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

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

1) 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) 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 a virus inference method based on clustered regularly interspaced short palindromic repeats (CRISPR). CRISPR is a prokaryotic nucleic acid restriction system that stores memory of previous exposure. Our method can reconstruct 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 method 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.

Introduction

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

There are three major approaches to sequence viral genomes: virus culture, virus-like particle filtration and metagenomics. Virus culture provides high sequence quality and depth with little contamination, and also allows for inference and detection of what should be considered contamination. However, it is limited to providing the sequence of viruses which replicate in host cells that can be cultured, a condition likely comprising a tiny fraction of the virome. Virus-like particle filtration[7] provides high quality sequences of a variety of viruses in a given sample;

64 however, filtration excludes large viruses and viruses that have already integrated into the genome
65 of their host (pro-viruses). In contrast, metagenomic[8] approaches sequence all genetic materials
66 extracted from a given organism or tissue. The advantage of a metagenomic approach is, with
67 sufficient depth of sequencing, presumably all genetic information of viruses in the sample can be
68 obtained. However, because metagenomic approaches obtain a mixture of cellular and viral
69 genomic information, it is necessary to subsequently distinguish viral from cellular sequences.

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

87 materials in specialized genomic loci between CRISPR direct repeats (DRs). The incorporated
88 sequences, called 'spacers', are identical in sequence to part of the previously infecting mobile
89 genetic element. Thus, the genetic information encoded in CRISPR spacers can be inferred as
90 likely viral, and distinguishable to the genetic material of the organism encoding CRISPR which
91 is most often cellular, but potentially also viral[13,14].

92 CRISPR spacers have been used to detect viral genomes[15–18] and predict viral hosts[19,20].
93 They have previously been extracted from assembled bacterial genomes to assess CRISPR 'dark
94 matter', finding that 80–90% of identifiable material matches known viral genomes[17,18]. In the
95 current study, we extended this conceptual approach to unassembled short-read metagenomic data.
96 CRISPR repeats are relatively easily identifiable, particularly compared to unknown viral
97 sequences. This trait enables us to search immense metagenomic datasets for reads comprised in
98 part as CRISPR DR sequences, and in other parts, unknown sequences inferred as CRISPR spacers.
99 By analysing these reads and the contigs assembled from them, we successfully extracted 11,391
100 terminally redundant (TR) CRISPR-targeted sequences ranging from 894 to 292,414 bases. These
101 sequences are expected to be complete or near-complete circular genomes which can be linked to
102 their predicted hosts. The sequences we inferred de novo in this manner included 257 complete
103 genomes of the crAssphage family[21,22], 11 genomes larger than 200 kilobases (kb), 766 genomes
104 of *Microviridae* species, 114 genomes of *Inoviridae* species and many novel, taxonomically
105 unplaced genomes. Furthermore, by comparing the spacer-associated metagenomic CRISPR DR
106 sequences to CRISPRs from reference bacterial and archaeal genomes, we predicted CRISPR
107 targeting hosts for 69.7% of the discovered genomes.

108

109 **Results and Discussion**

110 **Extraction of CRISPR-targeted sequences.** For this study, we analysed human gut metagenomes
111 as they serve as an 'ecosystem' with the most abundant metagenomic data available to our
112 knowledge. We downloaded 11,817 human gut metagenome datasets equivalent to 50.7 Tb from
113 the European Nucleotide Archive FTP server. FASTQ files were pre-processed and assembled to
114 180,068,349 contigs comprising 767.7 Gb of data (Supplementary Table 1). We discovered 11,223
115 unique CRISPR DRs from the assembled contigs which were used to extract CRISPR spacers from
116 raw reads, resulting in 1,969,721 unique CRISPR spacers. These spacers were then used as query
117 to reveal candidate protospacers (i.e. contigs containing the spacer sequence not within a CRISPR
118 locus). Spacers were mapped to CRISPR masked contigs with a 93% minimum sequence identity
119 threshold. To increase specificity, we checked that the 5' and 3' adjacent sequences of spacer-
120 mapped positions were not similar either to each other or the spacer-associated DR. A total of
121 164,590,387 candidate protospacer loci, attributed to 1,114,947 unique spacers (56.6% of all
122 unique spacers) were revealed, showing a substantially higher discovery rate compared with that
123 of a previous study (~7%[17]) which used National Center for Biotechnology Information (NCBI)
124 nucleotide sequences for protospacer discovery. Spacers were clustered based on protospacer co-
125 occurrence and used to extract contigs targeted by more than 30% of members of a spacer cluster.
126 Finally, 764,883 gapless CRISPR-targeted sequences (15.9 Gb) were extracted (Supplementary
127 Fig. 1 and 2). Of these, 11,391 unique sequences were identified as TR[23], which are expected if
128 originating as complete or near-complete circular mobile genetic elements. The size of CRISPR-
129 targeted TR sequences ranged from 894 to 292,414 bases (Supplementary table 2 and
130 Supplementary data).

131 To maximize recovery of novel genes, 3,099,178 protein coding genes were predicted from
132 CRISPR-targeted sequences, including both TR (400,566 genes) and non-TR sequences

133 (2,698,612 genes). Protein sequences were clustered based on a 30% sequence identity threshold,
134 resulting in 83,796 clusters with more than two members and comprising 2,996,113 genes (96.7%).
135 Each cluster was aligned and used to build a hidden Markov model (HMM), then used as query
136 against UniRef50 and RefSeq viral proteins. We found that 12,283 HMMs (14.7%), comprising
137 891,121 genes (28.8%), were similar to 60,296 different RefSeq viral protein sequences (E-value
138 < 10^-10). A total of 6,853 TR sequences (60%) encoded at least one of these genes, suggesting
139 that a substantial portion of TR sequences are viral genomes.
140
141 **Length distribution of CRISPR-targeted TR sequences.** Evaluation of TR sequence length
142 showed a multi-modal distribution with a distinct trough at 20 kb (Fig. 1a). For reference, we refer
143 to the 8,837 TR sequences shorter than 20 kb as 'small' and the 2,554 TR sequences longer than
144 20 kb as 'large'. This simple classification was previously used to infer capsid morphology[24,25].
145 Among large TR sequences, 1,509 sequences (59%) encode a set of tail, portal and terminase large-
146 subunit genes (score > 30), whereas only 50 small TR sequences (0.6%) encoded this set of genes
147 (Supplementary Fig. 3), indicating that the majority of large TR sequences are tailed phages, also
148 known as *Caudovirales*. In addition, 766 small TR sequences, range in size between 4,574 and
149 12,988 bases (average 5,766.14 bases), encoded both *Microviridae* major capsid and replication
150 initiator homologues[26]. 114 small TR sequences, range in size between 4,783 and 18,988 (average
151 6,715.6 bases), encoded both *Inoviridae* major coat and replication initiator homologues. We
152 consider that this portion of small TR sequences are likely viruses with a non-tailed morphology.
153
154 **Predicted CRISPR targeting hosts of TR sequences.** As our approach uses CRISPR spacers to
155 extract CRISPR-targeted sequences, we hypothesised that the relationship between virus and

