K and H.N. provided scientific advice on the entire study. I.I. designed the study and wrote the manuscript together with R.S.

Competing interests

The authors declare no competing financial interests.

629 **Figure legends**

630 **Fig. 1 | Properties of CRISPR-targeted TR sequences. a**, Length distribution of TR sequences.
631 We used portal, tail and terminase large subunit as marker genes of tailed phage; sequences
632 encoding all three of these genes were categorized as tailed phages. The dotted line at 20 kb
633 represents an arbitrary cut-off between small and large sequences. Sequences longer than 100 kb
634 are shown in the inset. **b**, Predicted CRISPR targeting host composition of TR sequences. Hosts
635 were predicted by mapping CRISPR DR sequences to the RefSeq database. Sequences containing
636 ≥10 protospacer loci but less than 90% associated DR taxa exclusiveness were considered targeted
637 by multiple phyla. If ≥10 protospacers could not be assigned to a taxon, the predicted targeting
638 host is denoted as not available (NA). **c**, Heterogeneous distribution of TR sequences targeted by
639 multiple phyla. Circle size approximately represents popularity of the respective host.
640 Bidirectional arrows connect the top two host phyla according to host-assigned protospacer counts
641 (i.e. protospacers most often associated with CRISPR DRs are assigned to these two phyla).
642 Numbers on arrows are counts of the number of TR sequences targeted by the connected phyla. **d**,
643 Predicted targeting host distribution by GC content. The dotted line indicates the low- and high-
644 GC content boundary at 55%.

645

646 **Fig. 2 | Bayesian phylogeny of *Microviridae* major capsid proteins.** A total of 159 representative
647 major capsid protein sequences from this study and 43 RefSeq major capsid protein sequences
648 were used for analysis. Taxa without a name denote the *Microviridae* species from this study and
649 taxa with text denote *Microviridae* species from RefSeq. Taxa were annotated based on predicted
650 targeting hosts. The phi X174 clade was selected as the outgroup.

