Dynamics of CRISPR-mediated virus-host interactions in the human gut microbiome

Adrián López-Beltrán[1], João Botelho[1], Jaime Iranzo[1,2,3]

[1] Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid (UPM) - Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria (INIA-CSIC), Madrid, Spain

[2] Centro de Astrobiología (CAB), CSIC-INTA, Madrid, Spain

[3] Institute for Biocomputation and Physics of Complex Systems (BIFI), University of Zaragoza, Zaragoza, Spain

Competing interests: The authors declare no competing financial interests.

Abstract

Arms races between mobile genetic elements and prokaryotic hosts are major drivers of ecological and evolutionary change in microbial communities. Prokaryotic defense systems such as CRISPR-Cas have the potential to regulate microbiome composition by modifying the interactions among bacteria, plasmids, and phages. Here, we used longitudinal metagenomic data from 130 healthy and diseased individuals to study how the interplay of genetic parasites and CRISPR-Cas immunity reflects on the dynamics and composition of the human gut microbiome. Based on the coordinated study of 80,000 CRISPR-Cas loci and their targets, we show that CRISPR-Cas immunity effectively modulates bacteriophage abundances in the gut. Acquisition of CRISPR-Cas immunity typically leads to a decrease in the abundance of lytic phages, but does not necessarily cause their complete disappearance. Much smaller effects are observed for lysogenic phages and plasmids. Conversely, phage-CRISPR interactions shape

bacterial microdiversity by producing weak selective sweeps that benefit immune host lineages. Interestingly, distal (and chronologically older) regions of CRISPR arrays are enriched in spacers that are potentially functional and target crass-like phages and local prophages. This suggests that exposure to reactivated prophages and other endemic viruses is a major selective pressure in the gut microbiome that drives the maintenance of long-lasting immune memory.

Keywords: microbial community, virome, metagenomics, CRISPR, gut microbiome, virus-bacteria interaction, lysogeny, microbial defense system.

36 **Introduction**

37 Virus-host interactions are major determinants of community dynamics, composition, and
38 evolution in microbial ecosystems (1-5). In two classic (and somewhat antagonistic) examples,
39 phage outbreaks can produce the collapse of aquatic bacterial populations and cholera
40 epidemics (6, 7), whereas sustained rates of phage-induced lysis positively affect ocean and
41 soil communities by facilitating carbon and nitrogen cycling (8, 9). More recent studies have
42 highlighted the potential of bacteriophages to shape host-associated microbiomes both
43 directly, through phage-driven selective pressures, and indirectly, through the alteration of the
44 host's immune response (10-13).

45 Prokaryotes have developed a diverse repertoire of specialized defense mechanisms against
46 viruses and other mobile genetic elements (MGE), such as plasmids and integrative and
47 conjugative/mobilizable elements (ICE/IME), many of which have only recently started to be
48 uncovered (14, 15). Among them, CRISPR-Cas systems are of particular interest as they provide
49 their bearers with adaptive immunity (16-18) The adaptive nature of CRISPR-Cas systems is
50 based on the acquisition of short DNA fragments from previously encountered MGE at the
51 leader end of the CRISPR array. Such fragments, called spacers, serve as highly specific probes
52 for the detection of the targeted MGE upon subsequent encounters. From a researcher's
53 perspective, CRISPR spacers constitute an ordered record that allow tracking virus-host
54 interactions through time (19, 20).

55 Despite experimental studies in the laboratory have provided a general understanding of how
56 CRISPR-Cas systems work, there are still many unknowns about their functioning in natural
57 ecosystems (21). The most prominent question regards the actual effect that CRISPR-based
58 immunity exerts on the composition and dynamics of microbial communities, both at the level
59 of the virome and the bacterial or archaeal host population. Studies in free-living model
60 ecosystems, such as hot springs, hypersaline lakes, and wastewater treatment plants, have

61 demonstrated that CRISPR-Cas contributes to shaping a highly dynamic network of interactions
62 among MGE and their prokaryotic hosts (19, 22-26). However, the paucity of appropriate
63 datasets has hampered the application of similar methods to complex host-associated
64 microbiomes. Some recurrent findings of previous studies are: (i) long-term coexistence of
65 MGE with a polymorphic population of hosts that includes immune and sensitive lineages; (ii)
66 high levels of local adaptation in the host CRISPR arrays; and (iii) high frequency of mutations
67 in regions of MGE targeted by CRISPR spacers. Coexistence of MGE and host lineages is
68 generally interpreted in terms of the diversity-promoting effects of clonal interference and
69 herd immunity (27-31), whereas coevolution of CRISPR and MGE suggests that CRISPR-Cas is
70 functional and the selective pressure that it imposes on MGE is strong enough to fuel
71 mutational escape (32, 33). Interestingly, these conclusions may not extend to host-associated
72 microbiomes. For example, single-sample measurements of *Prevotella* strains in the human gut
73 and *Pseudomonas aeruginosa* in the lung display complete immunodominance of a single host
74 lineage, possibly underscoring the occurrence of strong selective sweeps (34, 35).

75 A second major gap in our understanding of CRISPR-Cas systems concerns the time span of the
76 immune memory, which results from the interplay of spacer gain and loss within the CRISPR
77 array, mutation of the target MGE, and replacement of immune lineages in the population of
78 hosts. Mathematical models predict that the size of CRISPR arrays is controlled by a trade-off
79 between maintaining a large enough immune repertoire and avoiding the dilution effect of
80 expressing too many spacers. Likewise, theoretical arguments suggest that the rates of spacer
81 gain and loss must be finely tuned to keep CRISPR arrays sufficiently updated without losing
82 immune memory too quickly (36, 37). These models operate on the basic assumption that
83 distal spacers, generally assumed to be the oldest in the array (but see (38) for
84 counterexamples) provide the host with long-term immune memory. However, the empirical
85 and experimental evidence on this issue is inconclusive. On one hand, reduced transcription
86 rates (39-41) and broad conservation of distal spacers across strains (42) suggest that those

87 spacers do not play a relevant role in defense against local MGE. On the other hand, distal
88 spacers often match conserved regions of viral genomes, which supports an active role of
89 selection in maintaining these spacers for their higher potential in long-term immunity (43).

90 In this study, we leveraged a large longitudinal dataset from the integrative Human
91 Microbiome Project (44) to investigate the joint dynamics of MGE and CRISPR-Cas immunity in
92 the gut microbiome. We found evidence of effective (but incomplete) CRISPR-Cas immunity
93 against lytic and lysogenic phages, short- and long-term coexistence of multiple MGE and host
94 lineages, and selection for long-lasting immunity against endemic MGE provided by spacers
95 from distal regions of CRISPR arrays.

96 **Methods**

97 Metagenomic data

98 Metagenomic and metatranscriptomic reads, assembled contigs, and taxonomic profiles from
99 the Extended Human Microbiome Project – Inflammatory Bowel Disease (HMP2-IBD) were
100 downloaded from https://ibdmdb.org. The metagenomic dataset includes 1,638 samples from
101 130 individuals, collected at 2-week intervals for approximately 1 year (44). Samples from
102 participants with Crohn's disease and ulcerative colitis were classified as dysbiotic or non-
103 dysbiotic by applying the same criteria as in the original study.

104 Assembly, annotation, and clustering of CRISPR arrays

105 For each sample of the HMP2-IBD dataset, we assembled putative CRISPR-Cas loci from
106 metagenomic reads with the tool CasCollect v1.0 (45). CasCollect runs VSEARCH (46) in an
107 iterative manner to identify reads that possibly belong to CRISPR-Cas loci and then assembles
108 such reads with SPAdes (47). As baits for the first iteration of VSEARCH, we used the sequences
109 of known CRISPR-associated genes (as provided by default by CasCollect) and an in-house
110 database of CRISPR repeats that combined the default list of repeats provided by CasCollect,

111 additional repeats from CRISPRdb (48) and Makarova et al. (49) clustered at 90% identity and
112 coverage, and a set of previously unreported repeats obtained by running CRISPRCasTyper
113 v1.8 (50) on the publicly available, preassembled contigs from the HMP2-IBD study. CasCollect
114 was run with options --meta --noannotate --nucl. Putative CRISPR-Cas loci assembled by
115 CasCollect were quality-filtered and annotated with CRISPRCasTyper v1.8, obtaining 79,475
116 high-confidence CRISPR arrays of which 4,282 (representing around 5% of the total arrays)
117 included a Cas operon in their neighborhood. For comparison, we also searched for CRISPR-Cas
118 loci using a "classical" sequential approach, consisting of whole-metagenome assembly
119 followed by identification of CRISPR loci with CRISPRCasTyper. We tried this alternative
120 approach with two assemblies: the one originally generated by the HMP2-IBD project (sample-
121 level assembly with MEGAHIT, as described in (44)) and an in-house assembly with SPAdes at
122 the individual level (coassembly of all the samples from the same individual). We found that
123 the CRISPR arrays assembled with CasCollect were the most diverse in terms of repeats and
124 spacers (Figure S1 and Table S1). Moreover, CasCollect produced more arrays of all possible
125 lengths than the sample-level assembly approach (comparisons of array lengths with
126 individual-level coassemblies are not informative because of possible chimaeras and implicit
127 dereplication of arrays present in multiple samples). All in all, we concluded that CasCollect is
128 more sensitive than whole-genome assembly-based methods and selected it for all
129 downstream analyses.

130 Spacers and repeats from high-confidence CRISPR arrays were dereplicated using CD-Hit-Est
131 v4.8.1 (51) with a 100% identity and coverage threshold and option *-g* 1 (slow and accurate
132 mode), resulting in a total of 11,373 different repeats and 153,745 different spacers. The
133 directionality of transcription in a subset of 4,054 arrays with adjacent Cas genes and >5
134 spacers was predicted with Cas Orientation (52), CRISPRLoci v1.0 (53), and CRISPRDirection
135 v2.0 (54). Among these three prediction tools, Cas Orientation showed the best agreement
136 with the expected drop in transcription for middle and distal spacers and was applicable to a

137 larger fraction of arrays (Figure S2). Therefore, we chose Cas Orientation as the default
138 method to present the results in the text and main figures. To explore the distribution of the
139 number of spacers per array, we used the 'powerlaw' python package (55).

140 A strict nesting criterion was applied to identify groups (or lineages) of highly similar CRISPR
141 arrays present within and across samples of the same individual. A CRISPR array was
142 considered nested within another CRISPR array if both shared the same or highly similar
143 repeats (>90% identity and coverage, clustered using CD-HIT-Est with option -*g* 1) and all the
144 spacers of the former were also harbored by the latter. Group-representative arrays were
145 identified as those that were not nested within any other array. Finally, clusters of CRISPR
146 arrays were built by considering all the arrays nested within the same group representative.
147 This algorithm allowed us to retrieve clusters of highly similar CRISPR arrays while accounting
148 for incomplete assemblies and spacer gain and loss over time. Because of that, we refer to
149 these clusters as CRISPR lineages or lineages of arrays. Note that, although the strict nesting
150 criterion can leave partially overlapping arrays unmerged, this minimizes the risk of generating
151 heterogeneous clusters with lineages that share conserved spacers but behave independently
152 in the timescale of the study.