156 targeted host can be resolved for the majority of TR sequences. CRISPR DR sequences were
157 searched on RefSeq genomes and taxonomically assigned. Based on counts of protospacers
158 associated with taxonomically assigned DRs, 7,937 TR sequences (69.7%) were resolved to a
159 targeting host at the phylum level (Fig. 1b and Supplementary Fig. 4), and 6,083 TR sequences
160 (53.4%) were resolved to a targeting host at the order level (Supplementary Fig. 5 and
161 Supplementary table 2). The most frequent host was *Firmicutes*, followed by *Bacteroidetes* and
162 *Actinobacteria*. Notably, these are the most common bacteria in human intestine[27]. In addition,
163 1,418 TR sequences (12.5%) were predicted to be targeted by multiple phyla. Although some of
164 these TR sequences were targeted by multiple monoderm or diderm phyla exclusively, there was
165 an exceptional host ambiguity between *Firmicutes* and *Verrucomicrobia* which crosses the
166 monoderm-diderm boundary (Fig. 1c). We did not use CRISPR DR sequences shared between
167 cross-phyla species for targeting host prediction; targeting host ambiguity is unlikely because of
168 ambiguous taxonomic assignment of DR sequences. To determine whether the targeting host
169 predicted by this method corresponds with reported permissive hosts, we calculated GC content[28]
170 of the TR sequences (Fig. 1d). Among the 2,050 high-GC (GC% > 55%) TR sequences, 1,109
171 (54%) and 249 (12.1%) were predicted to be targeted by *Actinobacteria* and *Firmicutes*
172 respectively, and the targeting host was undetermined for 222 TR sequences (10.8%). The large
173 fraction of high-GC TR sequences predicted to be targeted by *Actinobacteria* likely indicate
174 genomic adaption of parasitic genetic elements which infect and routinely become targeted by
175 CRISPR systems of high-GC gram-positive bacteria.

176 Host range within a viral lineage was further assessed by phylogenetic analysis of TR
177 sequences determined to represent *Microviridae* species. The predicted targeting hosts of putative
178 *Microviridae* species from our study were *Bacteroides*, *Firmicutes* and *Proteobacteria*, indicating

179 a wide host range of *Microviridae* species. Molecular phylogeny of the *Microviridae* major capsid
180 protein clearly separates sequences based on their targeting host (Fig. 2). Interestingly, most
181 known *E. coli*-infecting viral species (such as phiX174) and other *Proteobacteria*-infecting species
182 (such as phi MH2K[29]) used as reference in this phylogeny split to two clades, and the clade
183 containing the latter was within a clade of TR sequences targeted by *Firmicutes*. The nested
184 phylogenetic structure of capsids encoded by TR sequences predicted to represent *Microviridae*
185 species may indicate that host-switching, a critically important topic in virus evolution[30,31], has
186 occurred within the viral lineage.

187 In this analysis, we predicted the CRISPR targeting host phylum of approximately 70% of
188 discovered sequences, suggesting that the majority of viruses in human gut are targeted by specific
189 host phyla. We also found that 12.5% of TR sequences were targeted by multiple phyla; however,
190 we are uncertain whether these elements actually infect multiple hosts at present, recently host-
191 switched, or some genomes become CRISPR-targeted as a result of abortive infections[32].
192 Importantly, expanding this approach to detect viral genomes from different environments may be
193 useful to infer the evolutionary history and genetic factors involved in host-switching.

194

195 **Comparison of CRISPR-targeted TR sequences to available viral and plasmid sequences.** TR
196 sequences identified in this study were compared to virus and plasmid genome databases,
197 including RefSeq[33] virus, RefSeq plasmid, IMG/VR[34,35] and GVD[11] (Fig. 3). IMG/VR is currently
198 the largest database of uncultivated viral genomes. GVD is a database of viral genomes discovered
199 from human gut metagenome datasets using VirSorter and VirFinder. Both of these databases rely
200 on protein homology, which complements our approach, to discover the viral genomes. Therefore,
201 we consider these databases are appropriate to validate our results. Among the 2,554 large TR

202 sequences, we found that 1,726 TR sequences (67.6%) were represented in RefSeq virus, IMG/VR
203 or GVD (Fig. 3a), using a threshold of 85% sequence identity with at least 75% aligned fraction
204 of the query sequence to a unique subject sequence. These sequences included 257 crAssphage
205 genomes, range in size between 92,182 and 100,327 (average 96,984.8 bases). Members of
206 *Bacteroidales* were predicted as targeting hosts of crAssphages in this study, findings which are
207 consistent with those hosts reported to propagate crAssphage in previous studies[21,22]. We also
208 found that 154 large TR sequences (6%) were listed in RefSeq plasmid, of which 7 were also listed
209 in the IMG/VR database. Based on our investigation, we determined that there is adequate
210 classification between plasmids and viruses for large genomes. Notably, we discovered 11 TR
211 sequences larger than 200 kb; 7 of which correspond to recently reported 'huge phage' genomes[14]
212 and 4 of these very large TR sequences were novel. Finally, we found that 681 large TR sequences
213 (26.7%) were novel. In summary, we conclude that many of the large TR sequences are already
214 represented in virus or plasmid databases, with the exception of those greater than 200 kb which
215 were only recently reported.

216 In contrast to the large TR sequences, most of the small TR sequences were not represented
217 in the databases searched (Fig. 3b). Among the 8,837 small TR sequences, 491 small TR sequences
218 (5.6%) were listed in IMG/VR, whereas 2,573 small TR sequences (29.1%) were listed in GVD.
219 Only 256 small TR sequences (2.9%) were listed in both IMG/VR and GVD. Finally, 5,518 small
220 TR sequences (62.4%) were novel, indicating that the majority of small viral or plasmid sequences
221 targeted by CRISPR are still unexplored, even in the intensively studied human gut metagenome.
222 However, IMG/VR filters out genomes shorter than 5 kb to minimize the rate of false positive
223 predictions[35], which likely explains the substantially lower represented sequences in IMG/VR for
224 small TR sequences. Moreover, among the 942 RefSeq plasmid-listed sequences, 444 small TR

225 sequences were listed in RefSeq virus, IMG/VR and/or GVD, suggesting a possible
226 misclassification between virus and plasmid within these databases. We investigated database
227 representation of CRISPR targeted TR sequences predicted to represent *Microviridae* and
228 *Inoviridae* species. Among the 766 putative *Microviridae* genomes from this study, we found that
229 639 genomes (83.4%) were represented in at least one virus database, and none were listed among
230 RefSeq plasmids. In contrast to *Microviridae*, among the 114 putative *Inoviridae* genomes from
231 this study, only 2 were listed in the virus databases and 11 were listed in the RefSeq plasmid
232 database. Next, we compared our putative *Inoviridae* genomes with recently reported *Inoviridae*
233 genomes discovered using a machine learning approach[36], and found that 43 genomes from our
234 study were highly similar to the genomes from the previous study, supporting our prediction that
235 these sequences are indeed *Inoviridae* genomes.

236 To determine whether our novel TR sequences are viruses, we searched sequences for
237 capsid genes using a high-sensitivity homology detection method. We built a capsid HMM
238 database using reference capsid proteins as baits which were then used to scan TR sequences. We
239 identified 1,582 small TR sequences (17.9%) encoded capsid genes (score > 30), suggesting that
240 at least this portion of TR sequences are viral genomes. However, capsid genes are unlikely to
241 share a single common ancestor[37], and it is unlikely that the reference sequences cover the entire
242 sequence diversity of capsids. In addition, some viruses do not encode capsids[38,39]. Therefore,
243 although this is a conservative estimate of the truly viral fraction of the taxonomically
244 unclassifiable sequences reported here, we are unable to strongly conclude whether the remaining
245 small TR sequences are viruses or not at this point.
246