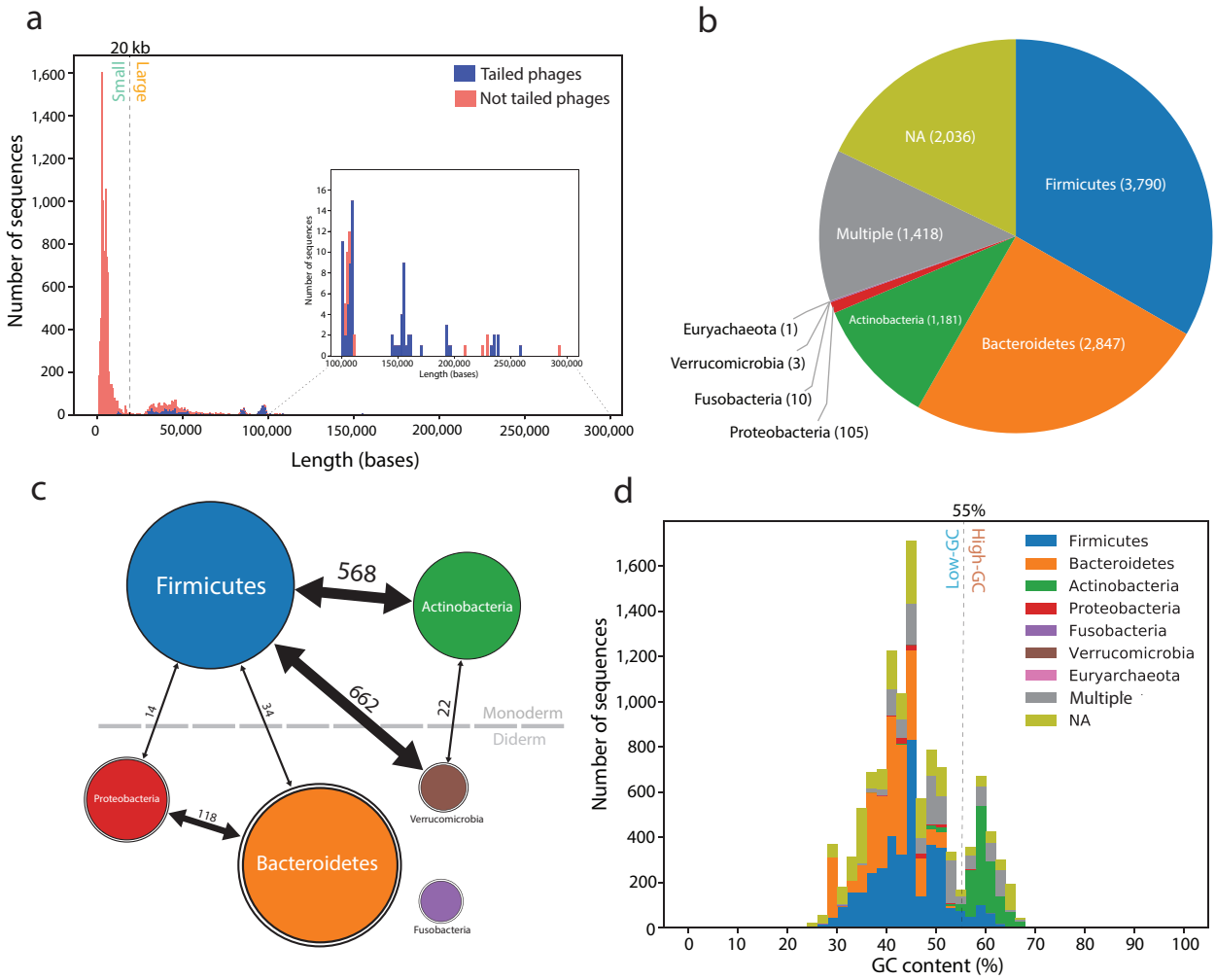
651

652 **Fig. 3 | Venn diagrams of database comparisons for (a) large and (b) small TR sequences.**
653 Each TR sequence was compared to RefSeq virus, RefSeq plasmid, IMG/VR and GVD using
654 BLASTN. Database hit minimum criteria was set to 85% sequence identity with 75% aligned
655 fraction of query sequence to a unique subject sequence.
656
657 **Fig. 4 | Hierarchical clustering of (a) large and (b) small TR sequences based on gene content.**
658 Heatmaps representing gene content of TR sequences, in which each row is a TR sequence and
659 each column is a gene cluster. Gray areas in the heatmap indicate sequences encoding a gene that
660 is homologous to the gene cluster. Note that one gene can be homologous to multiple gene clusters.
661 Sequences are annotated by database containing similar sequences, GC content, host and
662 detectable capsid genes. Gene clusters were annotated by searching corresponding HMMs to the
663 UniRef50 database. Several notable RefSeq-listed clusters are denoted on the right side of the
664 heatmaps.
665
666 **Fig. 5 | DGR locus and phylogeny of crAssphage core genes. a**, Comparison of the DGR locus
667 between a reference crAssphage genome (NC_024711.1), a representative from this study
668 (crAssphage.202.rt) and a baboon-derived crAssphages (Baboon19.1). The retroelement-insertion
669 region is enclosed by a dotted line. Sequence conservation of DGR and its target gene (tail collar
670 fibre) were calculated by aligning and building HMMs. Sequence conservation of the DGR target
671 gene was calculated separately based on whether the genome encodes DGR (red) or not (blue).
672 Information content was averaged over seven base-window sizes. **b**, Pairwise sequence alignment
673 of template (top) and variable (bottom) repeats from Baboon19.1. **c**, Unrooted Bayesian phylogeny

674 of crAssphage core genes. 55 core genes from 20 representative crAssphages genomes and one
675 reference genome were used for analysis. Those crAssphages with DGR appear in red.
676

677 **Fig. 6 | Number of mapped spacers according to various sequence identity thresholds.** All
678 unique CRISPR spacers were mapped to large TR sequences, small TR sequences and scrambled
679 sequences. The initially applied and relaxed sequence identity thresholds are denoted as red and
680 orange colored dashed lines respectively. The spacer mapping process is identical to the
681 protospacer discovery process (see Methods).
682