153 <u>Search for CRISPR targets in the local microbiome</u>

154 Protospacers in the local microbiome were identified by running nucleotide BLAST against
155 metagenomic contigs previously assembled by HMP2-IBD, with spacers as queries, options "-
156 dust no -word_size 8", and a 95% identity and coverage thresholds. We discarded hits to
157 known CRISPR arrays (identified with CRISPRCasTyper) and hits located within a 100bp
158 neighborhood of known CRISPR repeats (found with nucleotide BLAST, options "-dust no -
159 word_size 8", 95% identity and coverage).

160 One spacer in the dataset produced an extremely high number of hits to the local
161 metagenome (16,327 hits, while the second largest number was 121). Closer inspection

162 revealed that the spacer was highly similar to a transposase sequence found in bacterial
163 chromosomes and prophages. To avoid biases in downstream analyses, the spacer was
164 removed from the dataset.

165 Identification of spacer targets

166 The targets of CRISPR of spacers were inferred by performing a nucleotide BLAST search of all
167 spacers against a merged database of mobile genetic elements (MGEdb), setting BLAST
168 parameters for short queries (-dust no -word_size 8) and keeping all hits with >95% identity
169 and coverage. The merged MGEdb included 198,607 complete viral genomes from the NCBI
170 Genome database (downloaded on 27/01/2021 from ftp.ncbi.nlm.nih.gov/genomes/Viruses/)
171 and the NCBI Virus database (downloaded in interactive mode on 27/01/2021 from
172 https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&Completeness_
173 s=complete), a set of 1,887 gut phages recently identified by Benler et al. (56), a set of 13,274
174 ICE/IME from Botelho (57), and all sequences from mMGE v1.0 (58) (494,835 phages and
175 92,493 plasmids), ImmeDb (59) (201 phages and 4,727 ICE/IME, transposable elements, and
176 other unclassified MGE), ICEberg3 (60) (4,461 ICE/IME and other chromosomally integrated
177 elements), and IMG/PR (61) (699,973 plasmids). Based on the BLAST results, we classified
178 spacers according to their most likely target (virus, plasmid, or other mobile genetic elements).
179 Spacers with inconsistent target predictions were classified as “ambiguous”. Spacer-target
180 assignments were annotated according to the accuracy of the match as identical (100%
181 identity and coverage) or highly similar (identity and coverage above 95% but below 100%).
182 Moreover, viral targets were taxonomically annotated according to the information in MGEdb.
183 Spacers targeting phages from the infant gut were identified by running BLAST searches
184 against the database generated by Shah et al (62).

185 To infer the target of spacers that did not match any sequence in MGEdb but had protospacers
186 in the local microbiome, we extracted up to 15 kb neighborhoods (7.5 kb upstream and

187 downstream from the protospacer, which corresponds to the mean length of phage genomes
188 in MGEdb) from the HMP2-IBD contigs and ran a nucleotide BLAST search (option dc-
189 megablast) against MGEdb setting an e-value threshold of 10^{-8}. If the search returned hits from
190 multiple MGE, we discarded those with bitscores lower than 0.9 times the bitscore of the best
191 hit. Compared to just keeping the best hit, this approach preserves the uncertainty among
192 high-scoring candidates. Spacers with inconsistent predictions were classified as "ambiguous"
193 and those whose target remained unclassified after this second step were labeled as
194 "unknown".

195 Prediction of lytic or lysogenic phage lifestyle

196 Phages targeted by CRISPR spacers were classified as lytic or lysogenic based on the presence
197 of integrases in phage genomes. To that end, we extracted 49,645 phage genomes from
198 MGEdb and 110,712 metagenome-assembled contigs from HMP2-IBD that contained
199 protospacers and had been previously characterized as viral (see previous section). Then, we
200 ran Prodigal v2.6.3 (63) with the option *-p meta* to identify open reading frames and searched
201 for hits to viral integrases using the *hmmersearch* command of the HMMER v3.3.2 suite. We
202 included the following PFAM profiles for integrases: PF00239, PF00589, PF02899, PF07508,
203 PF12482, PF12834, PF13009, PF13102, PF13495, PF14659, PF16795, and PF18644. The search
204 returned 11,631 viral sequences with putative integrases (e-value $<10^{-8}$) that also contained
205 protospacers. The corresponding spacers (27,868 in total) were regarded as targeting lysogenic
206 phages.

207 Prevalence and abundance of CRISPR targets in the local microbiome

208 To study the quantitative dynamics of CRISPR targets in the microbiome, we focused on
209 spacer-protospacer pairs present in samples from the same individual (possibly at different
210 times) that fulfil the following conditions: (1) the first appearance of the spacer occurs later
211 than the first appearance of the CRISPR lineage with which the spacer is associated; and (2)

212 spacers are present in at least 2 samples. These two conditions allowed us to compare target
213 abundances before and after the first appearance of the spacer, while controlling for the
214 intermittent prevalence of hosts (using CRISPR lineages as proxies for the latter). For the same
215 reason, we excluded from these analyses all samples in which the CRISPR lineage was not
216 detected. After applying these filters, 1540 spacer-protospacer pairs remained, representing
217 1057 CRISPR lineages and 5369 protospacer occurrences.

218 To estimate the prevalence of targets, we counted the number of samples that contain the
219 protospacer before, at, and after the first appearance of the spacer, pooled these counts for all
220 spacer-protospacer pairs, and normalized them by the number of samples in each group.
221 Confidence intervals were calculated using the Clopper-Pearson method as implemented in
222 the *statsmodels* Python library. To quantify target abundances, we extracted 4kb
223 neighborhoods (2kb upstream and downstream) of protospacers from HMP2-IBD contigs and
224 mapped metagenomic reads to those neighborhoods with CoverM v0.6.1. The relatively small
225 size of these neighborhoods was chosen to minimize the risk of including flanking regions from
226 host genomes in the case of integrated MGE. Sample-wise relative abundances were
227 calculated as the number of mapped reads divided by the length of the mapped sequence
228 (generally 4kb, except for protospacers located near contig ends) and by the total number of
229 reads in the sample. The sample-wise abundance in samples in which the target was not
230 detected was set to 0. Finally, to facilitate the comparison of trends across targets, we
231 normalized the sample-wise abundances of each target by the maximum abundance observed
232 for that target in the whole time series.

233 <u>Taxonomic assignation of CRISPR-Cas systems</u>

234 To assign a taxonomic origin to CRISPR arrays, we leveraged the fact that CRISPR repeats are
235 taxon-specific (although the taxonomic level of such specificity varies among repeats) (64).
236 First, we ran CRISPRCasTyper on the preassembled HMP2-IBD metagenomic contigs to predict

237 CRISPR arrays. Then, we used CD-HIT to identify contigs from HMP2-IBD that contain CRISPR
238 arrays with the same repeats (90% identity and coverage) as those assembled by CasCollect.
239 Those contigs were provided as inputs to the program CAT/BAT v5.2.3 (65) for taxonomical
240 assignment at the lowest non-ambiguous available level. For CRISPR arrays whose repeats
241 could not be mapped to any HMP2-IBD contig, we directly ran CAT/BAT on the contigs
242 obtained with CasCollect. (Note, however, that the level of resolution obtained with CasCollect
243 contigs was generally lower due to their shorter length.) Finally, all CRISPR arrays that
244 remained unassigned and were part of a lineage with one or more assigned members were
245 given the same taxonomical origin as the other members of the lineage, as long as no
246 inconsistencies were found.

247 Spacer location bias and asymmetry

248 We studied the relative position of spacers that target phages from the local metagenome
249 (henceforth "locally adapted spacers") in a subset of 4,054 oriented arrays with >5 spacers,
250 separately considering identical and highly similar hits (CRISPR-Cas subtypes with <25 valid
251 arrays were not included in this analysis). To avoid artifacts associated with finite-size effects in
252 small arrays, we used a sliding-window approach to estimate the density of locally adapted
253 spacers along the array. We set a window width equal to 12% of the total length of the array
254 (equivalent to 2/3 of the inverse of the median number of spacers) and a step equal to 1.5% of
255 the total length of the array (1/8 of the window width). This choice results in a window size
256 slightly smaller than the typical span of a spacer (given by the inverse of the median number of
257 spacers per array), which provides a good balance between resolution and smoothness.

258 Quantification of transcription levels along CRISPR arrays

259 For 753 samples from HMP2-IBD with available metatranscriptomes, we mapped
260 metatranscriptomic reads to CRISPR assemblies obtained from the same samples with
261 minimap2 v2.24, setting the option '-x sr' for short reads (66). To improve specificity towards

262 spacers in CRISPR arrays, we also included contigs with protospacers in the reference genomic
263 dataset and removed all mapped reads with aligned lengths below 70 nucleotides using the
264 command FilterSamReads from gatk 4.1.2.0 (67). This 70-nucleotide threshold was chosen to
265 ensure that alignments covered the equivalent of one spacer and one repeat (the average
266 spacer and repeat lengths in our dataset are 34 and 31, respectively). The coverage per
267 position was calculated with the command genomeCoverageBed from BEDTools 2.28.0 (68).

268 Clustering and classification of CRISPR-spacer-target dynamics

269 To study the temporal dynamics of CRISPR lineages, spacers, and targets, we first identified all
270 the spacers that were observed for the first time later than their corresponding CRISPR
271 lineage. Besides this basic requirement, we did not impose any further condition to filter
272 potential acquisition events (for example, requiring that the new spacer was located at the
273 leading end) because predicted orientations were only available for a small fraction of arrays.
274 (Moreover, we could not discard that more than 1 spacers were incorporated between two
275 sampling points, which would make newly acquired spacers look as if they were internal.) We
276 codified the time series associated with each spacer as a string of characters, using letters A-F
277 to indicate all possible combinations for the presence or absence of that particular spacer, its
278 target, and its corresponding CRISPR lineage. “Empty” states (that is, those characterized by
279 the absence of spacer, CRISPR lineage, and target) were removed from both ends of the string
280 and collapsed to a single instance if appearing more than once between non-empty states (see
281 Supplementary Note S1 for a detailed justification of this procedure). To cluster similar time
282 series and obtain representative trajectories, we performed pairwise alignments of all strings
283 with the Biopython command pairwise2.align.globalms (69) using the following scores: 1 per
284 match, -2 per mismatch, -2 per gap opening, and -1 per gap extension (see Supplementary
285 Note S1 and Figure S3 for further details and discussion of alternative scoring schemes). After
286 normalizing the scores by the length of the longest member of each pair and discarding pairs

287 with negative scores, we built a time-series similarity network using normalized scores as
288 weights. Time series were then clustered with Infomap v0.21.0 (70), with options -2 -u -N 200
289 (two-level partition assuming undirected links and selecting the best of 200 trials).

290

291 **Results and discussion**

292 Assembly and characterization of CRISPR arrays in the human gut microbiome

293 We combined state-of-the-art methods to assemble CRISPR arrays from the gut microbiomes
294 of 130 individuals, recurrently sampled every 2 weeks over a 1-year period as part of the
295 integrative Human Microbiome Project (Figure 1a). After curation and removal of low-quality
296 arrays, we obtained a set of 79,475 CRISPR arrays, encompassing 11,373 unique repeats and
297 153,745 unique spacers. Of those, around 4,300 arrays were adjacent to a Cas operon, what
298 allowed us to predict the direction of transcription by applying three alternative criteria (see
299 Methods). We identified CRISPR arrays in 36 out of the 40 most abundant bacterial genera in
300 all gut samples, which collectively account for 94% of bacterial abundance (Suppl. Table S2).
301 The number of unique CRISPR arrays per sample broadly varies through time and across
302 microbiomes, with significantly lower numbers in microbiomes obtained from patients with
303 Crohn's disease and ulcerative colitis (Figure 1b-c). Such lower numbers are consistent with the
304 reduced microbial diversity associated with dysbiosis in inflammatory bowel diseases (44).