247 **Gene content based hierarchical clustering of CRISPR-targeted TR sequences.** To scrutinize
248 CRISPR-targeted TR sequences in a genomic context, we hierarchically clustered TR sequences
249 based on gene content. For this analysis, we selected the top 1,000 genes recurrently observed
250 from large and small TR sequences. The clustering results for large TR sequences (Fig. 4a) showed
251 that the majority of genomes, with a variety of gene contents, have already been listed in virus
252 databases. This finding further supports that these databases contain at least representatives similar,
253 at broad taxonomic levels, to the large viruses present in human gut. Specific gene contents were
254 observed for crAssphages, which are exclusively targeted by *Bacteroidetes*. In addition, RefSeq
255 plasmid hit sequences formed an exclusive cluster with conjugation-related genes. As the
256 conjugation proteins included those involved in pili formation and intercellular DNA transfer
257 proteins, these sequences are likely F plasmids. TR sequences predicted to be targeted by
258 monoderm and diderm hosts clustered separately, indicating there is little gene flow between them.
259 Except for the likely F plasmid sequences, most of these sequences encode capsids. Combined
260 with the fact that most of the large TR sequences encoded tailed phage essential genes, this result
261 further supports the conclusion that the majority of the large TR sequences are tailed phages.
262 In contrast to the large TR sequences, we found that the small TR sequences remained
263 largely enigmatic when clustered by gene content. Few clusters were represented in IMG/VR, and
264 although GVD covers a relatively broader range, representatives are still missing or sparse for
265 some clusters. Several clusters were also listed in both plasmid and virus databases. As many
266 clusters did not encode detectable capsid genes, both clusters representing *Microviridae* evince
267 capsid genes. Another pattern shown by this analysis was that high-GC content TR sequences were
268 frequently observed with *Actinobacteria* as the predicted targeting host.
269

270 **Diversity-generating retroelements of crAssphage lineage.** Recently, one crAssphage genome
271 encoding a diversity-generating retroelement (DGR) was reported[40]. DGRs are retroelements
272 found in prokaryotes and viruses[41]. They minimally consist of an error-prone reverse transcriptase
273 and a non-coding region called a template repeat. The transcript of a template repeat is reverse
274 transcribed, and this complementary DNA (cDNA) product includes mutations from adenine to a
275 random nucleotide (A-to-N). The cDNA is incorporated to a specific position of the target gene
276 called the variable repeat; thus, this element effectively alters the coding sequence of the target
277 gene which may benefit from rapid adaptation, such as expansion to new hosts and
278 environments[42,43]. Among the 257 putative crAssphage genomes identified in our study, 98
279 genomes encoded DGRs including a reverse transcriptase gene with an associated non-coding
280 region recognizable as a putative template repeat showing A-to-N (adenine-to-any base) mutations
281 to the tail-collar coding gene. Using these abundant putative crAssphage sequences, we further
282 investigated this locus to uncover the origin of this retroelement. The position of the DGR locus
283 was consistent among DGR-encoding crAssphages (Supplementary Fig. 6), suggesting that DGRs
284 in crAssphages observed so far are orthologous and likely originate from a common ancestral DGR
285 insertion. The reverse transcriptase gene and template repeat within this module showed high
286 sequence conservation, whereas the variable repeat within the adjacent gene that encodes tail-
287 collar fibre showed high diversity, indicating that this DGR was recently generating diversity (Fig.
288 5a). Notably, however, the tail-collar fibre of putative crAssphage sequences without DGR were
289 similarly diverse. We used the protein sequence of DGR reverse transcriptase to query non-human-
290 derived crAssphage-like genomes[44], and only one crAssphage-like genome, which was derived by
291 sequencing baboon faeces, encoded this element (Fig. 5a). Pair-wise alignment of template and
292 variable repeats within the baboon crAssphage-like genome showed A-to-N substitutions that are

293 likely facilitated by the DGR (Fig. 5b), which indicates the DGR in the baboon crAssphage-like
294 genome is also functional. We then constructed a Bayesian phylogeny of crAssphage genomes
295 (Fig. 5c). We found that the distribution of reverse transcriptase-encoding genomes (shown in red
296 text in Fig. 5c) did not form a cluster, a finding which is most consistent with multiple losses of
297 this retroelement, and potentially explaining the high diversity of the tail collar fibre even in
298 assembled sequences that do not currently contain a DGR. A deeper-rooted molecular phylogeny
299 with distant crAssphages-like genomes would be required to uncover more details about the origin
300 of this retroelement.
301
302 **Remnant CRISPR spacers and overall contribution of CRISPR targeted sequences to**
303 **identified spacers.** Viruses and other mobile genetic elements can escape from CRISPR targeting
304 by acquiring mutations to protospacer loci[45,46]. While the corresponding spacers are thus no longer
305 effective, they can remain in the host genome. To investigate these potential 'remnant' CRISPR
306 spacers, we mapped all unique CRISPR spacers to TR sequences as well as scrambled sequences
307 with various sequence identity thresholds (Fig. 6). The scrambled sequences were used to monitor
308 false positive matches arising by chance (see Methods). Based on the observation of incremental
309 false positive matches of spacers to scrambled sequences, 84% is a sequence identity threshold at
310 which some putative remnant spacers can be mapped, but very few false-positive matches. At an
311 84% sequence identity threshold, 269,808 spacers and 126,616 spacers were mapped to large and
312 small TR sequences, respectively. Taken together, 20.13% of all unique spacers (396,424 spacers)
313 were mapped to TR sequences. Compared to the stringent identity threshold initially applied (93%),
314 91.91% more spacers were mapped to small TR sequences with relaxed threshold. By comparison,
315 only 42.68% more spacers were mapped to large TR sequences with relaxed threshold. These

316 results suggest that a substantial fraction of CRISPR spacers imperfectly matches to protospacers
317 within circular mobile genetic elements in human gut. This potentially reflects an 'escape mutation'
318 phenomenon. By percentage, small TR sequences can be explained in this manner than large TR
319 sequences.

320 While we focused TR sequences in this study because of high confidence of the genomic
321 completeness, spacers could be derived from incompletely reconstructed or non-circular genomes.
322 To estimate the contribution of discovered CRISPR targeted sequences to identified CRISPR
323 spacers, all unique spacers were mapped to all representative CRISPR targeted sequences with
324 84% sequence identity threshold. Under these conditions, 971,224 spacers (49.31% of all unique
325 spacers) were mapped.

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

330

331 **Conclusion**

332 In the current study, we demonstrated that CRISPR spacers can be used as a tool to detect viral
333 genomes and other mobile genetic elements from metagenome sequences. Using spacers to
334 confidently infer which sequences are targeted by CRISPR, we substantially expanded the
335 diversity of mobile genetic elements identifiable from the human gut metagenome, which has
336 already been a topic of intense investigation for virus discovery. Comparing the sequences
337 predicted by this approach against viral databases showed that our method effectively detects viral
338 genomes without requiring similarity to any known viral sequence. Although the majority of large

339 (>20 kb) genomes were predicted as tailed phages with high confidence based on sequence
340 homology, we found that the majority of small (<20 kb) genomes remain unclassified because of
341 a lack of similar genomes in annotated databases. We showed that the source of nearly half of the
342 CRISPR spacers encoded by residents of the human gut remains unknown, suggesting that
343 additional protospacer reservoirs, whether extinct or simply unsampled, are still to be characterized.
344 Applying this conceptual advance to additional metagenomic datasets will increase the breadth of
345 the lens through which we can study the diversity of Earth's virome.
346

347 **Materials and Methods**

348 **Materials.** Sequencing data were selected based on NCBI metadata. Filtering parameters for the
349 query were as follows: layout = PAIRED, platform = ILLUMINA, selection = RANDOM, strategy
350 = WGS, source = METAGENOME, NCBI Taxonomy = 408170 (human gut metagenome) and
351 minimum library size = 1 Gb. If a sample contained multiple runs, we selected the run with the
352 most bases for simplicity of the analysis pipeline and to avoid possible bias to protospacer counts
353 from nearly identical metagenomes.
354

355 **Database versions and download dates.** RefSeq Release 98 was downloaded on 10 January 2020
356 and IMG/VR Release Jan. 2018 was downloaded on 21 October 2019. GVD was downloaded on
357 11 March 2020 and UniRef50 was downloaded on 16 December 2019.
358

