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 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…

34 **Introduction**

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

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

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

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

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

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

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

94 **Methods**

95 Metagenomic data

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

102 Assembly, annotation, and clustering of CRISPR arrays

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

109 additional repeats from CRISPRdb (48) and Makarova et al. (49) clustered at 90% identity and
110 coverage, and a set of previously unreported repeats obtained by running CRISPRCasTyper
111 v1.8 (50) on the publicly available, preassembled contigs from the HMP2-IBD study. CasCollect
112 was run with options --meta --noannotate --nucl. Putative CRISPR-Cas loci assembled by
113 CasCollect were quality-filtered and annotated with CRISPRCasTyper v1.8, obtaining 79,475
114 high-confidence CRISPR arrays of which 4,282 (representing around 5% of the total arrays)
115 included a Cas operon in their neighborhood. Spacers and repeats from high-confidence
116 CRISPR arrays were dereplicated using CD-Hit-Est v4.8.1 (51) with a 100% identity and
117 coverage threshold and option *-g* 1 (slow and accurate mode), resulting in a total of 11,373
118 different repeats and 153,745 different spacers. The directionality of transcription in a subset
119 of 4,054 arrays with adjacent Cas genes and >5 spacers was predicted with Cas Orientation
120 (52), CRISPRLoci v1.0 (53), and CRISPRDirection v2.0 (54). Among these three prediction tools,
121 Cas Orientation showed the best agreement with the expected drop in transcription for middle
122 and distal spacers and was applicable to a larger fraction of arrays (Suppl. Fig. S1). Therefore,
123 we chose Cas Orientation as the default method to present the results in the text and main
124 figures. To explore the distribution of the number of spacers per array, we used the 'powerlaw'
125 python package (55).

126 A strict nesting criterion was applied to identify groups (or lineages) of highly similar CRISPR
127 arrays present within and across samples of the same individual. A CRISPR array was
128 considered nested within another CRISPR array if both shared the same or highly similar
129 repeats (>90% identity and coverage, clustered using CD-HIT-Est with option *-g* 1) and all the
130 spacers of the former were also harbored by the latter. Group-representative arrays were
131 identified as those that were not nested within any other array. Finally, clusters of CRISPR
132 arrays were built by considering all the arrays nested within the same group representative.
133 This algorithm allowed us to retrieve clusters of highly similar CRISPR arrays while accounting

134 for incomplete assemblies and spacer gain and loss over time. Because of that, we refer to
135 these clusters as CRISPR lineages or lineages of arrays.

136 Search for CRISPR targets in the local microbiome

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

143 One spacer in the dataset produced an extremely high number of hits to the local
144 metagenome (16,327 hits, while the second largest number was 121). Closer inspection
145 revealed that the spacer was highly similar to a transposase sequence found in bacterial
146 chromosomes and prophages. To avoid biases in downstream analyses, the spacer was
147 removed from the dataset.

148 Identification of spacer targets

149 The targets of CRISPR of spacers were inferred by performing a nucleotide BLAST search of all
150 spacers against a merged database of mobile genetic elements (MGEdb), setting BLAST
151 parameters for short queries (-dust no -word_size 8) and keeping all hits with >95% identity
152 and coverage. The merged MGEdb included 198,607 complete viral genomes from the NCBI
153 Genome database (downloaded on 27/01/2021 from ftp.ncbi.nlm.nih.gov/genomes/Viruses/)
154 and the NCBI Virus database (downloaded in interactive mode on 27/01/2021 from
155 https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&Completeness_
156 s=complete), a set of 1,887 gut phages recently identified by Benler et al. (56), a set of 13,274
157 ICE/IME from Botelho (57), and all sequences from mMGE v1.0 (58) (494,835 phages and
158 92,493 plasmids), ImmeDb (59) (201 phages and 4,727 ICE/IME, transposable elements, and

159 other unclassified MGE), ICEberg3 (60) (4,461 ICE/IME and other chromosomally integrated
160 elements), and IMG/PR (61) (699,973 plasmids). Based on the BLAST results, we classified
161 spacers according to their most likely target (virus, plasmid, or other mobile genetic elements).
162 Spacers with inconsistent target predictions were classified as "ambiguous". Spacer-target
163 assignments were annotated according to the accuracy of the match as identical (100%
164 identity and coverage) or highly similar (identity and coverage above 95% but below 100%).
165 Moreover, viral targets were taxonomically annotated according to the information in MGEdb.

166 To infer the target of spacers that did not match any sequence in MGEdb but had protospacers
167 in the local microbiome, we extracted 15 kb neighborhoods (7.5 kb upstream and downstream
168 from the protospacer) from the HMP2-IBD contigs and ran a nucleotide BLAST search against
169 MGEdb (default options) setting an e-value threshold of 10^{-8}. For each spacer, we predicted
170 the target based on all hits with scores greater than 0.9 times the highest score. Spacer whose
171 target remained unclassified after this second step were labeled as "unknown".