305 The CRISPR array sizes follow a heavy-tailed distribution that is well-fit by a truncated power
306 law with exponent close to -2 and a cutoff at around 100 spacers (Figure 1d). Similar
307 distributions, although with higher cutoffs, have been reported for CRISPR arrays from
308 complete genomes (71, 72). Thus, the high heterogeneity of array lengths observed in this
309 dataset likely reflects a true property of CRISPR systems in the gut microbiome, rather than an
310 excess of fragmentary or poorly-assembled arrays.

311 We applied a strict nesting criterion to group CRISPR arrays into 41,742 clusters, of which
312 9,870 contain >1 array (Figure 1e). These clusters represent lineages of arrays from the same
313 microbiome (possibly sampled at different times) that are likely related by sequential gain and
314 loss of spacers. Array lineages serve a triple purpose in this study. First, lineages are pivotal to
315 assess the consistency of CRISPR assemblies by comparing spacer order and orientation in
316 arrays from the same lineage. Second, lineages permit a robust characterization of CRISPR
317 diversity by minimizing the impact of incomplete arrays and abundance biases. Third, lineages
318 constitute a convenient proxy for the presence of subpopulations of hosts in the microbial
319 community.

320 Within-lineage comparisons show that spacer order and orientation are highly consistent
321 across pairs of related arrays (Figure 1f). Such high consistencies are not a by-product of the
322 method used to group arrays into lineages, which was solely based on spacer presence or
323 absence and did not discriminate between spacers and their reverse complements. Overall,
324 internal rearrangements only affect 3.6% of all array pairs, whereas spacer inversions are
325 virtually inexistent (<0.01%). Predicted array orientations are also generally consistent with
326 respect to spacer order, with inconsistencies observed in <1% of the pairs. Because lineages
327 include arrays separately assembled in different samples, within-lineage consistency strongly
328 suggests that CRISPR assemblies are reliable in terms of spacer order and orientation.

329 The most abundant CRISPR-Cas systems in the human gut microbiome belong to subtypes I-B,
330 I-C, I-E, and II-C (Figure 2). Subtypes I-B, I-C, and II-C are most often associated with Firmicutes
331 and Bacteroidetes, whereas subtype I-E is observed in Firmicutes and Proteobacteria.
332 Alterations in the composition of the microbiome significantly affect the profile of CRISPR
333 subtypes in Crohn's disease and ulcerative colitis (Figure 2 and Suppl. Table S2). In agreement
334 with previous studies (44, 73), dysbiosis leads to a decrease in the relative proportion of
335 subtypes I-B, I-C, I-E, and II-C that can be attributed to the depletion of *Faecalibacterium*

336 *prausnitzii* (I-C and I-E) and several groups of Bacteroidetes (I-B, I-C, and II-C). Among the less
337 abundant subtypes, IV-A (typically carried by large conjugative plasmids (74)) stands out for its
338 strong association with normal (that is, no dysbiotic) samples.

339 Selection promotes long-term immunity against endemic MGE

340 To identify the potential targets of CRISPR immunity, we mapped the spacers in CRISPR arrays
341 to a large collection of phages, plasmids, and other MGE obtained from state-of-the-art
342 (meta)genomic databases (see Methods). Overall, 41% of the spacers show high similarity
343 (zero or one mismatch) to known phages and only 4% map to plasmids and other MGE (Figure
344 3a). These values are notably higher than those reported by global surveys of CRISPR spacers
345 from complete prokaryotic genomes (75, 76) but close to those obtained for gut bacteria (43,
346 77), probably reflecting the higher coverage of MGE from the human gut in public databases.
347 Around 6% of the spacers target sequences from the local microbiome (defined as the union of
348 all samples collected from the same individual). We refer to these spacers as "locally adapted"
349 because they can provide immunity against MGE that are locally observed in the 1-year period
350 covered by the dataset. The most frequent targets of locally adapted spacers are lysogenic
351 phages (43%), followed by lytic and other non-lysogenic phages (28%), plasmids (16%), and
352 other MGE (1%). Around 10% of the spacers (labeled as "ambiguous" in Figure 3a-b) match
353 sequences found in multiple classes of MGE. All in all, we could assign a unique or ambiguous
354 origin to 99% of all locally adapted spacers. Therefore, the disproportionate number of hits to
355 phages and plasmids compared to other MGE is probably genuine, despite the admittedly
356 biased representation of different classes of MGE in public databases. Comparing among
357 CRISPR subtypes (Figure 3b), local adaptation is more frequent in subtypes I-C and II-C (7.7-
358 8.5% of spacers are locally adapted, compared to 2.5-4.2% in other subtypes; χ^2 test $p < 10^{-10}$).
359 Moreover, subtypes II-A, II-C, and III-A target a higher proportion of plasmids than class I
360 systems (12% in I-C vs 25% in II-C; Fisher's exact test $p < 10^{-10}$).

361 Locally adapted spacers are not uniformly distributed along the CRISPR array (Figure 3c). On
362 the contrary, they tend to concentrate near both ends of the array. The enrichment at the
363 leader (or proximal) end is consistent with recent spacer acquisitions targeting MGE from the
364 local microbiome. However, the excess of locally adapted spacers near the trailer (or distal)
365 end, especially evident in subtypes I-C and I-E, is somewhat paradoxical. In fact, distal spacers
366 are generally the oldest in the array and their role in immunity is a matter of debate (78, 79).
367 Three lines of evidence reject the possibility that locally adapted spacers near the distal end
368 represent a technical artifact. First, array orientation and spacer order are highly consistent
369 across arrays from the same lineage, despite the arrays being independently assembled from
370 different samples. Second, the same trend is observed when the orientation of the arrays is
371 inferred with three independent methods (Cas Orientation, CRISPRDirection, and CRISPRLoci;
372 Figure S2). Third, in most subtypes, mismatches between the spacer and target sequences
373 become more frequent towards the trailer end, as expected for correctly assembled arrays.
374 CRISPR subtypes I-C and I-E contain 85% of all locally adapted spacers within the set of high-
375 quality arrays (those with adjacent Cas genes and 5 or more spacers). Therefore, despite the
376 concentration of locally adapted spacers near the trailer end is not observed in all subtypes, it
377 arguably represents a dominant trend in the gut microbiome.

378 We conceive two possible biological explanations for the observed trend in the distribution of
379 locally adapted spacers. The first is that CRISPR-based immunity leads to the complete
380 elimination of MGE targeted by spacers in the middle section of the array, whereas MGE
381 targeted by distal, less efficient spacers can persist in the microbiome. In this scenario, most
382 middle spacers would not be identified as “locally adapted” simply because they succeeded in
383 eliminating their targets. Although CRISPR immunity can effectively prevent MGE proliferation
384 in culture (80, 81), it is less likely that complete elimination of MGE occurs in complex, spatially
385 extended microbial communities (26, 82) (and the human gut virome appears indeed highly
386 stable over time (83)). More generally, this scenario does not account for the possibility that

387 the same MGE is targeted by multiple spacers located at different positions in different arrays,
388 which would effectively decouple within-array spacer positions and population-level MGE
389 abundances.

390 The second (and more likely) explanation is that locally adapted spacers near the trailing end
391 are a consequence of sustained selection for long-term immunity against endemic MGE. It is
392 widely accepted that intense selection and frequent spacer loss via homologous recombination
393 shape the composition of CRISPR arrays in natural populations, promoting the maintenance of
394 functional spacers (42). Therefore, if endemic MGE impose a sustained burden on the
395 population, it can be expected that older spacers providing with effective long-term immunity
396 against such MGE accumulate at the distal part of the array. This phenomenon has been
397 observed in *Leptospirillum* strains from an acid mine drainage where the same phage
398 population persisted for at least 5 years (23, 84). We propose a similar explanation in the case
399 of the adult human gut, where bacteria coexist with a very stable virome dominated by crass-
400 like phages (83, 85-87) and are exposed to continuous (but low-intensity) induction of
401 prophages (88, 89). In support of the long-term memory role of distal CRISPR spacers, we
402 found that the fraction of spacers targeting lysogenic and crass-like phages systematically
403 increases from the leading to the trailing end of the array (Figure 3d-e).

404 Considering the high rates of lytic activity and CRISPR spacer turnover that characterize the
405 infant gut, and the possible contribution of early phage-bacteria interactions to shaping the
406 adult gut microbiome (13, 62, 90, 91), an intriguing possibility might be that long-term memory
407 spacers were selected during the consolidation of a coexisting community of MGE and bacteria
408 in the first years of age. To test that hypothesis, we investigated the relative abundance of
409 spacers targeting phages from the infant gut in leading, middle, and trailing regions of CRISPR
410 arrays (Figure 3f). Interestingly, spacers that target infant gut phages do not concentrate at
411 the trailing end (except in subtype II-A), but they represent 60-80% of all the phage-targeting

412 spacers in the leading region of the array. These results confirm that exposure to phages in the
413 infant gut is a major trigger of CRISPR adaptation, though it does not seem to affect selection
414 for long-term immune memory.

415 To provide efficient long-term immune memory, spacers near the trailing end of the array
416 must be expressed at sufficient levels to form effector complexes. Indeed, lack of expression
417 due to early termination of transcription limits the functionality of distal spacers in
418 experimental conditions (39-41). By analyzing metatranscriptomes, we found that distal
419 spacers are moderately expressed, with transcript abundances around 30-40% of those
420 observed for spacers near the leading end (Figure S4). Distal spacers may also contribute to
421 long-term immunity through primed adaptation, that is, by promoting the acquisition of
422 additional spacers from cognate MGE (92). Primed adaptation is widespread in type I and II
423 CRISPR systems (93, 94) and operates for spacers that do not exactly match their target (95) or
424 are expressed at low levels (96). Because spacers acquired through this mechanism are often
425 biased in terms of genomic location and strand specificity, indirect evidence for primed
426 adaptation can be obtained by testing for non-random patterns in the relative location of
427 protospacers within the target genome. Following that strategy, we identified several
428 instances of MGE targeted by >1 spacer from the same array, which could correspond to
429 primed adaptation. However, the number of instances observed was insufficient for statistical
430 confirmation.

431 CRISPR-Cas immunity modulates microbiome composition

432 The longitudinal nature of the HMP2-IBD metagenomic survey allowed us to investigate how
433 CRISPR immunity affects the dynamics of MGE and bacterial hosts in the gut microbiome. We
434 first studied if the appearance of a new spacer in a preexisting CRISPR lineage (possibly
435 implying a spacer acquisition event) was associated with changes in the prevalence and
436 abundance of the target MGE (Figures 4a and S5a). Consistent with the expectation that

437 exposure to MGE induces immune adaptation, we found that putative spacer acquisitions
438 occur in samples in which the abundance of the target MGE is higher than average. In sharp
439 contrast, subsequent observations of the spacer are associated with significantly decreased
440 MGE abundances. Notably, such decrease is not observed in populations that have seemingly
441 lost the spacer but in which the CRISPR lineage (and therefore the host) is still detected. As
442 expected, these trends are strongest in the case of lytic and lysogenic phages, but also
443 significant for plasmids. Despite its quantitative effect on MGE abundance, CRISPR-Cas
444 immunity does not generally lead to the complete elimination of its targets. Indeed, around
445 50% of lytic and lysogenic phages targeted by CRISPR-Cas can still be detected following the
446 acquisition of a cognate spacer in the population (Figures 4b and S5b). In the case of plasmids,
447 acquisition of cognate spacers does not produce a significant drop in prevalence. All in all,
448 these observations imply that CRISPR-Cas immunity modulates phage and plasmid abundances
449 in the human gut, while not compromising their long-term persistence. This conclusion is
450 consistent with previous observations from free-living communities and with the theoretical
451 prediction that CRISPR-Cas immunity can stabilize complex phage-host populations through
452 eco-evolutionary feedback (28, 97-99).