359 **Metagenome assembly.** All downloaded paired FASTQ files were pre-processed based on
360 guidance in BBTools[47] (version 38.73). Adapters, phi X and human sequences were removed using
361 BBDuk and BBMap. Sequencing errors were corrected using Tadpole. Each pre-processed pair of

362 FASTQ files was assembled using SPAdes[48] (version 3.12) with -meta option. Contigs smaller
363 than 1 kb were discarded.
364
365 **Detection of CRISPR and spacer extraction.** Assembled contigs were scanned with
366 CRISPRDetect[49] (version 2.2) to extract CRISPR DRs which were deduplicated using CD-HIT-
367 EST[50] (version 4.7) and used to mask the raw reads. We extracted CRISPR spacers from the raw
368 reads to maximize spacer capture from the library. Sequences between the masked regions within
369 the raw reads were considered CRISPR spacers which were extracted by a simple Python program
370 (available in our source code repository) and then deduplicated.
371
372 **Detection of protospacer loci.** All DRs were mapped to contigs using BBMap with a 93%
373 minimum sequence identity. DR mapped positions and their flanking 60 bases were masked as
374 CRISPR loci. Next, identified spacers were mapped to all CRISPR masked contigs with a 93%
375 minimum sequence identity. Rather than excluding all contigs with CRISPR loci, we exclusively
376 masked CRISPR loci to identify viruses that encode CRISPR systems which are themselves
377 targeted by other CRISPR systems. To increase specificity, we aligned the 5' and 3' adjacent
378 regions of spacer-mapped positions. These adjacent sequences were also aligned to the DR
379 sequence associated with the mapped spacer. We discarded those loci in which any alignment
380 score divided by length was higher than 0.5, using the alignment parameters: match = 1, mismatch
381 = -1, gap = -1 and gap extension = -1. The remaining positions were considered authentic
382 protospacer loci.
383

384 **Co-occurrence-based spacer clustering.** We clustered spacers in two steps (Supplementary Fig.
385 2). First, we clustered protospacer loci located within 50 kb of another protospacer locus. We then
386 further clustered spacers based on the co-occurrence of protospacers represented as a graph. In this
387 graph, protospacers are nodes and the edges represent co-occurrence of connected protospacers.
388 The weights of edges were observed counts of co-occurrence of the connected protospacers
389 defined in the previous clustering. Graph communities were detected using Markov Clustering
390 algorithm[51] (options: -I 4 -pi 0.4; version 14-137). Clusters with a size lower than 10 and a global
391 clustering coefficient lower than 0.5 were discarded. Finally, 12,749 clusters comprising 591,189
392 spacers were derived.
393

394 **Extraction of CRISPR-targeted sequences.** Contiguous regions of contigs targeted by more than
395 10% of members of a spacer cluster were marked as a bed file using BEDTools[52]. To join the
396 fragmented clusters, adjacent regions within 1 kb were concatenated. Marked regions were
397 extracted, and sequences containing assembly gap were discarded. Finally, both ends of each
398 extracted sequence were compared to identify TR sequences using a python program utilizing the
399 Biopython[53] package (available in our source code repository).
400

401 **Deduplication of CRISPR-targeted sequences.** TR sequences were clustered using PSI-CD-HIT
402 (options: -c 0.95 -aS 0.95 -aL 0.95 -G 1 -g 1 -prog blastn -circle 1). The remaining CRISPR
403 targeted sequences were clustered twice using linclust[54] (options: --cluster-mode 2 --cov-mode 1
404 -c 0.9 --min-seq-id 0.95), then clustered again using PSI-CD-HIT (options: -c 0.9 -aS 0.95 -G 1 -
405 g 1 -prog blastn -circle 1).
406

407 **Gene prediction and annotation of CRISPR-targeted sequence.** Protein-coding genes were
408 predicted from both TR and non-TR sequences using Prodigal (version 2.6.3) with the -p meta
409 option. To recover truncated genes from TR sequences, each TR sequence was concatenated in
410 silico and unique predicted genes were selected. Predicted protein sequences with partial frags
411 were discarded. The remaining protein sequences were clustered based on a 30% sequence identity
412 threshold and aligned using mmseqs[55] (version e1a1c1226ef22ac3d0da8e8f71adb8fd2388a249).
413 HMMs were constructed from each aligned sequence using hmmbuild from HMMER[56] (version
414 3.2.1). Constructed HMMs were then used as query to search against UniRef50 and RefSeq viral
415 proteins (E value < 10e-10) using hmmsearch.
416

417 **Capsid HMM database construction.** We selected capsids and coat proteins from RefSeq viral
418 protein sequences. Sequences containing 'capsid' and 'coat' in the description were selected and
419 then manually inspected to remove sequences unlikely to actually be capsid, such as containing
420 the term 'associate' or phrase 'assembly chaperone'. A total of 5,069 sequences were clustered
421 based on a 30% sequence identity, resulting in 1,324 representative sequences. Each representative
422 sequence was used as bait and underwent five jackhmmer iterations on UniRef50 database. HMMs
423 from the final iteration were used to build the capsid HMM database.
424

425 **Targeting host prediction.** DRs were mapped to RefSeq bacteria and archaea genomes using
426 BBMap. A locus with more than three consecutive DR hits within 100 bases was considered an
427 authentic CRISPR locus associated with the mapped DR. DRs mapped to multiple taxa in a given
428 taxonomic level were not taxonomically assigned for that level. DRs assigned to taxa were used
429 to predict the targeting host. We counted the protospacers linked to taxonomically assigned DRs

430 within a TR sequence. If the count of a given taxon was ≥10 and higher than 90% exclusiveness,
431 we considered that the corresponding taxon is a targeting host of a given contig. Host prediction
432 was done for each taxonomic level: species, genus, family, order, class, phylum and domain.
433

434 **Gene content-based hierarchical clustering of TR sequences.** TR sequences were scanned by
435 HMMs derived from the clustering results of the predicted protein sequences using both TR and
436 non-TR sequences. The scanned result was represented by a binary matrix (score > 60). We
437 selected the top 1,000 genes recurrently observed within TR sequences. The matrix was
438 hierarchically clustered with ComplexHeatmap[57] (version 2.5.3), and then annotated with database
439 hits, host, GC content, capsids and predicted gene functions. Clustering was separately performed
440 for large and small TR sequences.
441

442 **Phylogenetic analysis of *Microviridae* major capsid proteins.** Representative *Microviridae*
443 major capsid protein sequences were selected by clustering all capsids based on a 85% sequence
444 identity threshold throughout the entire length of the protein. Representative and reference protein
445 sequences were aligned using MAFFT[58] (version 7.310) and then trimmed using trimAI[59] (version
446 1.4). Aligned sequences were used for Bayesian phylogenetic analysis using MrBayes[60] (version
447 3.2.7). A mixed substitution model with a uniform prior that converged to Blosum62 (posterior
448 probability = 1.000) was selected. All other priors were set to default. Two chains of Markov chain
449 Monte Carlo (MCMC) with identical priors run over ten million generations and sampled every
450 500 generations. The standard deviation of split frequencies approached close to zero (0.007837)
451 over the run.
452

453 **Phylogenetic analysis of crAssphage core genes.** 55 core crAssphage genes which were encoded
454 in all crAssphage genomes from our study and the reference crAssphage genome were selected.
455 Representative crAssphage genomes were seelcted by clustering RNA polymerase coding
456 sequences based on a 97.5% sequence identity threshold. Each gene was aligned using MAFFT
457 and then trimmed using trimAI. Aligned core genes were then concatenated and used for Bayesian
458 phylogenetic analysis. A GTR + I + Γ substitution model was selected. Two chains of MCMC with
459 identical priors were run over ten million generations and sampled every 500 generations. The
460 standard deviation of split frequencies approached close to zero (0.000499) over the run.
461
462 **Generation of scrambled sequences.** Scrambled sequences are random sequences with identical
463 lengths to the TR sequences. The sequences are generated based on sampled nucleotide
464 frequencies from the TR sequences using Biopython package (available in our source code
465 repository).
466
467 **Code availability.** The source codes used in this study are available at https://github.com/ryota-
468 sugimoto/virome_scripts
469

470 **References**

471 1. Wommack, K. E. & Colwell, R. R. Virioplankton: viruses in aquatic ecosystems.
472 Microbiol Mol Biol Rev 64, 69-114 (2000).