683 **Fig. 1**

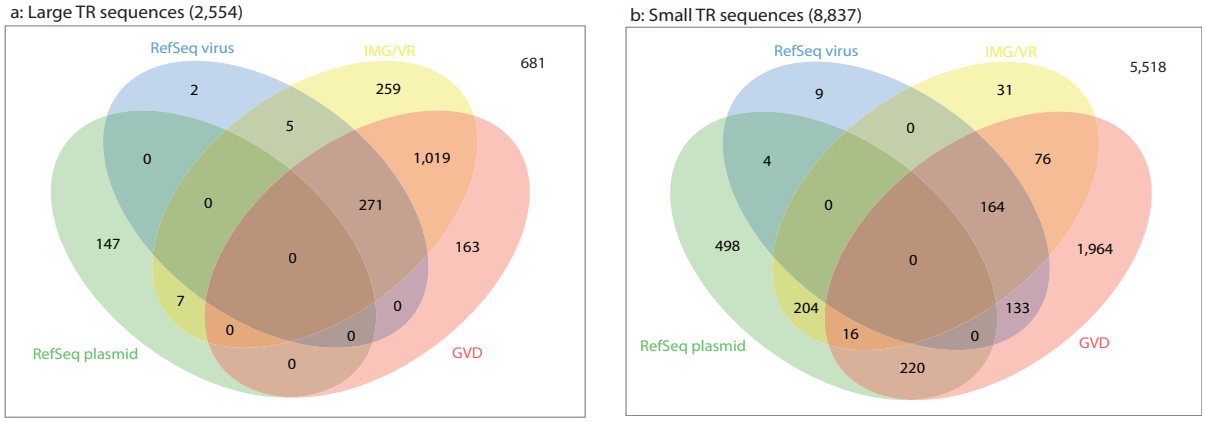

a
20 kb
Small
Large
Tailed phages
Not tailed phages
Number of sequences
Length (bases)
b
NA (2,036)
Firmicutes (3,790)
Multiple (1,418)
Actinobacteria (1,181)
Bacteroidetes (2,847)
Euryachaeota (1)
Verrucomicrobia (3)
Fusobacteria (10)
Proteobacteria (105)
c
Firmicutes
568
Actinobacteria
662
14
34
22
Monoderm
Diderm
Proteobacteria
118
Bacteroidetes
Verrucomicrobia
Fusobacteria
d
55%
Low-GC
High-GC
Firmicutes
Bacteroidetes
Actinobacteria
Proteobacteria
Fusobacteria
Verrucomicrobia
Euryarchaeota
Multiple
NA
Number of sequences
GC content (%)