453 We then investigated if the interplay of CRISPR immunity and phage predation affects host
454 diversity at the population level. To that end, we divided CRISPR lineages into those that are
455 possibly involved in effective immune interactions (by harboring at least one locally adapted
456 spacer) and those that do not show any evidence of local adaptation. The former tend to be
457 more diverse, that is, they encompass more array variants per lineage, likely reflecting higher
458 rates of spacer acquisition and loss (Figure 4c). Moreover, the diversity of CRISPR arrays in
459 these lineages is lower in samples in which targeted phages are present (Figure 4d). This
460 observation underscores the occurrence of weak MGE-induced selective sweeps in the host
461 population, that transiently favor those hosts bearing CRISPR arrays with cognate spacers
462 (Figure 4e). The weakness of these selective sweeps could in part be explained by clonal

463 interference among multiple immune lineages, as previously suggested for hot spring
464 communities (27), and by the spatially structured nature of the gut microbiome.

465 Temporal dynamics of MGE and CRISPR arrays

466 To better understand the interplay between CRISPR-Cas immunity and the dynamics of MGE
467 and hosts, we selected 1,769 time series for single spacer-target pairs whose dynamics are
468 compatible with possible spacer acquisition events at the population level. Then, we applied
469 an unsupervised clustering algorithm to group similar trajectories based on the temporal
470 patterns of presence or absence of spacers, targets, and CRISPR lineages (the latter serving as
471 proxies for host subpopulations). Further details about this procedure are provided in the
472 Methods section, Supplementary Note S1, and Figure S3, where we show that the final clusters
473 are stable with respect to small methodological variations. Despite the potentially
474 combinatorial diversity of individual trajectories, we found that 40% of them belonged to one
475 of the 10 largest clusters. The inspection of such representative clusters reveals qualitative
476 differences across MGE in their temporal dynamics (Figures 5 and S6). Phages are
477 preferentially associated with episodic dynamics (TC-1, 3, 5, and 7) involving type I-C CRISPR-
478 Cas systems. In turn, plasmids display recurrency (TC-13 and 14) and are most often targeted
479 by type II-C systems. Although array incompleteness could, in principle, introduce artifacts in
480 time series in the form of apparent spacer loss with lineage persistence, such transitions were
481 rare or absent from the major trajectory clusters (Figure 5a).

482 One of the most frequent trajectories in the data set (TC-1) corresponds to a single
483 observation of a CRISPR lineage followed by a simultaneous observation of a new spacer and
484 the cognate phage, in what might represent a spacer acquisition event. Other representative
485 trajectories involve variations of the same theme, with loss of the phage following spacer
486 acquisition (TC-5) or missing the intermediate acquisition event (TC-2 and 3). In contrast, the
487 most frequent trajectories associated with plasmids represent two alternative scenarios, one

488 dominated by the plasmid (TC-13) and the other characterized by the stable coexistence of
489 plasmid-bearing and plasmid-resistant host subpopulations (TC-14). The latter is reminiscent of
490 the global population structure of *Pseudomonas aeruginosa*, in which spacers from active
491 CRISPR-Cas systems target MGE integrated in closely related genomes that lack CRISPR (100),
492 and in hot spring populations of *Hydrogenobacter sp.*, where the spacers of single-cell
493 assembled genomes target MGE found in other cells from the same population (101). Our
494 results suggest that such crosstalk between CRISPR-Cas systems and plasmids that coexist in
495 the host population is also widespread in the human gut.

496

497 **Conclusions**

498 By analyzing a large metagenomic dataset encompassing 130 individual gut microbiomes
499 periodically sampled for up to 1 year, we found that locally adapted spacers targeting
500 lysogenic and crass-like phages are often located in the distal region of CRISPR arrays. This
501 finding strongly suggests that selection promotes the maintenance of long-term memory
502 spacers against endemic MGE, and that periodic reactivation of prophages is an important
503 stressor for the human gut microbiome. Interestingly, leader and trailer regions of CRISPR
504 arrays appear to be specialized in short- and long-term memory, respectively. We expect that
505 selection for long-lasting immunity and polar allocation of long-term memory spacers are not
506 unique to the human gut microbiome, but general features of microbial communities with
507 stable viromes. Testing such prediction constitutes an interesting subject of future research.

508 A second major finding of this study was that the interaction of phages and CRISPR-Cas
509 immunity quantitatively affects the composition of the human gut microbiome. Phage and
510 plasmid abundances tend to decrease after the first appearance of cognate CRISPR spacers in
511 the microbiome, although such decrease does not generally lead to their complete elimination.
512 On the other side, in the presence of MGE, bacterial populations experience weak and

513 incomplete selective sweeps favoring immune lineages. The coexistence of phages and diverse
514 host lineages despite effective CRISPR-Cas immunity suggests that spatial heterogeneity, clonal
515 interference, and fluctuating selection could have a major role in explaining phage-bacteria
516 interaction dynamics in the gut. Disentangling the contribution of these mechanisms warrants
517 further investigation, as they may differentially affect the efficacy of phage therapy and other
518 intervention strategies aimed at controlling the gut microbiome.

519

520 **Acknowledgements**

521 A.L-B. is supported by the Agencia Estatal de Investigación of Spain (Grant No. PRE2020-
522 092935). J.B. is supported by the Maria Zambrano grant of the Spanish Ministry of Universities
523 (Grant No. UP2021-035), and the Severo Ochoa Program for Centres of Excellence in R&D of
524 the Agencia Estatal de Investigación of Spain (Grant No. CEX2020-000999-S (2022–2025) to the
525 CBGP). J.I is supported by the Ramón y Cajal Programme of the Spanish Ministry of Science
526 (Grant No. RYC-2017–22524); the Agencia Estatal de Investigación of Spain (Grant No.
527 PID2019-106618GA-I00), the Severo Ochoa Programme for Centres of Excellence in R&D of the
528 Agencia Estatal de Investigación of Spain (Grant No. SEV-2016–0672 (2017–2021) to the
529 CBGP); and the Comunidad de Madrid (through the call Research Grants for Young
530 Investigators from Universidad Politécnica de Madrid, Grant No. M190020074JIIS).

531

532 **Data availability**

533 All relevant data produced in this study have been deposited in the Zenodo repository (DOI:
534 10.5281/zenodo.10555110) and are freely available.

535 **References**

536

537 1. Diaz-Munoz SL, Koskella B. Bacteria-phage interactions in natural environments. Adv
538 Appl Microbiol. 2014;89:135-83.
539 2. Fernandez L, Rodriguez A, Garcia P. Phage or foe: an insight into the impact of viral
540 predation on microbial communities. The ISME journal. 2018;12(5):1171-9.
541 3. Chevallereau A, Pons BJ, van Houte S, Westra ER. Interactions between bacterial and
542 phage communities in natural environments. Nat Rev Microbiol. 2022;20(1):49-62.
543 4. Brown TL, Charity OJ, Adriaenssens EM. Ecological and functional roles of
544 bacteriophages in contrasting environments: marine, terrestrial and human gut. Curr Opin
545 Microbiol. 2022;70:102229.
546 5. Koskella B, Brockhurst MA. Bacteria-phage coevolution as a driver of ecological and
547 evolutionary processes in microbial communities. FEMS Microbiol Rev. 2014;38(5):916-31.
548 6. Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, Breitbart M, et al. Viral and
549 microbial community dynamics in four aquatic environments. The ISME journal. 2010;4(6):739-
550 51.
551 7. Faruque SM, Islam MJ, Ahmad QS, Faruque AS, Sack DA, Nair GB, et al. Self-limiting
552 nature of seasonal cholera epidemics: Role of host-mediated amplification of phage.
553 Proceedings of the National Academy of Sciences of the United States of America.
554 2005;102(17):6119-24.
555 8. Suttle CA. Marine viruses--major players in the global ecosystem. Nat Rev Microbiol.
556 2007;5(10):801-12.
557 9. Braga LPP, Spor A, Kot W, Breuil MC, Hansen LH, Setubal JC, et al. Impact of phages on
558 soil bacterial communities and nitrogen availability under different assembly scenarios.
559 Microbiome. 2020;8(1):52.
560 10. Koskella B. Phage-mediated selection on microbiota of a long-lived host. Curr Biol.
561 2013;23(13):1256-60.
562 11. Manrique P, Dills M, Young MJ. The Human Gut Phage Community and Its Implications
563 for Health and Disease. Viruses. 2017;9(6).
564 12. Gogokhia L, Buhrke K, Bell R, Hoffman B, Brown DG, Hanke-Gogokhia C, et al.
565 Expansion of Bacteriophages Is Linked to Aggravated Intestinal Inflammation and Colitis. Cell
566 Host Microbe. 2019;25(2):285-99 e8.
567 13. Shamash M, Maurice CF. Phages in the infant gut: a framework for virome
568 development during early life. The ISME journal. 2022;16(2):323-30.
569 14. Millman A, Melamed S, Leavitt A, Doron S, Bernheim A, Hor J, et al. An expanded
570 arsenal of immune systems that protect bacteria from phages. Cell Host Microbe.
571 2022;30(11):1556-69 e5.
572 15. Georjon H, Bernheim A. The highly diverse antiphage defence systems of bacteria. Nat
573 Rev Microbiol. 2023;21(10):686-700.
574 16. Watson BNJ, Steens JA, Staals RHJ, Westra ER, van Houte S. Coevolution between
575 bacterial CRISPR-Cas systems and their bacteriophages. Cell Host Microbe. 2021;29(5):715-25.
576 17. Meaden S, Biswas A, Arkhipova K, Morales SE, Dutilh BE, Westra ER, et al. High viral
577 abundance and low diversity are associated with increased CRISPR-Cas prevalence across
578 microbial ecosystems. Curr Biol. 2022;32(1):220-7 e5.
579 18. Chabas H, Muller V, Bonhoeffer S, Regoes RR. Epidemiological and evolutionary
580 consequences of different types of CRISPR-Cas systems. PLoS computational biology.
581 2022;18(7):e1010329.