473 2. HERSHEY, A. D. & CHASE, M. Independent functions of viral protein and nucleic acid
474 in growth of bacteriophage. J Gen Physiol 36, 39-56 (1952).

475 3. Fiers, W. *et al*. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and
476 secondary structure of the replicase gene. Nature 260, 500-507 (1976).

477 4. Sanger, F. *et al*. Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265, 687-
478 695 (1977).

479 5. Koonin, E. V., Senkevich, T. G. & Dolja, V. V. The ancient Virus World and evolution
480 of cells. Biol Direct 1, 29 (2006).

481 6. Krupovic, M., Dolja, V. V. & Koonin, E. V. Origin of viruses: primordial replicators
482 recruiting capsids from hosts. Nature Reviews Microbiology 17, 449-458 (2019).

483 7. Minot, S. *et al*. The human gut virome: inter-individual variation and dynamic response
484 to diet. Genome research 21, 1616-1625 (2011).

485 8. Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. Molecular
486 biological access to the chemistry of unknown soil microbes: a new frontier for natural products.
487 Chem Biol 5, R245-9 (1998).

488 9. Roux, S., Enault, F., Hurwitz, B. L. & Sullivan, M. B. VirSorter: mining viral signal from
489 microbial genomic data. PeerJ 3, e985 (2015).

490 10. Ren, J., Ahlgren, N. A., Lu, Y. Y., Fuhrman, J. A. & Sun, F. VirFinder: a novel k-mer
491 based tool for identifying viral sequences from assembled metagenomic data. Microbiome 5, 69
492 (2017).

493 11. Gregory, A. C., Zablocki, O., Howell, A., Bolduc, B. & Sullivan, M. B. The human gut
494 virome database. (2019).

495 12. Barrangou, R. *et al*. CRISPR provides acquired resistance against viruses in prokaryotes.
496 Science 315, 1709-1712 (2007).

497 13. Seed, K. D., Lazinski, D. W., Calderwood, S. B. & Camilli, A. A bacteriophage encodes
498 its own CRISPR/Cas adaptive response to evade host innate immunity. Nature 494, 489-491
499 (2013).

500 14. Al-Shayeb, B. *et al*. Clades of huge phages from across Earth's ecosystems. Nature 578,
501 425-431 (2020).

502 15. Snyder, J. C., Bateson, M. M., Lavin, M. & Young, M. J. Use of Cellular CRISPR
503 (Clusters of Regularly Interspaced Short Palindromic Repeats) Spacer-Based Microarrays for
504 Detection of Viruses in Environmental Samples. Applied and Environmental Microbiology 76,
505 7251-7258 (2010).

506 16. Zhang, Q., Rho, M., Tang, H., Doak, T. G. & Ye, Y. CRISPR-Cas systems target a
507 diverse collection of invasive mobile genetic elements in human microbiomes. Genome Biol 14,
508 R40 (2013).

509 17. Shmakov, S. A. *et al*. The CRISPR spacer space is dominated by sequences from species-
510 specific mobilomes. MBio 8, e01397-17 (2017).

511 18. Shmakov, S. A., Wolf, Y. I., Savitskaya, E., Severinov, K. V. & Koonin, E. V. Mapping
512 CRISPR spaceromes reveals vast host-specific viromes of prokaryotes. Commun Biol 3, 321
513 (2020).

514 19. Paez-Espino, D. *et al*. Uncovering Earth's virome. Nature 536, 425-430 (2016).

515 20. Stern, A., Mick, E., Tirosh, I., Sagy, O. & Sorek, R. CRISPR targeting reveals a reservoir
516 of common phages associated with the human gut microbiome. Genome Res 22, 1985-1994
517 (2012).

518 21. Dutilh, B. E. *et al*. A highly abundant bacteriophage discovered in the unknown
519 sequences of human faecal metagenomes. Nat Commun 5, 4498 (2014).

520 22. Yutin, N. *et al*. Discovery of an expansive bacteriophage family that includes the most
521 abundant viruses from the human gut. Nat Microbiol 3, 38-46 (2018).

522 23. STREISINGER, G., EDGAR, R. S. & DENHARDT, G. H. CHROMOSOME
523 STRUCTURE IN PHAGE T4. I. CIRCULARITY OF THE LINKAGE MAP. Proc Natl Acad
524 Sci U S A 51, 775-779 (1964).

525 24. Cui, J., Schlub, T. E. & Holmes, E. C. An allometric relationship between the genome
526 length and virion volume of viruses. J Virol 88, 6403-6410 (2014).

527 25. Hua, J. *et al*. Capsids and Genomes of Jumbo-Sized Bacteriophages Reveal the
528 Evolutionary Reach of the HK97 Fold. mBio 8, (2017).

529 26. Roux, S., Krupovic, M., Poulet, A., Debroas, D. & Enault, F. Evolution and diversity of
530 the Microviridae viral family through a collection of 81 new complete genomes assembled from
531 virome reads. PloS one 7, e40418 (2012).

532 27. Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K. & Knight, R. Diversity,
533 stability and resilience of the human gut microbiota. Nature 489, 220-230 (2012).

534 28. Almpanis, A., Swain, M., Gatherer, D. & McEwan, N. Correlation between bacterial
535 G+C content, genome size and the G+C content of associated plasmids and bacteriophages.
536 Microb Genom 4, (2018).

537 29. Brentlinger, K. L. *et al*. Microviridae, a family divided: isolation, characterization, and
538 genome sequence of φMH2K, a bacteriophage of the obligate intracellular parasitic bacterium
539 Bdellovibrio bacteriovorus. Journal of bacteriology 184, 1089-1094 (2002).

540 30. Sharp, P. M. & Hahn, B. H. Origins of HIV and the AIDS pandemic. Cold Spring Harb
541 Perspect Med 1, a006841 (2011).

542 31. Zhou, P. *et al*. A pneumonia outbreak associated with a new coronavirus of probable bat
543 origin. nature 579, 270-273 (2020).

544 32. Watson, B. N. J. *et al*. Type IF CRISPR-Cas resistance against virulent phages results in
545 abortive infection and provides population-level immunity. Nature communications 10, 1-8
546 (2019).

547 33. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq): a
548 curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids
549 Res 35, D61-5 (2007).

550 34. Paez-Espino, D. *et al*. IMG/VR: a database of cultured and uncultured DNA Viruses and
551 retroviruses. Nucleic Acids Res 45, D457-D465 (2017).

552 35. Paez-Espino, D. *et al*. IMG/VR v.2.0: an integrated data management and analysis
553 system for cultivated and environmental viral genomes. Nucleic Acids Res 47, D678-D686
554 (2019).

555 36. Roux, S. *et al*. Cryptic inoviruses revealed as pervasive in bacteria and archaea across
556 Earth's biomes. Nat Microbiol 4, 1895-1906 (2019).

557 37. Krupovic, M. & Koonin, E. V. Multiple origins of viral capsid proteins from cellular
558 ancestors. Proc Natl Acad Sci U S A 114, E2401-E2410 (2017).

559 38. Lindsay, J. A., Ruzin, A., Ross, H. F., Kurepina, N. & Novick, R. P. The gene for toxic
560 shock toxin is carried by a family of mobile pathogenicity islands in Staphylococcus aureus.
561 Molecular microbiology 29, 527-543 (1998).