684

Fig. 2

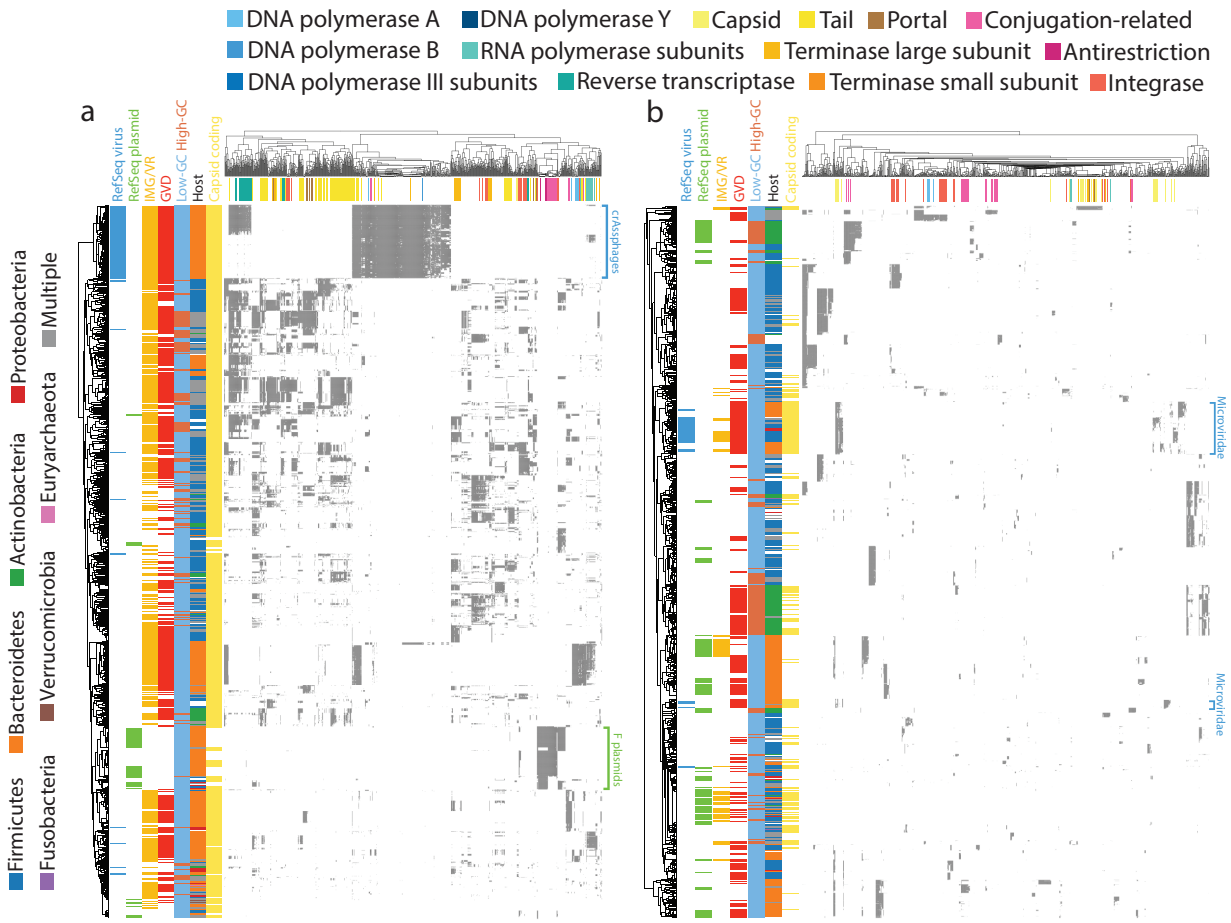
Host phylum
Bacteroidetes
Firmicutes
Proteobacteria
Gokushovirinae Bog1183 53
Chimpanzee faeces associated microphage 1
Chimpanzee faeces associated microphage 3
Escherichia phage EC6098
Alces alces faeces associated microvirus MP11 5517
Bdellovibrio phage phiMH2K
Lynx canadensis associated microvirus CLP 9413
Chimpanzee faeces associated microphage 2
Microviridae phi-CA82
Spiroplasma virus SpV4
Alces alces faeces associated microvirus MP10 5560
Alces alces faeces associated microvirus MP12 5423
Marine gokushovirus
Gokushovirinae Bog5712 52
Gokushovirinae Fen7875 21
Gokushovirinae Fen672 31
Gokushovirinae Bog8989 22
Chlamydia phage CPG1
Chlamydia phage 2
Chlamydia virus Chp1
Microviridae Bog1249 12
Microviridae Fen4707 41
Alces alces faeces associated microvirus MP21 4718
Microviridae Fen7940 21
Parabacteroides phage YZ-2015b
Parabacteroides phage YZ-2015a
Microviridae Fen685 11
Microviridae Fen7895 21
Microviridae Bog5275 51
Microviridae Fen7786 21
Microviridae Fen418 41
Microviridae Fen2266 11
Microviridae Bog9017 22
Microviridae Fen7918 21
Alces alces faeces associated microvirus MP15 5067
Lepus americanus faeces associated microvirus SHP1 6472
Microviridae IME-16
Alces alces faeces associated microvirus MP3 6497
Escherichia phage G4
Escherichia phageaID21
Escherichia virus phiX174
Escherichia phage St-1
Eel River basin pequenovirus
0.3

687 **Fig. 3**


a: Large TR sequences (2,554)
RefSeq virus
IMG/VR
681
2
259
5
0
1,019
0
271
147
0
163
7
0
0
0
0
RefSeq plasmid
GVD
b: Small TR sequences (8,837)
RefSeq virus
IMG/VR
5,518
9
31
0
4
76
0
164
498
0
1,964
204
133
16
0
220
RefSeq plasmid
GVD