582 19. Martinez Arbas S, Narayanasamy S, Herold M, Lebrun LA, Hoopmann MR, Li S, et al.
583 Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics
584 revealed using time-series integrated meta-omics. Nature microbiology. 2021;6(1):123-35.
585 20. Cornuault JK, Moncaut E, Loux V, Mathieu A, Sokol H, Petit MA, et al. The enemy from
586 within: a prophage of Roseburia intestinalis systematically turns lytic in the mouse gut, driving
587 bacterial adaptation by CRISPR spacer acquisition. The ISME journal. 2020;14(3):771-87.
588 21. Westra ER, van Houte S, Gandon S, Whitaker R. The ecology and evolution of microbial
589 CRISPR-Cas adaptive immune systems. Philos Trans R Soc Lond B Biol Sci.
590 2019;374(1772):20190101.
591 22. Emerson JB, Andrade K, Thomas BC, Norman A, Allen EE, Heidelberg KB, et al. Virus-
592 host and CRISPR dynamics in Archaea-dominated hypersaline Lake Tyrrell, Victoria, Australia.
593 Archaea. 2013;2013:370871.
594 23. Sun CL, Thomas BC, Barrangou R, Banfield JF. Metagenomic reconstructions of
595 bacterial CRISPR loci constrain population histories. The ISME journal. 2016;10(4):858-70.
596 24. Pauly MD, Bautista MA, Black JA, Whitaker RJ. Diversified local CRISPR-Cas immunity to
597 viruses of Sulfolobus islandicus. Philos Trans R Soc Lond B Biol Sci. 2019;374(1772):20180093.
598 25. Medvedeva S, Liu Y, Koonin EV, Severinov K, Prangishvili D, Krupovic M. Virus-borne
599 mini-CRISPR arrays are involved in interviral conflicts. Nature communications.
600 2019;10(1):5204.
601 26. Guerrero LD, Perez MV, Orellana E, Piuri M, Quiroga C, Erijman L. Long-run bacteria-
602 phage coexistence dynamics under natural habitat conditions in an environmental
603 biotechnology system. The ISME journal. 2021;15(3):636-48.
604 27. Held NL, Herrera A, Cadillo-Quiroz H, Whitaker RJ. CRISPR associated diversity within a
605 population of Sulfolobus islandicus. PLoS One. 2010;5(9).
606 28. Iranzo J, Lobkovsky AE, Wolf YI, Koonin EV. Evolutionary dynamics of the prokaryotic
607 adaptive immunity system CRISPR-Cas in an explicit ecological context. Journal of bacteriology.
608 2013;195(17):3834-44.
609 29. Iranzo J, Lobkovsky AE, Wolf YI, Koonin EV. Immunity, suicide or both? Ecological
610 determinants for the combined evolution of anti-pathogen defense systems. BMC evolutionary
611 biology. 2015;15:43.
612 30. Payne P, Geyrhofer L, Barton NH, Bollback JP. CRISPR-based herd immunity can limit
613 phage epidemics in bacterial populations. Elife. 2018;7.
614 31. Iranzo J, Faure G, Wolf YI, Koonin EV. Game-Theoretical Modeling of Interviral Conflicts
615 Mediated by Mini-CRISPR Arrays. Front Microbiol. 2020;11:381.
616 32. Andersson AF, Banfield JF. Virus population dynamics and acquired virus resistance in
617 natural microbial communities. Science. 2008;320(5879):1047-50.
618 33. Kosmopoulos JC, Campbell DE, Whitaker RJ, Wilbanks EG. Horizontal Gene Transfer
619 and CRISPR Targeting Drive Phage-Bacterial Host Interactions and Coevolution in "Pink Berry"
620 Marine Microbial Aggregates. Applied and environmental microbiology. 2023;89(7):e0017723.
621 34. England WE, Kim T, Whitaker RJ. Metapopulation Structure of CRISPR-Cas Immunity in
622 Pseudomonas aeruginosa and Its Viruses. mSystems. 2018;3(5).
623 35. Devoto AE, Santini JM, Olm MR, Anantharaman K, Munk P, Tung J, et al. Megaphages
624 infect Prevotella and variants are widespread in gut microbiomes. Nature microbiology.
625 2019;4(4):693-700.
626 36. Bradde S, Nourmohammad A, Goyal S, Balasubramanian V. The size of the immune
627 repertoire of bacteria. Proceedings of the National Academy of Sciences of the United States of
628 America. 2020;117(10):5144-51.
629 37. Weissman JL, Fagan WF, Johnson PLF. Selective Maintenance of Multiple CRISPR Arrays
630 Across Prokaryotes. CRISPR J. 2018;1(6):405-13.
631 38. Ou Y, McInerney JO. High Frequency of Dynamic Rearrangements In Crispr loci.
632 bioRxiv. 2022:2022.05.19.492656.

633 39. Richter H, Zoephel J, Schermuly J, Maticzka D, Backofen R, Randau L. Characterization
634 of CRISPR RNA processing in Clostridium thermocellum and Methanococcus maripaludis.
635 Nucleic acids research. 2012;40(19):9887-96.
636 40. Elmore JR, Yokooji Y, Sato T, Olson S, Glover CV, 3rd, Graveley BR, et al. Programmable
637 plasmid interference by the CRISPR-Cas system in Thermococcus kodakarensis. RNA Biol.
638 2013;10(5):828-40.
639 41. Nickel L, Weidenbach K, Jager D, Backofen R, Lange SJ, Heidrich N, et al. Two CRISPR-
640 Cas systems in Methanosarcina mazei strain Go1 display common processing features despite
641 belonging to different types I and III. RNA Biol. 2013;10(5):779-91.
642 42. Garrett SC. Pruning and Tending Immune Memories: Spacer Dynamics in the CRISPR
643 Array. Front Microbiol. 2021;12:664299.
644 43. Stern A, Mick E, Tirosh I, Sagy O, Sorek R. CRISPR targeting reveals a reservoir of
645 common phages associated with the human gut microbiome. Genome research.
646 2012;22(10):1985-94.
647 44. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et
648 al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature.
649 2019;569(7758):655-62.
650 45. Podlevsky JD, Hudson CM, Timlin JA, Williams KP. CasCollect: targeted assembly of
651 CRISPR-associated operons from high-throughput sequencing data. NAR genomics and
652 bioinformatics. 2020;2(3):lqaa063.
653 46. Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: a versatile open source tool
654 for metagenomics. PeerJ. 2016;4:e2584.
655 47. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes De Novo
656 Assembler. Current protocols in bioinformatics. 2020;70(1):e102.
657 48. Grissa I, Vergnaud G, Pourcel C. The CRISPRdb database and tools to display CRISPRs
658 and to generate dictionaries of spacers and repeats. BMC Bioinformatics. 2007;8:172.
659 49. Makarova KS, Wolf YI, Iranzo J, Shmakov SA, Alkhnbashi OS, Brouns SJJ, et al.
660 Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. Nat
661 Rev Microbiol. 2020;18(2):67-83.
662 50. Russel J, Pinilla-Redondo R, Mayo-Munoz D, Shah SA, Sorensen SJ. CRISPRCasTyper:
663 Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. CRISPR J.
664 2020;3(6):462-9.
665 51. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation
666 sequencing data. Bioinformatics. 2012;28(23):3150-2.
667 52. Milicevic O, Repac J, Bozic B, Djordjevic M, Djordjevic M. A Simple Criterion for
668 Inferring CRISPR Array Direction. Frontiers in microbiology. 2019;10:2054.
669 53. Alkhnbashi OS, Mitrofanov A, Bonidia R, Raden M, Tran VD, Eggenhofer F, et al.
670 CRISPRloci: comprehensive and accurate annotation of CRISPR-Cas systems. Nucleic acids
671 research. 2021;49(W1):W125-W30.
672 54. Biswas A, Fineran PC, Brown CM. Accurate computational prediction of the transcribed
673 strand of CRISPR non-coding RNAs. Bioinformatics. 2014;30(13):1805-13.
674 55. Alstott J, Bullmore E, Plenz D. Powerlaw: a Python package for analysis of heavy-tailed
675 distributions. PLoS One. 2014;9(1):e85777.
676 56. Benler S, Yutin N, Antipov D, Rayko M, Shmakov S, Gussow AB, et al. Thousands of
677 previously unknown phages discovered in whole-community human gut metagenomes.
678 Microbiome. 2021;9(1):78.
679 57. Botelho J. Defense systems are pervasive across chromosomally integrated mobile
680 genetic elements and are inversely correlated to virulence and antimicrobial resistance.
681 Nucleic acids research. 2023;51(9):4385-97.
682 58. Lai S, Jia L, Subramanian B, Pan S, Zhang J, Dong Y, et al. mMGE: a database for human
683 metagenomic extrachromosomal mobile genetic elements. Nucleic acids research.
684 2021;49(D1):D783-D91.

59. Jiang X, Hall AB, Xavier RJ, Alm EJ. Comprehensive analysis of chromosomal mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. PLoS One. 2019;14(12):e0223680.
60. Wang M, Liu G, Liu M, Tai C, Deng Z, Song J, et al. ICEberg 3.0: functional categorization and analysis of the integrative and conjugative elements in bacteria. Nucleic acids research. 2024;52(D1):D732-D7.
61. Camargo AP, Call L, Roux S, Nayfach S, Huntemann M, Palaniappan K, et al. IMG/PR: a database of plasmids from genomes and metagenomes with rich annotations and metadata. Nucleic acids research. 2024;52(D1):D164-D73.
62. Shah SA, Deng L, Thorsen J, Pedersen AG, Dion MB, Castro-Mejia JL, et al. Expanding known viral diversity in the healthy infant gut. Nature microbiology. 2023;8(5):986-98.
63. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119.
64. Kunin V, Sorek R, Hugenholtz P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome biology. 2007;8(4):R61.
65. von Meijenfeldt FAB, Arkhipova K, Cambuy DD, Coutinho FH, Dutilh BE. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. Genome biology. 2019;20(1):217.
66. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094-100.
67. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010;20(9):1297-303.
68. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2.
69. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422-3.
70. Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences of the United States of America. 2008;105(4):1118-23.
71. Pavlova YS, Paez-Espino D, Morozov AY, Belalov IS. Searching for fat tails in CRISPR-Cas systems: Data analysis and mathematical modeling. PLoS computational biology. 2021;17(3):e1008841.
72. Toms A, Barrangou R. On the global CRISPR array behavior in class I systems. Biology direct. 2017;12(1):20.
73. Kostic AD, Xavier RJ, Gevers D. The microbiome in inflammatory bowel disease: current status and the future ahead. Gastroenterology. 2014;146(6):1489-99.
74. Pinilla-Redondo R, Mayo-Munoz D, Russel J, Garrett RA, Randau L, Sorensen SJ, et al. Type IV CRISPR-Cas systems are highly diverse and involved in competition between plasmids. Nucleic acids research. 2020;48(4):2000-12.
75. Shmakov SA, Sitnik V, Makarova KS, Wolf YI, Severinov KV, Koonin EV. The CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes. MBio. 2017;8(5).
76. Shmakov SA, Wolf YI, Savitskaya E, Severinov KV, Koonin EV. Mapping CRISPR spaceromes reveals vast host-specific viromes of prokaryotes. Commun Biol. 2020;3(1):321.
77. Soto-Perez P, Bisanz JE, Berry JD, Lam KN, Bondy-Denomy J, Turnbaugh PJ. CRISPR-Cas System of a Prevalent Human Gut Bacterium Reveals Hyper-targeting against Phages in a Human Virome Catalog. Cell Host Microbe. 2019;26(3):325-35 e5.
78. McGinn J, Marraffini LA. Molecular mechanisms of CRISPR-Cas spacer acquisition. Nat Rev Microbiol. 2019;17(1):7-12.