562 39. Krupovic, M., Ravantti, J. J. & Bamford, D. H. Geminiviruses: a tale of a plasmid
563 becoming a virus. BMC Evol Biol 9, 112 (2009).

564 40. Morozova, V. *et al*. First crAss-Like Phage Genome Encoding the Diversity-Generating
565 Retroelement (DGR). Viruses 12, (2020).

566 41. Paul, B. G. *et al*. Retroelement-guided protein diversification abounds in vast lineages of
567 Bacteria and Archaea. Nat Microbiol 2, 17045 (2017).

568 42. Liu, M. *et al*. Reverse transcriptase-mediated tropism switching in Bordetella
569 bacteriophage. Science 295, 2091-2094 (2002).

570 43. Guo, H., Arambula, D., Ghosh, P. & Miller, J. F. Diversity-generating Retroelements in
571 Phage and Bacterial Genomes. Microbiol Spectr 2, (2014).

572 44. Edwards, R. A. *et al*. Global phylogeography and ancient evolution of the widespread
573 human gut virus crAssphage. Nat Microbiol 4, 1727-1736 (2019).

574 45. Deveau, H. *et al*. Phage response to CRISPR-encoded resistance in Streptococcus
575 thermophilus. Journal of bacteriology 190, 1390-1400 (2008).

576 46. Fineran, P. C. *et al*. Degenerate target sites mediate rapid primed CRISPR adaptation.
577 Proc Natl Acad Sci U S A 111, E1629-38 (2014).

578 47. Bushnell, B. BBTools software package. URL http://sourceforge. net/projects/bbmap
579 (2014).

580 48. Bankevich, A. *et al*. SPAdes: A New Genome Assembly Algorithm and Its Applications
581 to Single-Cell Sequencing. Journal of Computational Biology 19, 455-477 (2012).

582 49. Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C. & Brown, C. M.
583 CRISPRDetect: a flexible algorithm to define CRISPR arrays. BMC genomics 17, 356 (2016).

584 50. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of
585 protein or nucleotide sequences. Bioinformatics 22, 1658-1659 (2006).

586 51. Van Dongen, S. M. Graph clustering by flow simulation. (2000).

587 52. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing
588 genomic features. Bioinformatics 26, 841-842 (2010).

589 53. Cock, P. J. A. *et al*. Biopython: freely available Python tools for computational molecular
590 biology and bioinformatics. Bioinformatics 25, 1422-1423 (2009).

591 54. Steinegger, M. & Söding, J. Linclust: clustering billions of protein sequences per day on
592 a single server. bioRxiv 104034 (2017).

593 55. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for
594 the analysis of massive data sets. Nature biotechnology 35, 1026-1028 (2017).

595 56. Eddy, S. R. & team, H. M. M. E. R. D. HMMER User's Guide 2019).

596 57. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in
597 multidimensional genomic data. Bioinformatics 32, 2847-2849 (2016).

598 58. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid
599 multiple sequence alignment based on fast Fourier transform. Nucleic acids research 30, 3059-
600 3066 (2002).

601 59. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated
602 alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972-1973 (2009).

603 60. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees.
604 Bioinformatics 17, 754-755 (2001).

605

Acknowledgements

The major part of analysis has been done on the supercomputer system in National Institute of Genetics. R.S. is supported by funds of Database Center for Life Science, Research Organization of Information and Systems, and KAKENHI.

Author contributions

R.S. designed the studies and performed most of the analyses. L.N. assisted in contig construction and alignment. 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.

620 **Figure legends**

621 **Fig. 1 | Properties of CRISPR-targeted TR sequences. a**, Length distribution of TR sequences.
622 We used portal, tail and terminase large subunit as marker genes of tailed phage; sequences
623 encoding all three of these genes were categorized as tailed phages. The dotted line at 20 kb
624 represents an arbitrary cut-off between small and large sequences. Sequences longer than 100 kb
625 are shown in the inset. **b**, Predicted CRISPR targeting host composition of TR sequences. Hosts
626 were predicted by mapping CRISPR DR sequences to the RefSeq database. Sequences containing
627 ≥10 protospacer loci but less than 90% associated DR taxa exclusiveness were considered targeted
628 by multiple phyla. If ≥10 protospacers could not be assigned to a taxon, the predicted targeting
629 host is denoted as not available (NA). **c**, Predicted targeting host distribution by GC content. The
630 dotted line indicates the low- and high-GC content boundary at 55%. **d**, Heterogeneous distribution
631 of TR sequences targeted by multiple phyla. Circle size approximately represents popularity of the
632 respective host. Bidirectional arrows connect the top two host phyla according to host-assigned
633 protospacer counts (i.e. protospacers most often associated with CRISPR DRs are assigned to these
634 two phyla). Numbers on arrows are counts of the number of TR sequences targeted by the
635 connected phyla.