172 Prediction of lytic or lysogenic phage lifestyle

173 Phages targeted by CRISPR spacers were classified as lytic or lysogenic based on the presence
174 of integrases in phage genomes. To that end, we extracted 49,645 phage genomes from
175 MGEdb and 110,712 metagenome-assembled contigs from HMP2-IBD that contained
176 protospacers. Then, we ran Prodigal v2.6.3 (62) with the option *-p meta* to identify open
177 reading frames and searched for hits to viral integrases using the *hmmersearch* command of
178 the HMMER v3.3.2 suite. We included the following PFAM profiles for integrases: PF00239,
179 PF00589, PF02899, PF07508, PF12482, PF12834, PF13009, PF13102, PF13495, PF14659,
180 PF16795, and PF18644. The search returned 11,631 viral sequences with putative integrases
181 (e-value $<10^{-8}$) that also contained protospacers. The corresponding spacers (27,868 in total)
182 were regarded as targeting lysogenic phages.

183 Prevalence and abundance of CRISPR targets in the local microbiome

184 To study the quantitative dynamics of CRISPR targets in the microbiome, we focused on
185 spacer-protospacer pairs present in samples from the same individual (possibly at different
186 times) that fulfil the following conditions: (1) the first appearance of the spacer occurs later
187 than the first appearance of the CRISPR lineage with which the spacer is associated; and (2)
188 spacers are present in at least 2 samples. These two conditions allowed us to compare target
189 abundances before and after the first appearance of the spacer, while controlling for the
190 intermittent prevalence of hosts (using CRISPR lineages as proxies for the latter). For the same
191 reason, we excluded from these analyses all samples in which the CRISPR lineage was not
192 detected. After applying these filters, 1540 spacer-protospacer pairs remained, all of them
193 involving phages as targets.

194 To estimate the prevalence of targets, we counted the number of samples that contain the
195 protospacer before, at, and after the first appearance of the spacer, pooled these counts for all
196 spacer-protospacer pairs, and normalized them by the number of samples in each group.
197 Confidence intervals were calculated using the Clopper-Pearson method as implemented in
198 the *statsmodels* Python library. To quantify target abundances, we extracted 4kb
199 neighborhoods (2kb upstream and downstream) of protospacers from HMP2-IBD contigs and
200 mapped metagenomic reads to those neighborhoods with CoverM v0.6.1. Sample-wise
201 relative abundances were calculated as the number of mapped reads divided by the length of
202 the mapped sequence (generally 4kb, except for protospacers located near contig ends) and by
203 the total number of reads in the sample. The sample-wise abundance in samples in which the
204 target was not detected was set to 0. Finally, to facilitate the comparison of trends across
205 targets, we normalized the sample-wise abundances of each target by the maximum
206 abundance observed for that target in the whole time series.

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

208 To assign a taxonomic origin to CRISPR arrays, we leveraged the fact that CRISPR repeats are
209 taxon-specific (although the taxonomic level of such specificity varies among repeats) (63).
210 First, we ran CRISPRCasTyper on the preassembled HMP2-IBD metagenomic contigs to predict
211 CRISPR arrays. Then, we used CD-HIT to identify contigs from HMP2-IBD that contain CRISPR
212 arrays with the same repeats (90% identity and coverage) as those assembled by CasCollect.
213 Those contigs were provided as inputs to the program CAT/BAT v5.2.3 (64) for taxonomical
214 assignment at the lowest non-ambiguous available level. For CRISPR arrays whose repeats
215 could not be mapped to any HMP2-IBD contig, we directly ran CAT/BAT on the contigs
216 obtained with CasCollect. (Note, however, that the level of resolution obtained with CasCollect
217 contigs was generally lower due to their shorter length.) Finally, all CRISPR arrays that
218 remained unassigned and were part of a lineage with one or more assigned members were
219 given the same taxonomical origin as the other members of the lineage, as long as no
220 inconsistencies were found.

221 Spacer location bias and asymmetry

222 We studied the relative position of spacers that target phages from the local metagenome
223 (henceforth "locally adapted spacers") in a subset of 4,054 oriented arrays with >5 spacers,
224 separately considering identical and highly-similar hits (CRISPR-Cas subtypes with <25 valid
225 arrays were not included in this analysis). To avoid artifacts associated with finite-size effects in
226 small arrays, we used a sliding-window approach to estimate the density of active spacers
227 along the array. We set a window width equal to 12% of the total length of the array
228 (equivalent to 2/3 of the inverse of the median number of spacers) and a step equal to 1.5% of
229 the total length of the array (1/8 of the window width).