736 79. McGinn J, Marraffini LA. CRISPR-Cas Systems Optimize Their Immune Response by
737 Specifying the Site of Spacer Integration. Mol Cell. 2016;64(3):616-23.
738 80. Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in
739 staphylococci by targeting DNA. Science. 2008;322(5909):1843-5.
740 81. van Houte S, Ekroth AK, Broniewski JM, Chabas H, Ashby B, Bondy-Denomy J, et al. The
741 diversity-generating benefits of a prokaryotic adaptive immune system. Nature.
742 2016;532(7599):385-8.
743 82. Watson BNJ, Capria L, Alseth EO, Pons B, Biswas A, Lenzi L, et al. CRISPR-Cas in
744 Pseudomonas aeruginosa provides transient population-level immunity against high phage
745 exposures. The ISME journal. 2024.
746 83. Shkoporov AN, Clooney AG, Sutton TDS, Ryan FJ, Daly KM, Nolan JA, et al. The Human
747 Gut Virome Is Highly Diverse, Stable, and Individual Specific. Cell Host Microbe.
748 2019;26(4):527-41 e5.
749 84. Weinberger AD, Sun CL, Plucinski MM, Denef VJ, Thomas BC, Horvath P, et al.
750 Persisting viral sequences shape microbial CRISPR-based immunity. PLoS computational
751 biology. 2012;8(4):e1002475.
752 85. Manrique P, Bolduc B, Walk ST, van der Oost J, de Vos WM, Young MJ. Healthy human
753 gut phageome. Proceedings of the National Academy of Sciences of the United States of
754 America. 2016;113(37):10400-5.
755 86. Koonin EV, Dolja VV, Krupovic M. The healthy human virome: from virus-host
756 symbiosis to disease. Curr Opin Virol. 2021;47:86-94.
757 87. Smith L, Goldobina E, Govi B, Shkoporov AN. Bacteriophages of the Order Crassvirales:
758 What Do We Currently Know about This Keystone Component of the Human Gut Virome?
759 Biomolecules. 2023;13(4).
760 88. Sutcliffe SG, Reyes A, Maurice CF. Bacteriophages playing nice: Lysogenic
761 bacteriophage replication stable in the human gut microbiota. iScience. 2023;26(2):106007.
762 89. Dion MB, Shah SA, Deng L, Thorsen J, Stokholm J, Krogfelt KA, et al. Escherichia coli
763 CRISPR arrays from early life fecal samples preferentially target prophages. The ISME journal.
764 2024.
765 90. Wu J, Zhang H, Gan R, Xia Y, Zhang F, Wang D, et al. CRISPR dynamics during the
766 interaction between bacteria and phage in the first year of life. Microb Genom. 2023;9(7).
767 91. Redgwell TA, Thorsen J, Petit M-A, Deng L, Vestergaard GA, Russel J, et al. Prophages in
768 the infant gut are largely induced, and may be functionally relevant to their hosts. bioRxiv.
769 2021:2021.06.25.449885.
770 92. Sternberg SH, Richter H, Charpentier E, Qimron U. Adaptation in CRISPR-Cas Systems.
771 Mol Cell. 2016;61(6):797-808.
772 93. Nussenzweig PM, Marraffini LA. Molecular Mechanisms of CRISPR-Cas Immunity in
773 Bacteria. Annu Rev Genet. 2020;54:93-120.
774 94. Nicholson TJ, Jackson SA, Croft BI, Staals RHJ, Fineran PC, Brown CM. Bioinformatic
775 evidence of widespread priming in type I and II CRISPR-Cas systems. RNA Biol. 2019;16(4):566-
776 76.
777 95. Datsenko KA, Pougach K, Tikhonov A, Wanner BL, Severinov K, Semenova E. Molecular
778 memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system.
779 Nature communications. 2012;3:945.
780 96. Deecker SR, Ensminger AW. Type I-F CRISPR-Cas Distribution and Array Dynamics in
781 Legionella pneumophila. G3. 2020;10(3):1039-50.
782 97. Haerter JO, Sneppen K. Spatial structure and Lamarckian adaptation explain extreme
783 genetic diversity at CRISPR locus. mBio. 2012;3(4):e00126-12.
784 98. Koonin EV, Wolf YI. Evolution of the CRISPR-Cas adaptive immunity systems in
785 prokaryotes: models and observations on virus-host coevolution. Mol Biosyst. 2015;11(1):20-7.

786 99. Weissman JL, Holmes R, Barrangou R, Moineau S, Fagan WF, Levin B, et al. Immune
787 loss as a driver of coexistence during host-phage coevolution. The ISME journal.
788 2018;12(2):585-97.
789 100. Wheatley RM, MacLean RC. CRISPR-Cas systems restrict horizontal gene transfer in
790 Pseudomonas aeruginosa. The ISME journal. 2021;15(5):1420-33.
791 101. Bowers RM, Nayfach S, Schulz F, Jungbluth SP, Ruhl IA, Sheremet A, et al. Dissecting
792 the dominant hot spring microbial populations based on community-wide sampling at single-
793 cell genomic resolution. The ISME journal. 2022;16(5):1337-47.

794

795

796 **Figure legends**

797 **Figure 1:** Assembly and quality-control of CRISPR arrays in the human gut microbiome. (a)
798 Schematics of the study design based on metagenomic data from the HMP2-IBD cohort.
799 CRISPR arrays were assembled de novo with CasCollect. (b) Number of CRISPR arrays per
800 sample. (c) Distribution of the number of CRISPR arrays per sample in patients with Crohn's
801 disease (CD), ulcerative colitis (UC), and participants with no inflammatory bowel disease
802 (nonIBD). The central lines indicate the medians for each group. Kruskal-Wallis omnibus test
803 for differences among groups ($\chi^2 = 148$; $p < 10^{-10}$). Mann-Whitney tests for differences between
804 nonIBD and disease ($Z = 10.4$; $p < 10^{-10}$); and between normal and dysbiosis ($Z = 6.76$; $p < 10^{-10}$).
805 (d) Distribution of the number of spacers per CRISPR array, in all arrays (grey) and arrays
806 located in the vicinity of Cas genes (blue). The dashed line indicates the best fit to a truncated
807 power-law. (e) Distribution of the number of CRISPR arrays per lineage (note the logarithmic
808 scale in the y-axis, also in (d)). (f) Quantification of the number of pairs of CRISPR arrays from
809 the same lineage that show inconsistencies in spacer order, spacer orientation, and array
810 orientation.

811 **Figure 2:** Distribution of CRISPR subtypes among taxa and association with dysbiosis. The area
812 of each circle is proportional to the number of CRISPR lineages in the whole dataset. Colors
813 indicate the relative association with dysbiosis. Subtypes with <50 CRISPR lineages are not
814 represented. "Other/undetermined" includes genera (or phyla) with <50 CRISPR lineages and
815 lineages with ambiguous classification at the genus (or phylum) level. CRISPR subtypes on the
816 x-axis are listed from left to right based on their total abundance.

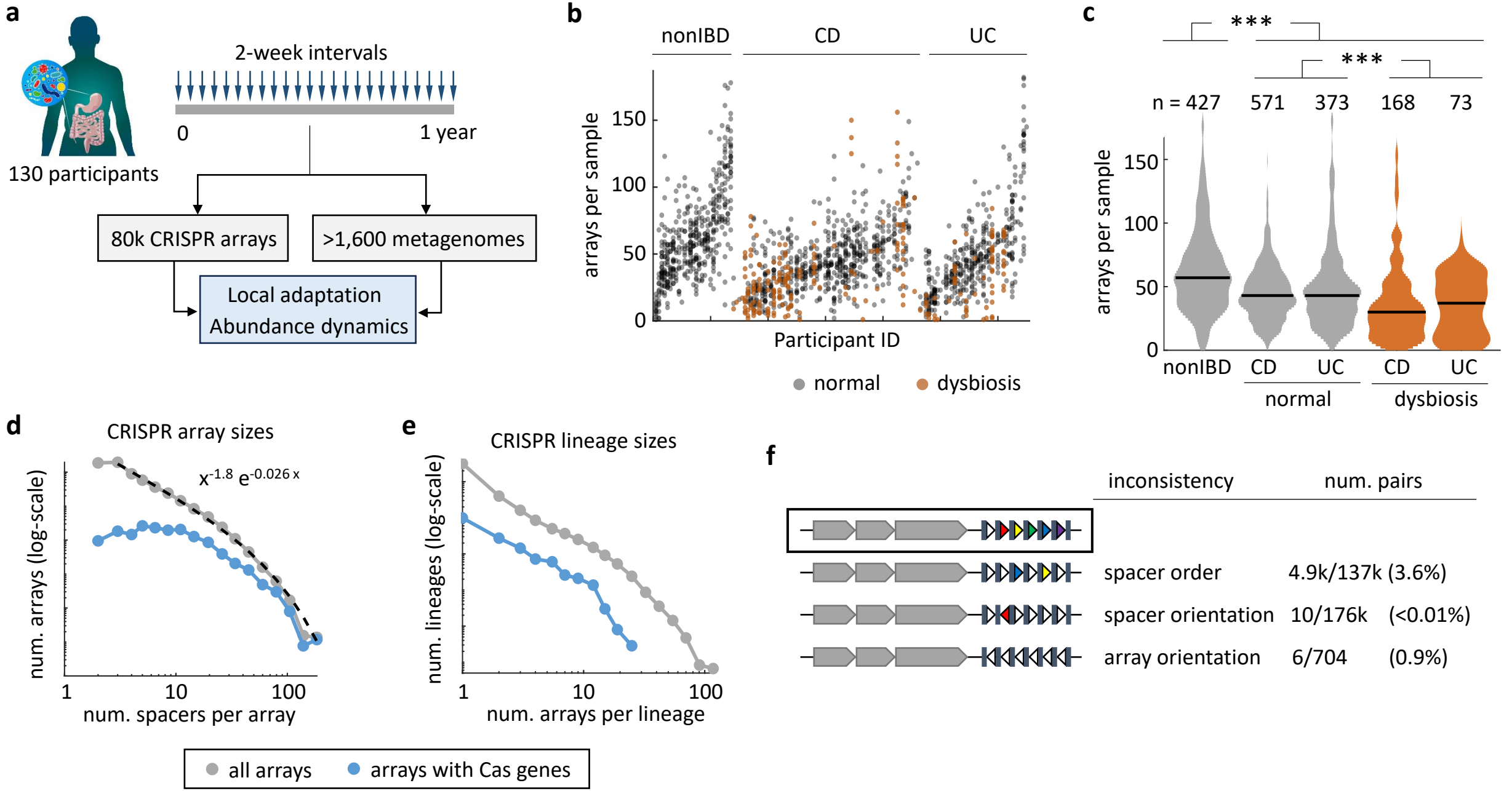
817 **Figure 3:** Targets of CRISPR-Cas immunity in the gut microbiome. (a) Relative representation of
818 different classes of mobile genetic elements among the targets of CRISPR-Cas systems. Left pie
819 chart: hits to a merged database of known mobile genetic elements (MGE db). Right:
820 protospacers from the local microbiome, defined as the combination of all samples obtained