636

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

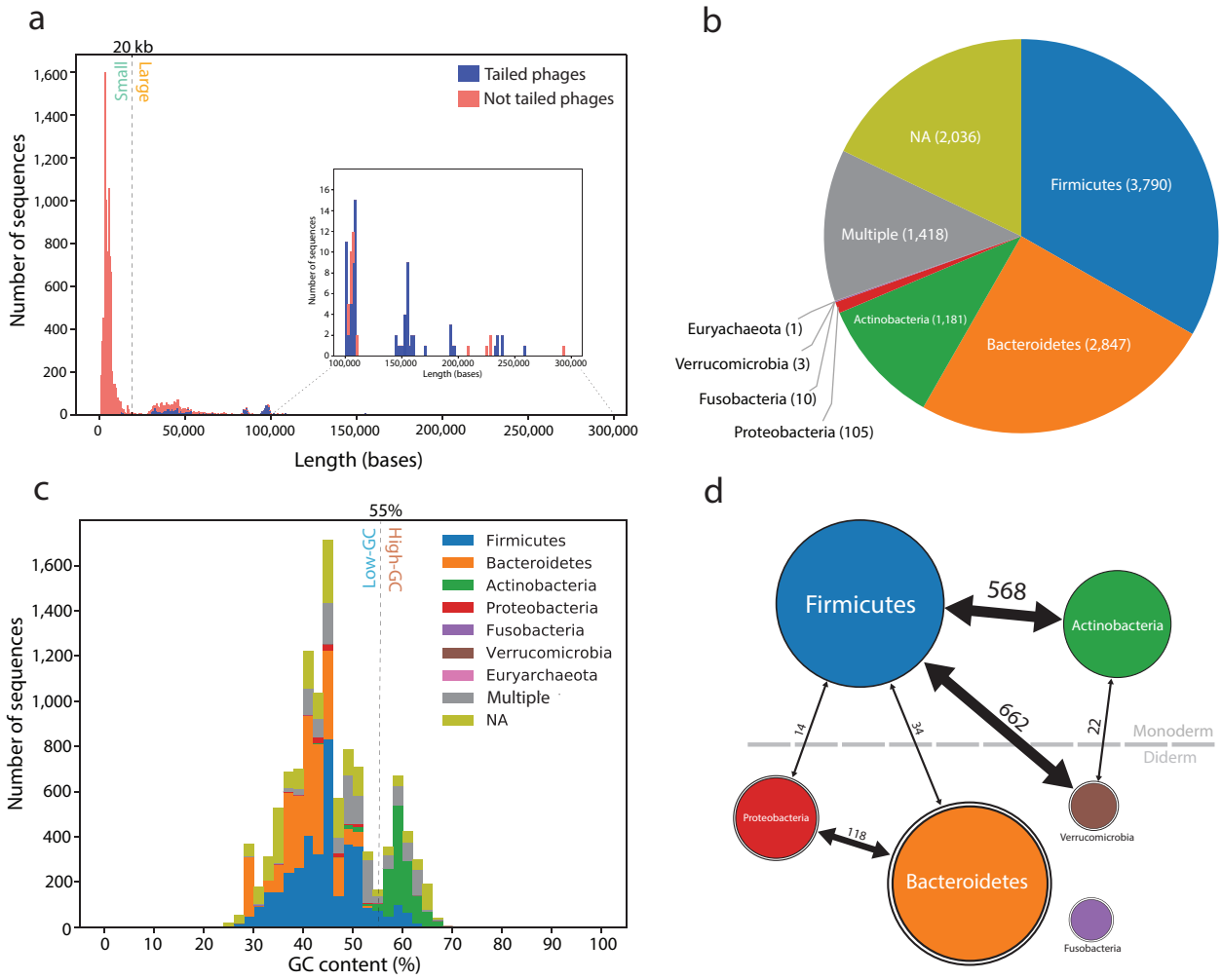
642

643 **Fig. 3 | Venn diagrams of database comparisons for (a) large and (b) small TR sequences.**
644 Each TR sequence was compared to RefSeq virus, RefSeq plasmid, IMG/VR and GVD using
645 BLASTN. Database hit minimum criteria was set to 85% sequence identity with 75% aligned
646 fraction of query sequence to a unique subject sequence.
647
648 **Fig. 4 | Hierarchical clustering of (a) large and (b) small TR sequences based on gene content.**
649 Heatmaps representing gene content of TR sequences, in which each row is a TR sequence and
650 each column is a gene cluster. Gray areas in the heatmap indicate sequences encoding a gene that
651 is homologous to the gene cluster. Note that one gene can be homologous to multiple gene clusters.
652 Sequences are annotated by database containing similar sequences, GC content, host and
653 detectable capsid genes. Gene clusters were annotated by searching corresponding HMMs to the
654 UniRef50 database. Several notable RefSeq-listed clusters are denoted on the right side of the
655 heatmaps.
656
657 **Fig. 5 | DGR locus and phylogeny of crAssphage core genes. a**, Comparison of the DGR locus
658 between a reference crAssphage genome (NC_024711.1), a representative from this study
659 (crAssphage.202.rt) and a baboon-derived crAssphages (Baboon19.1). The retroelement-insertion
660 region is enclosed by a dotted line. Sequence conservation of DGR and its target gene (tail collar
661 fibre) were calculated by aligning and building HMMs. Sequence conservation of the DGR target
662 gene was calculated separately based on whether the genome encodes DGR (red) or not (blue).
663 Information content was averaged over seven base-window sizes. **b**, Pairwise sequence alignment
664 of template (top) and variable (bottom) repeats from Baboon19.1. **c**, Unrooted Bayesian phylogeny

665 of crAssphage core genes. 55 core genes from 20 representative crAssphages genomes and one
666 reference genome were used for analysis. Those crAssphages with DGR appear in red.
667

668 **Fig. 6 | Number of mapped spacers according to various sequence identity thresholds.** All
669 unique CRISPR spacers were mapped to large TR sequences, small TR sequences and scrambled
670 sequences. The initially applied and relaxed sequence identity thresholds are denoted as red and
671 orange colored dashed lines respectively. The spacer mapping process is identical to the
672 protospacer discovery process (see Methods).
673

674 **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
55%
Low-GC
High-GC
Firmicutes
Bacteroidetes
Actinobacteria
Proteobacteria
Fusobacteria
Verrucomicrobia
Euryarchaeota
Multiple
NA
Number of sequences
GC content (%)
d
Firmicutes
568
Actinobacteria
662
14
34
22
Monoderm
Diderm
Proteobacteria
118
Bacteroidetes
Verrucomicrobia
Fusobacteria

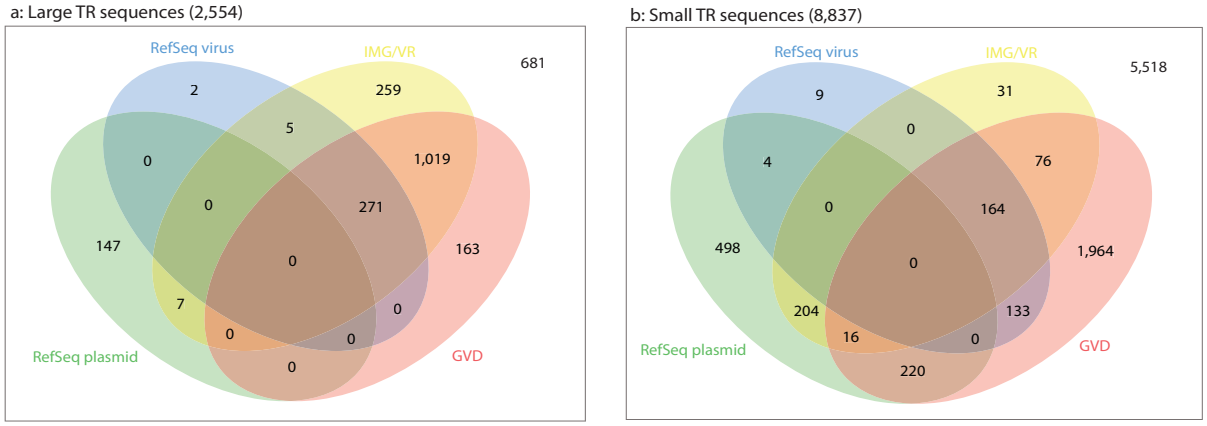
675

676 **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 Fen6723 1
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 31
Microviridae Bog5735 51
Microviridae Fen7184 21
Microviridae Fen418 41
Microviridae Fen2069 11
Microviridae Bog9917 22
Microviridae Fen7918 31
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 phageID21
Escherichia virus phiX174
Escherichia phage St-1
Eel River basin pequenovirus
0.3