230 Quantification of transcription levels along CRISPR arrays

231 For 753 samples from HMP2-IBD with available metatranscriptomes, we mapped
232 metatranscriptomic reads to CRISPR assemblies obtained from the same samples with

233 minimap2 v2.24, setting the option '-x sr' for short reads (65). To improve specificity towards
234 spacers in CRISPR arrays, we also included contigs with protospacers in the reference genomic
235 dataset and removed all mapped reads with aligned lengths below 70 nucleotides using the
236 command FilterSamReads from gatk 4.1.2.0 (66). This 70-nucleotide threshold was chosen to
237 ensure that alignments covered the equivalent of one spacer and one repeat (the average
238 spacer and repeat lengths in our dataset are 34 and 31, respectively). The coverage per
239 position was calculated with the command genomeCoverageBed from BEDTools 2.28.0 (67).

240 Clustering and classification of CRISPR-spacer-target dynamics

241 To study the temporal dynamics of CRISPR lineages, spacers, and targets, we first identified all
242 the spacers that were observed for the first time later than their corresponding CRISPR
243 lineage. Then, we codified the time series associated with those spacers as strings of
244 characters, with letters A-F indicating all possible combinations for the presence or absence of
245 the spacer, its target, and its corresponding CRISPR lineage. "Empty" states (that is, those
246 characterized by the absence of spacer, CRISPR lineage, and target) were removed from both
247 ends of the string and collapsed to a single instance if appearing more than once between non-
248 empty states. We performed pairwise alignments of all strings with the Biopython command
249 pairwise2.align.globalms (68) using the following scores: 2 per match, -4 per mismatch, -4 per
250 gap opening, and -2 per gap extension. After normalizing the scores by the length of the
251 longest member of each pair and discarding pairs with negative scores, we built a time-series
252 similarity network using normalized scores are weights. Time series were then clustered with
253 Infomap v0.21.0 (69), with options -2 -u -N 200 (two-level partition assuming undirected links
254 and selecting the best of 200 trials).

255

256 **Results and discussion**

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

258 We combined state-of-the-art methods to assemble CRISPR arrays from the gut microbiomes
259 of 130 individuals, recurrently sampled every 2 weeks over a 1-year period as part of the
260 integrative Human Microbiome Project (Figure 1a). After curation and removal of low-quality
261 arrays, we obtained a set of 79,475 CRISPR arrays, encompassing 11,373 unique repeats and
262 153,745 unique spacers. Of those, around 4,300 arrays were adjacent to a Cas operon, what
263 allowed us to predict the direction of transcription by applying three alternative criteria (see
264 Methods). We identified CRISPR arrays in 36 out of the 40 most abundant bacterial genera in
265 all gut samples, which collectively account for 94% of bacterial abundance (Suppl. Table S1).
266 The number of unique CRISPR arrays per sample broadly varies through time and across
267 microbiomes, with significantly lower numbers in microbiomes obtained from patients with
268 Crohn's disease and ulcerative colitis (Figure 1b-c). Such lower numbers are consistent with the
269 reduced microbial diversity associated with dysbiosis in inflammatory bowel diseases (44).

270 The CRISPR array sizes follow a heavy-tailed distribution that is well-fit by a truncated power
271 law with exponent close to -2 and a cutoff at around 100 spacers (Figure 1d). Similar
272 distributions, although with higher cutoffs, have been reported for CRISPR arrays from
273 complete genomes (70, 71). Thus, the high heterogeneity of array lengths observed in this
274 dataset likely reflects a true property of CRISPR systems in the gut microbiome, rather than an
275 excess of fragmentary or poorly-assembled arrays.

276 We applied a strict nesting criterion to group CRISPR arrays into 41,742 clusters, of which
277 9,870 contain >1 array (Figure 1e). These clusters represent lineages of arrays from the same
278 microbiome (possibly sampled at different times) that are likely related by sequential gain and
279 loss of spacers. Array lineages serve a triple purpose in this study. First, lineages are pivotal to
280 assess the consistency of CRISPR assemblies by comparing spacer order and orientation in
281 arrays from the same lineage. Second, lineages permit a robust characterization of CRISPR
282 diversity by minimizing the impact of incomplete arrays and abundance biases. Third, lineages

283 constitute a convenient proxy for the presence of subpopulations of hosts in the microbial
284 community.