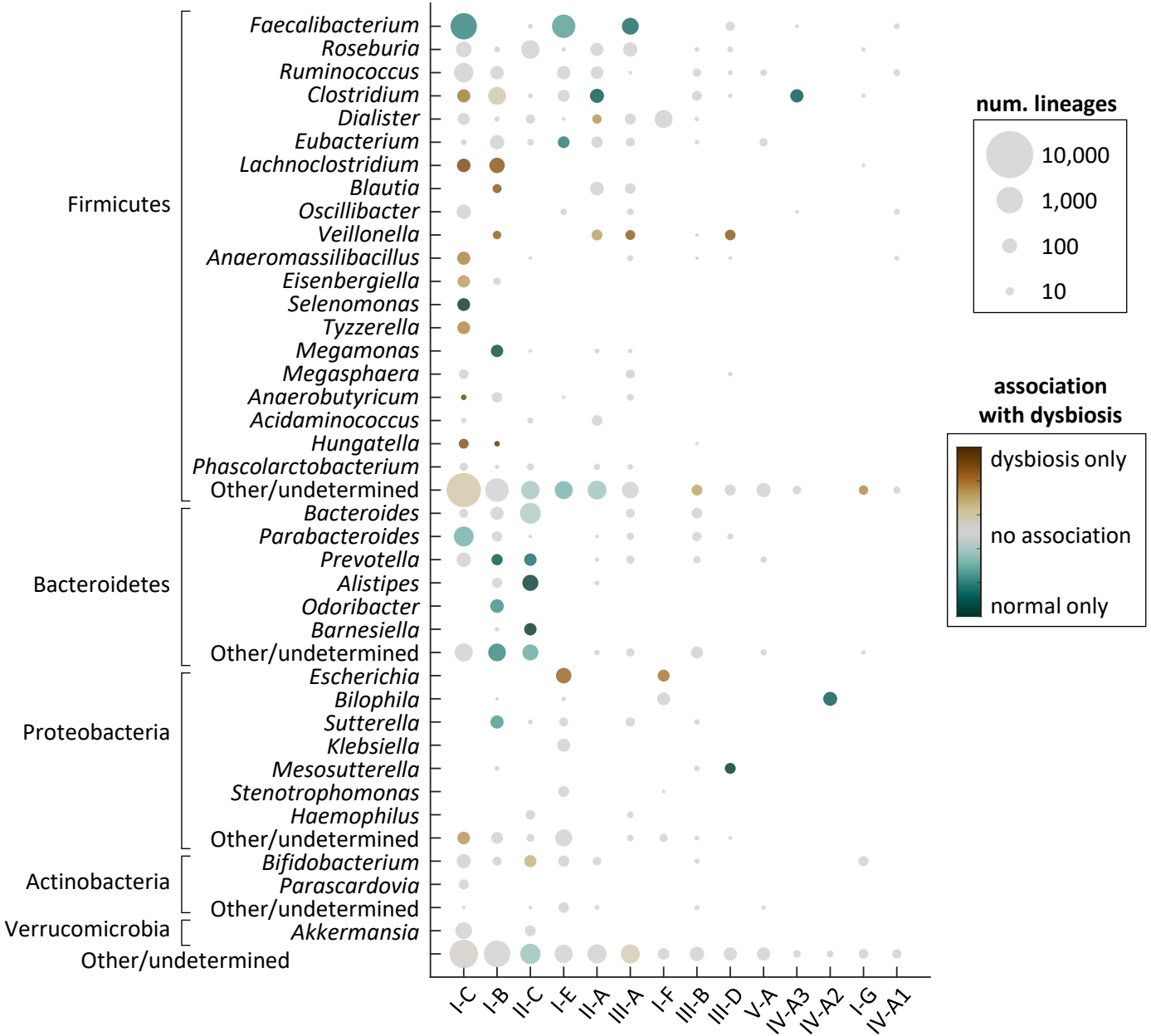
821 from the same participant. The segment for lytic viruses also includes chronic (non-lysogenic)
822 viruses (b) Distribution and functional profile of CRISPR targets across CRISPR-Cas subtypes.
823 Locally adapted spacers (bar plot on the left) are those that target sequences from the local
824 microbiome. (c) Relative position of locally adapted spacers within the CRISPR array. L: leading
825 end; T: trailing end. (d-f) Fraction of spacers targeting lysogenic (d), crass-like (e), and infant
826 gut (f) phages (with respect to all spacers that target phages) in the leading, middle, and distal
827 sections of the array. Whiskers represent 95% confidence intervals calculated with the
828 Clopper-Pearson method. In (b) and (c), only those subtypes with >100 locally adapted spacers
829 are shown. Panels (d-f), only include subtypes with >100 spacers targeting lysogenic phages
830 (d), crass-like phages (e), and infant gut phages (f) within high-quality, locally adapted arrays
831 (oriented and with ≥5 spacers). Despite the high fraction of locally adapted spacers, subtype II-
832 C was excluded from (c-f) due to the unavailability of accurate predictions for array
833 orientation.

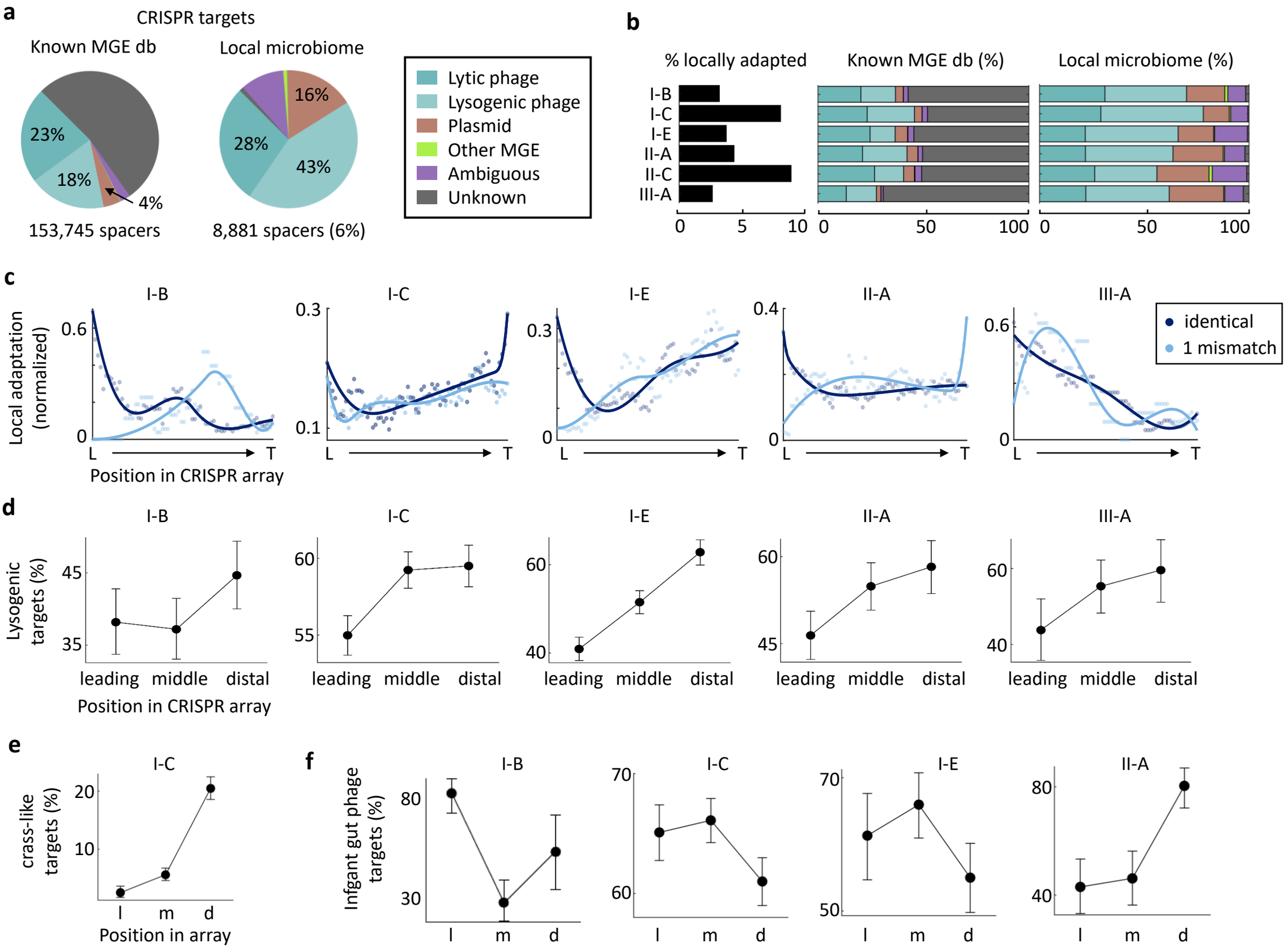
834 **Figure 4:** Effect of CRISPR-Cas immunity on MGE abundance and host diversity (see also Figure
835 S5). (a) Median normalized abundance of targeted MGE (plasmids, lytic/non-lysogenic phages,
836 and lysogenic phages) before the spacer was first observed in the host population ("Before"),
837 in the sample in which the spacer was observed for the first time ("Acquisition"), in
838 subsequent samples in which the spacer was present ("After – With spacer"), and in
839 subsequent samples that contained the same CRISPR lineage but not the spacer ("After – No
840 spacer"). Error bars indicate the 95% confidence intervals for the median, based on 2,000
841 bootstraps. All comparisons shown above the plot are statistically significant with $p < 10^{-4}$
842 (Mann-Whitney test). Note that comparisons are conservative with respect to array
843 incompleteness, because apparent loss of spacers in incomplete arrays would lead to
844 misclassification of samples (from "After – With spacer" to "After – No spacer"), therefore
845 reducing the observed differences. (b) Prevalence of targeted MGE, calculated as the fraction
846 of samples in which the target is detected. Error bars indicate the 95% confidence intervals

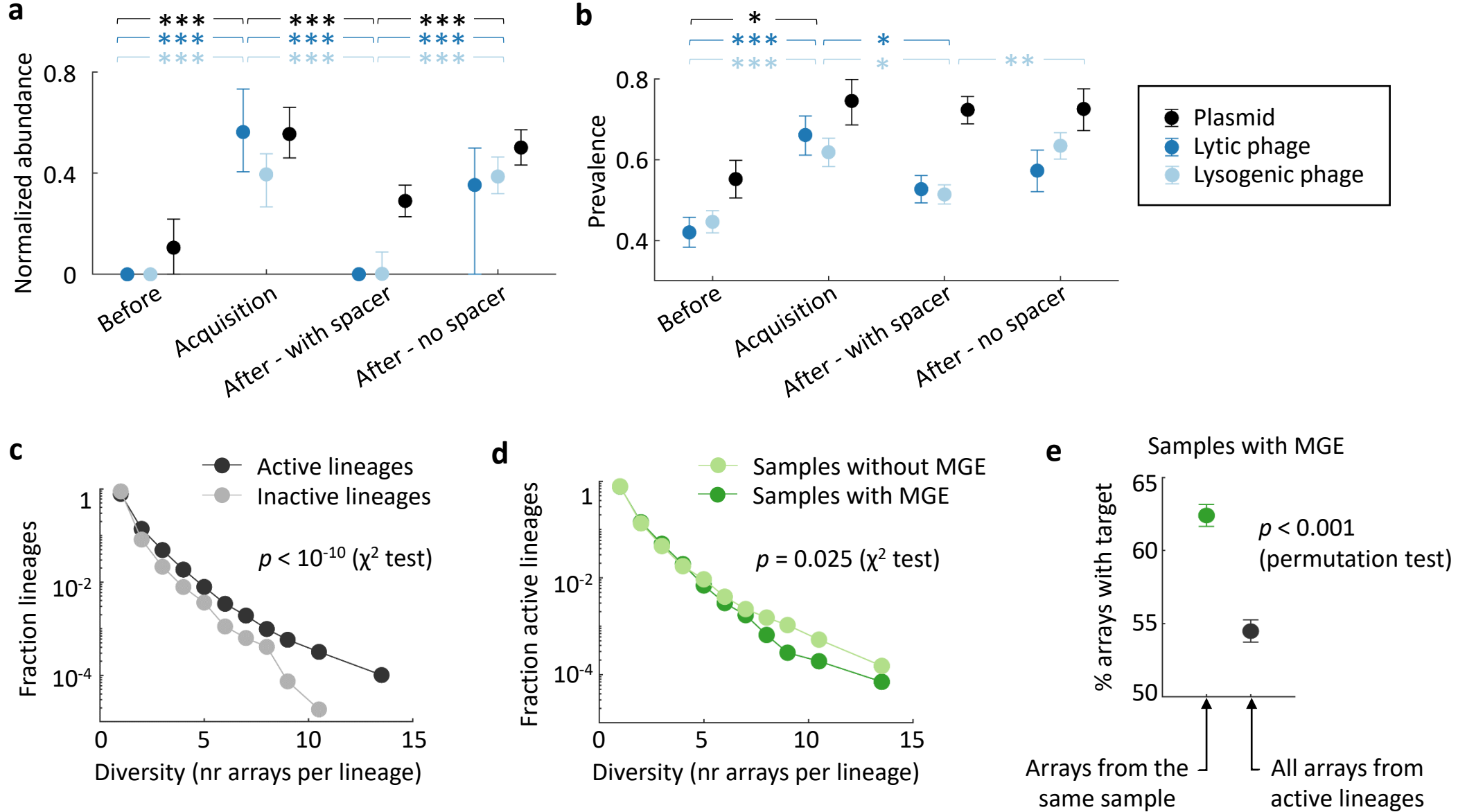
847 (Clopper-Pearson method). Statistical significance based on Fisher's exact test: $p < 0.05$ (*), $p <$
848 0.01 (**), $p < 0.001$ (***). (c) Diversity distribution (number of distinct arrays) of CRISPR
849 lineages that target at least one MGE from the local microbiome ("active" lineages) and those
850 that do not show any sign of local adaptation ("inactive" lineages). (d) Number of arrays per
851 lineage in the presence and absence of MGE. For a given active lineage, samples with MGE are
852 those that contain at least one MGE targeted by at least one spacer from that lineage. (e)
853 Percentage of arrays per lineage that target MGE from the same sample, compared to the
854 expected value for randomly chosen arrays. The randomized set only includes active lineages
855 and samples with MGE targeted by those lineages. Error bars indicate the 95% confidence
856 intervals based on 2,000 bootstraps.