688

689 **Fig. 4**


DNA polymerase A
DNA polymerase B
DNA polymerase III subunits
DNA polymerase Y
RNA polymerase subunits
Reverse transcriptase
Capsid
Tail
Terminase large subunit
Terminase small subunit
Portal
Conjugation-related
Antirestriction
Integrase
a
b
RefSeq virus
RefSeq plasmid
IMG/VR
GVD
Low-GC
High-GC
Host
Capsid coding
Firmicutes
Fusobacteria
Bacteroidetes
Verrucomicrobia
Actinobacteria
Euryarchaeota
Proteobacteria
Multiple
crAssphages
F plasmids
Microviridae
Microviridae

690

Fig. 5

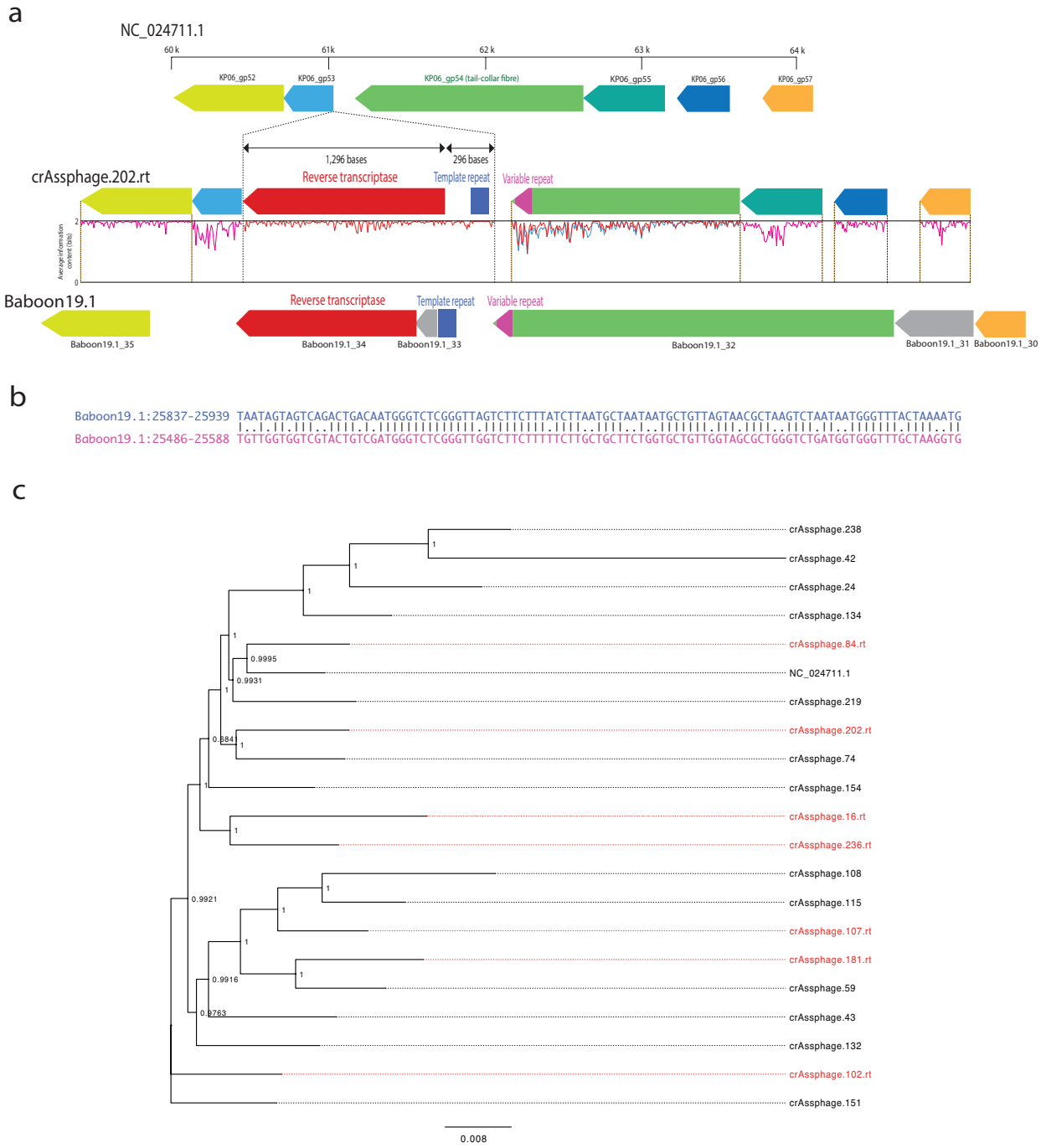
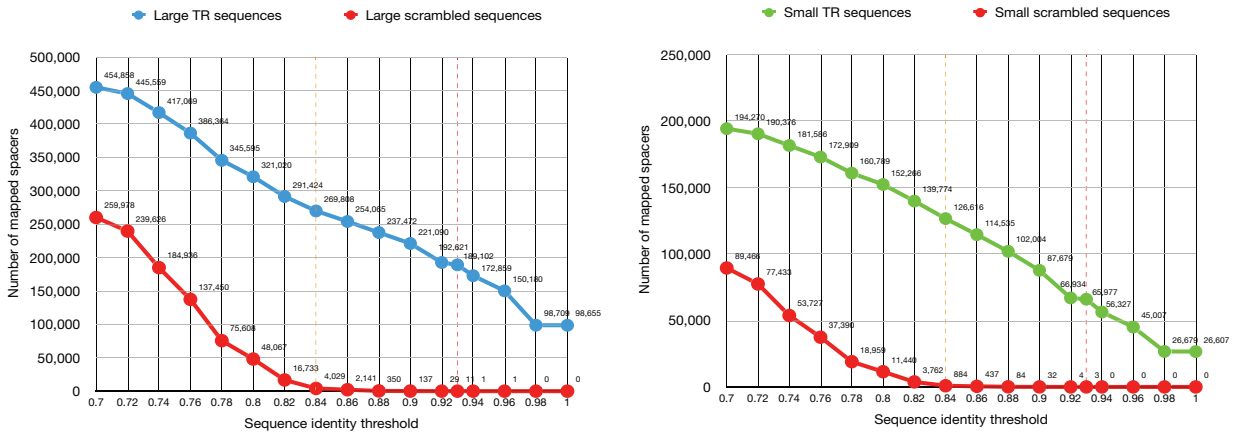