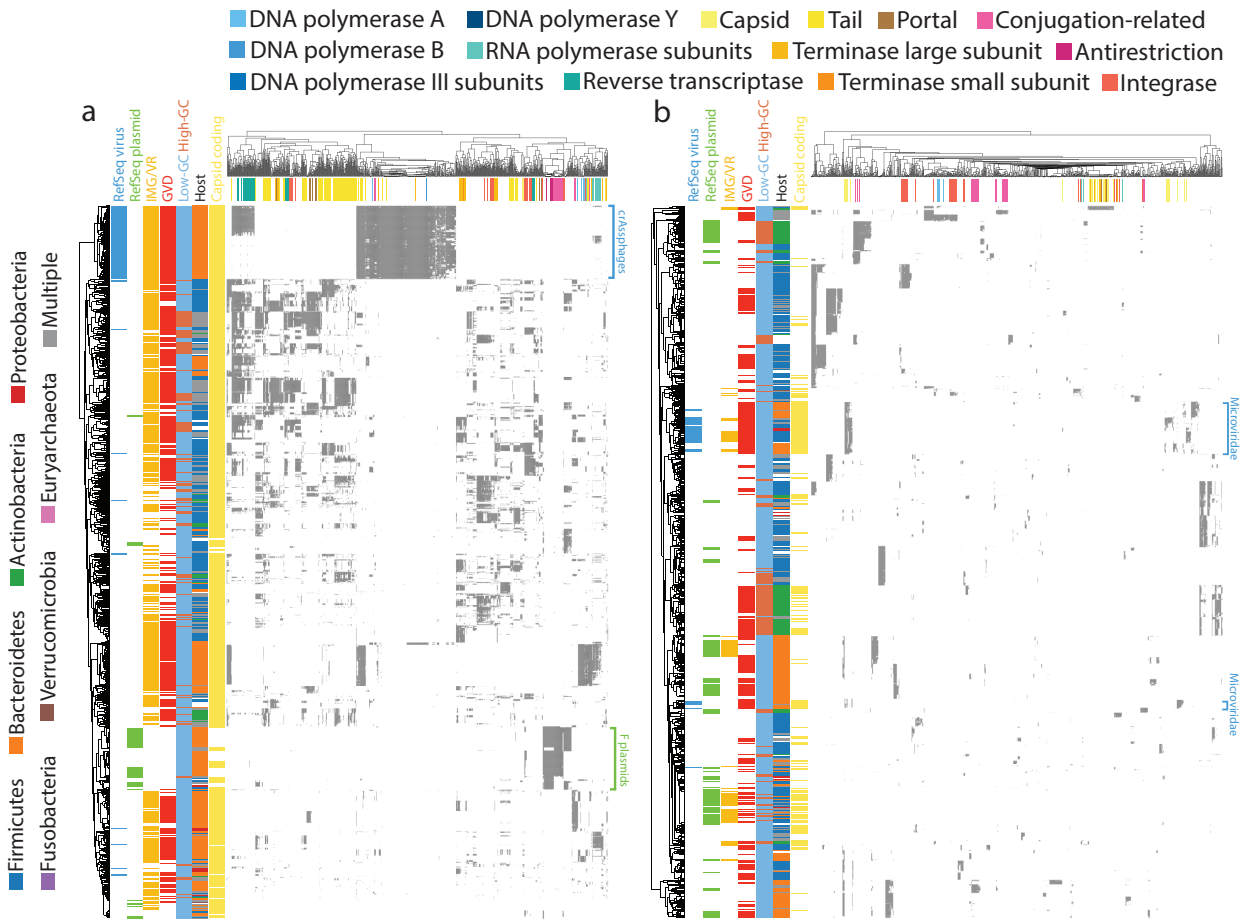
677

678 **Fig. 3**


a: Large TR sequences (2,554)
RefSeq virus
IMG/VR
681
2
259
5
0
1,019
0
271
147
163
0
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
1,964
0
204
133
16
0
220
RefSeq plasmid
GVD

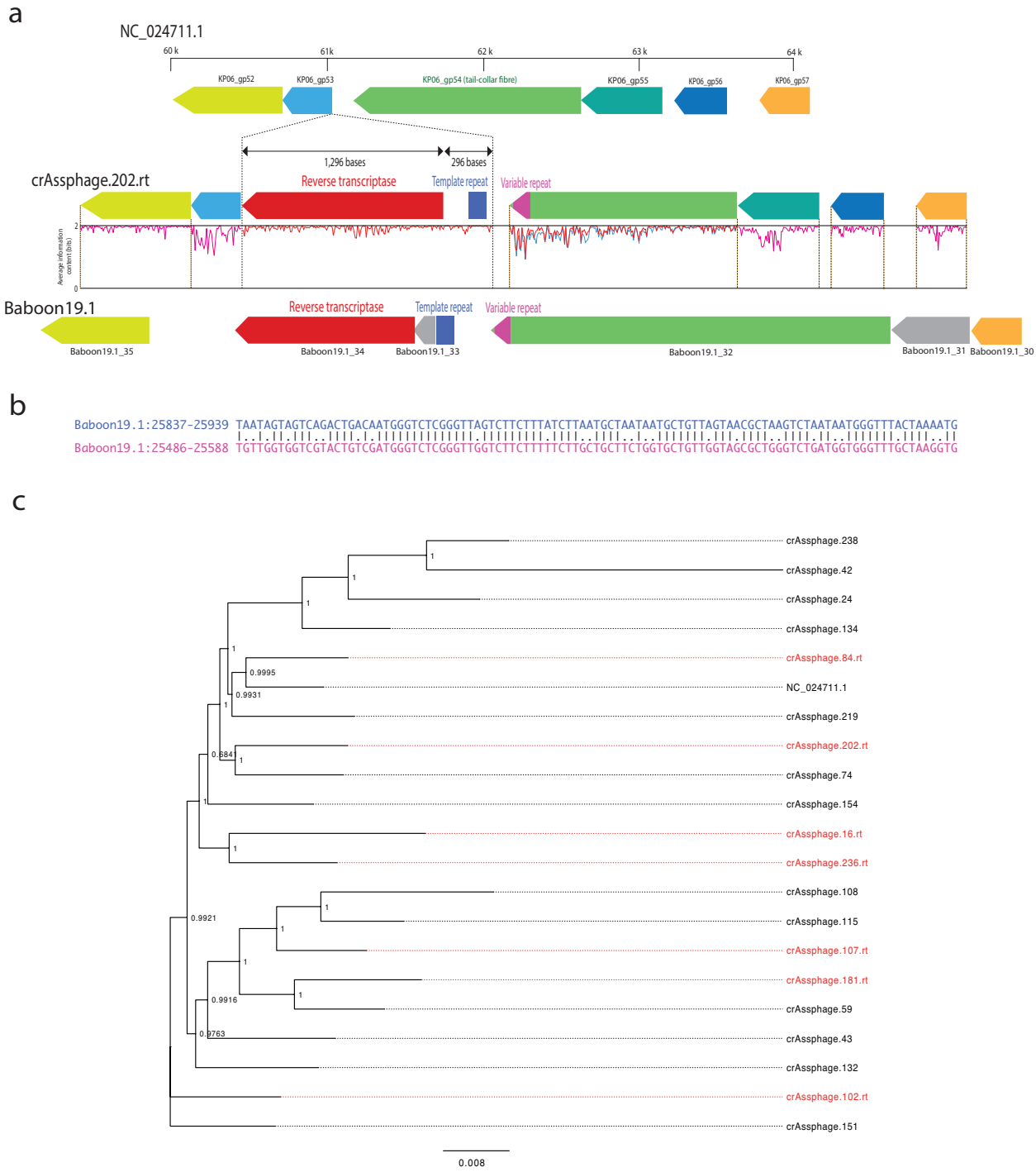
679

680 **Fig. 4**


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

681

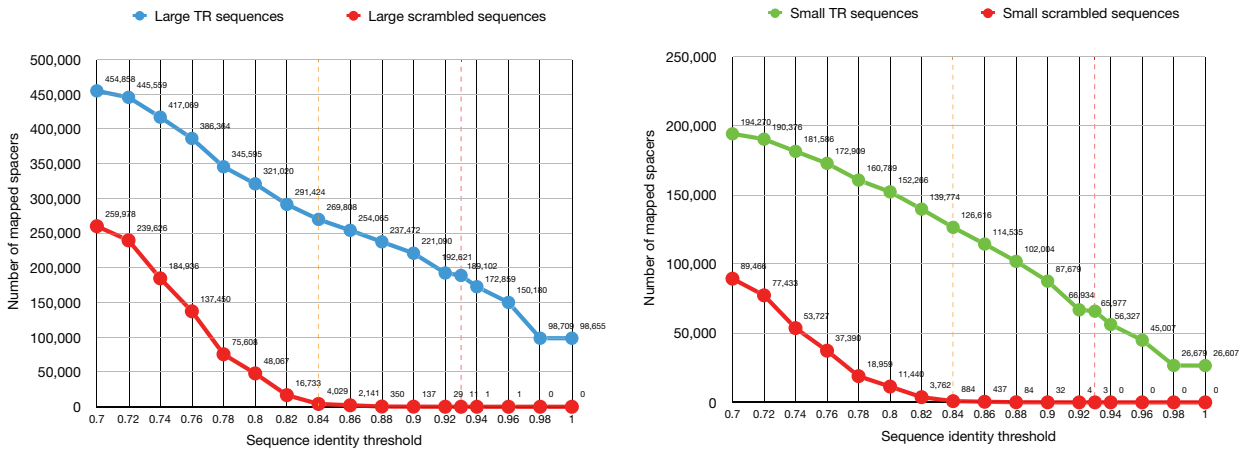
682 **Fig. 5**



683

684

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