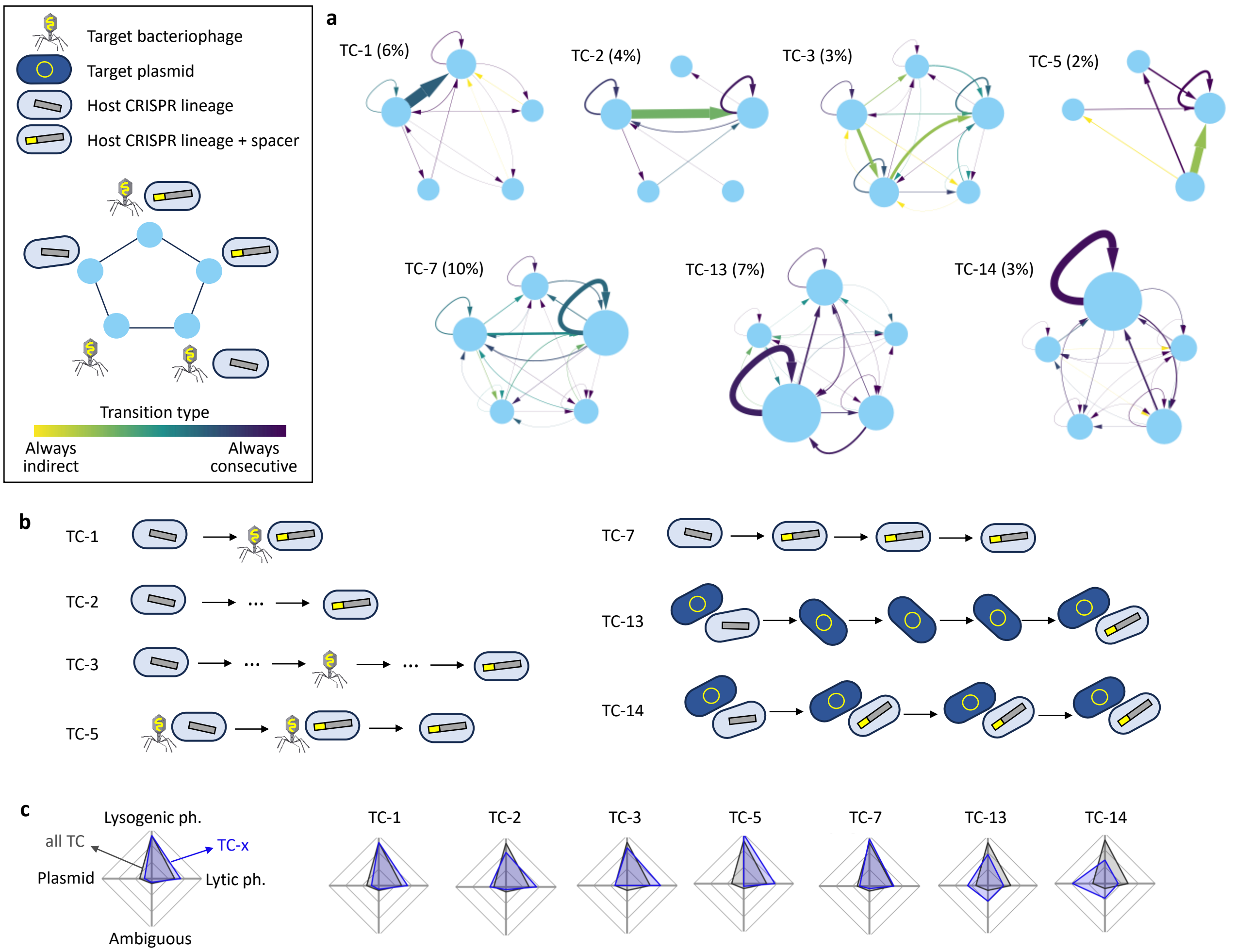
857 **Figure 5:** Temporal dynamics of MGE and CRISPR arrays. (a) Network representation of 7 of the
858 most frequent trajectory clusters (TC) representing potential immune adaptation in the host
859 population. The percentages in parentheses indicate the relative frequency of each trajectory.
860 Each node represents a state, defined by the presence or absence of a spacer, its cognate
861 MGE, and the CRISPR lineage in which the spacer was acquired (legend in the upper left box).
862 Node sizes are proportional to the number of times that a state is observed in a trajectory.
863 Arrow widths represent transition probabilities between states and the color scale indicates
864 the fraction of direct vs indirect transitions (direct transitions correspond to two consecutive
865 samples, indirect transitions involve intermediate samples in which neither the MGE nor the
866 CRISPR lineage were observed). (b) Simplified representation of typical trajectories associated
867 with the TC in (a). (c) Relative association of lytic (or non-lysogenic) phages, lysogenic phages,
868 plasmids, and ambiguous MGE with each TC.


a
2-week intervals
0
1 year
130 participants
80k CRISPR arrays
>1,600 metagenomes
Local adaptation
Abundance dynamics
b
nonIBD
CD
UC
arrays per sample
150
100
50
0
Participant ID
normal
dysbiosis
c
***
***
n = 427
571
373
168
73
arrays per sample
150
100
50
0
nonIBD
CD
UC
CD
UC
normal
dysbiosis
d
CRISPR array sizes
$x^{-1.8}\ e^{-0.026\,x}$
num. arrays (log-scale)
1
10
100
num. spacers per array
e
CRISPR lineage sizes
num. lineages (log-scale)
1
10
100
num. arrays per lineage
all arrays
arrays with Cas genes
f
inconsistency
num. pairs
spacer order
4.9k/137k (3.6%)
spacer orientation
10/176k
(<0.01%)
array orientation
6/704
(0.9%)


Firmicutes
Faecalibacterium
Roseburia
Ruminococcus
Clostridium
Dialister
Eubacterium
Lachnoclostridium
Blautia
Oscillibacter
Veillonella
Anaeromassilibacillus
Eisenbergiella
Selenomonas
Tyzzerella
Megamonas
Megasphaera
Anaerobutyricum
Acidaminococcus
Hungatella
Phascolarctobacterium
Other/undetermined
Bacteroidetes
Bacteroides
Parabacteroides
Prevotella
Alistipes
Odoribacter
Barnesiella
Other/undetermined
Proteobacteria
Escherichia
Bilophila
Sutterella
Klebsiella
Mesosutterella
Stenotrophomonas
Haemophilus
Other/undetermined
Actinobacteria
Bifidobacterium
Parascardovia
Other/undetermined
Verrucomicrobia
Akkermansia
Other/undetermined
I-C
I-B
II-C
I-E
II-A
III-A
I-F
III-B
III-D
V-A
IV-A3
IV-A2
I-G
IV-A1
num. lineages
10,000
1,000
100
10
association with dysbiosis
dysbiosis only
no association
normal only


a
CRISPR targets
Known MGE db
Local microbiome
23%
18%
4%
16%
28%
43%
153,745 spacers
8,881 spacers (6%)
Lytic phage
Lysogenic phage
Plasmid
Other MGE
Ambiguous
Unknown
b
% locally adapted
Known MGE db (%)
Local microbiome (%)
I-B
I-C
I-E
II-A
II-C
III-A
0
5
10
0
50
100
0
50
100
c
I-B
I-C
I-E
II-A
III-A
Local adaptation (normalized)
0.6
0
0.3
0.1
0.3
0
0.4
0
0.6
0
L
T
Position in CRISPR array
identical
1 mismatch
d
I-B
I-C
I-E
II-A
III-A
Lysogenic targets (%)
45
35
60
55
60
40
60
45
60
40
leading
middle
distal
Position in CRISPR array
e
I-C
crass-like targets (%)
20
10
l
m
d
Position in array
f
I-B
I-C
I-E
II-A
Infgant gut phage targets (%)
80
30
70
60
70
50
80
40
l
m
d


a
Normalized abundance
0.8
0.4
0
Before
Acquisition
After - with spacer
After - no spacer
b
Prevalence
0.8
0.6
0.4
Before
Acquisition
After - with spacer
After - no spacer
Plasmid
Lytic phage
Lysogenic phage
c
Active lineages
Inactive lineages
$p < 10^{-10}$ (χ^2 test)
Fraction lineages
1
10^{-2}
10^{-4}
0
5
10
15
Diversity (nr arrays per lineage)
d
Samples without MGE
Samples with MGE
$p = 0.025$ (χ^2 test)
Fraction active lineages
1
10^{-2}
10^{-4}
0
5
10
15
Diversity (nr arrays per lineage)
e
Samples with MGE
% arrays with target
65
60
55
50
$p < 0.001$
(permutation test)
Arrays from the same sample
All arrays from active lineages


Target bacteriophage
Target plasmid
Host CRISPR lineage
Host CRISPR lineage + spacer
Transition type
Always indirect
Always consecutive
a
TC-1 (6%)
TC-2 (4%)
TC-3 (3%)
TC-5 (2%)
TC-7 (10%)
TC-13 (7%)
TC-14 (3%)
b
TC-1
TC-2
TC-3
TC-5
TC-7
TC-13
TC-14
c
Lysogenic ph.
all TC
TC-x
Plasmid
Lytic ph.
Ambiguous
TC-1
TC-2
TC-3
TC-5
TC-7
TC-13
TC-14