a
NC_024711.1
60 k
61k
62 k
63 k
64 k
KP06_gp52
KP06_gp53
KP06_gp54 (tail-collar fibre)
KP06_gp55
KP06_gp56
KP06_gp57
1,296 bases
296 bases
crAssphage.202.rt
Reverse transcriptase
Template repeat
Variable repeat
Average information content (bits)
Baboon19.1
Baboon19.1_35
Baboon19.1_34
Baboon19.1_33
Baboon19.1_32
Baboon19.1_31
Baboon19.1_30
b
Baboon19.1:25837-25939 TAATAGTAGTCAGACTGACAATGGGTCTCGGGTTAGTCTTCTTTATCTTAATGCTAATAATGCTGTTAGTAACGCTAAGTCTAATAATGGGTTTACTAAAATG
Baboon19.1:25486-25588 TGTTGGTGGTCGTACTGTCGATGGGTCTCGGGTTGGTCTTCTTTTCTTGCTGCTTCTGGTGCTGTTGGTAGCGCTGGGTCTGATGGTGGGTTTGCTAAGGTG
c
crAssphage.238
crAssphage.42
crAssphage.24
crAssphage.134
crAssphage.84.rt
NC_024711.1
crAssphage.219
crAssphage.202.rt
crAssphage.74
crAssphage.154
crAssphage.16.rt
crAssphage.236.rt
crAssphage.108
crAssphage.115
crAssphage.107.rt
crAssphage.181.rt
crAssphage.59
crAssphage.43
crAssphage.132
crAssphage.102.rt
crAssphage.151
0.9995
0.9931
0.9921
0.9916
0.008

Fig. 6


Large TR sequences
Large scrambled sequences
Number of mapped spacers
Sequence identity threshold
Small TR sequences
Small scrambled sequences
Number of mapped spacers
Sequence identity threshold