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

294 The most abundant CRISPR-Cas systems in the human gut microbiome belong to subtypes I-B,
295 I-C, I-E, and II-C (Figure 2). Subtypes I-B, I-C, and II-C are most often associated with Firmicutes
296 and Bacteroidetes, whereas subtype I-E is observed in Firmicutes and Proteobacteria.
297 Alterations in the composition of the microbiome significantly affect the profile of CRISPR
298 subtypes in Crohn's disease and ulcerative colitis (Figure 2 and Suppl. Table S1). In agreement
299 with previous studies (44, 72), dysbiosis leads to a decrease in the relative proportion of
300 subtypes I-B, I-C, I-E, and II-C that can be attributed to the depletion of *Faecalibacterium*
301 *prausnitzii* (I-C and I-E) and several groups of Bacteroidetes (I-B, I-C, and II-C). Among the less
302 abundant subtypes, IV-A (typically carried by large conjugative plasmids (73)) stands out for its
303 strong association with normal (that is, no dysbiotic) samples.

304 Selection promotes long-term immunity against endemic MGE

305 To identify the potential targets of CRISPR immunity, we mapped the spacers in CRISPR arrays
306 to a large collection of phages, plasmids, and other MGE obtained from state-of-the-art
307 (meta)genomic databases (see Methods). Overall, 41% of the spacers show high similarity

308 (zero or one mismatch) to known phages and only 4% map to plasmids and other MGE (Figure
309 3a). These values are notably higher than those reported by global surveys of CRISPR spacers
310 from complete prokaryotic genomes (74, 75) but close to those obtained for gut bacteria (43,
311 76), probably reflecting the higher coverage of MGE from the human gut in public databases.
312 Around 6% of the spacers target sequences from the local microbiome (defined as the union of
313 all samples collected from the same individual). We refer to these spacers as "locally adapted"
314 because they can provide immunity against MGE that are locally observed in the 1-year period
315 covered by the dataset. The most frequent targets of locally adapted spacers are lysogenic
316 phages (43%), followed by lytic and other non-lysogenic phages (28%), plasmids (16%), and
317 other MGE (1%). Around 10% of the spacers (labeled as "ambiguous" in Figure 3a-b) match
318 sequences found in multiple classes of MGE. All in all, we could assign a unique or ambiguous
319 origin to 99% of all locally adapted spacers. Therefore, the disproportionate number of hits to
320 phages and plasmids compared to other MGE is probably genuine, despite the admittedly
321 biased representation of different classes of MGE in public databases. Comparing among
322 CRISPR subtypes (Figure 3b), local adaptation is more frequent in subtypes I-C and II-C (7.7-
323 8.5% of spacers are locally adapted, compared to 2.5-4.2% in other subtypes; χ^2 test $p < 10^{-10}$).
324 Moreover, subtypes II-A, II-C, and III-A target a higher proportion of plasmids than class I
325 systems (12% in I-C vs 25% in II-C; Fisher's exact test $p < 10^{-10}$).

326 Locally adapted spacers are not uniformly distributed along the CRISPR array (Figure 3c). On
327 the contrary, they tend to concentrate near both ends of the array. The enrichment at the
328 leader (or proximal) end is consistent with recent spacer acquisitions targeting MGE from the
329 local microbiome. However, the excess of locally adapted spacers near the trailer (or distal)
330 end, especially evident in subtypes I-C and I-E, is somewhat paradoxical. In fact, distal spacers
331 are generally the oldest in the array and their role in immunity is a matter of debate (77, 78).
332 Three lines of evidence reject the possibility that locally adapted spacers near the distal end
333 represent a technical artifact. First, array orientation and spacer order are highly consistent

334 across arrays from the same lineage, despite the arrays being independently assembled from
335 different samples. Second, the same trend is observed when the orientation of the arrays is
336 inferred with three independent methods (Cas Orientation, CRISPRDirection, and CRISPRLoci;
337 Suppl. Fig. S1). Third, in most subtypes, mismatches between the spacer and target sequences
338 become more frequent towards the trailer end, as expected for correctly assembled arrays.

339 We conceive two possible biological explanations for the observed trend in the distribution of
340 locally adapted spacers. The first is that CRISPR-based immunity leads to the complete
341 elimination of MGE that are targeted by spacers in the middle section of the array, whereas
342 MGE targeted by distal, less efficient spacers can persist in the microbiome. In this scenario,
343 most middle spacers would not be identified as "locally adapted" simply because they
344 succeeded in eliminating their targets. Although CRISPR immunity can effectively prevent MGE
345 proliferation in culture (79, 80), it is less likely that complete elimination of MGE occurs in
346 complex, spatially extended microbial communities (26, 81) (and the human gut virome
347 appears indeed highly stable over time (82)). More generally, this scenario does not account
348 for the possibility that the same MGE is targeted by multiple spacers located at different
349 positions in different arrays, which would effectively decouple within-array spacer positions
350 and population-level MGE abundances.

351 The second (and more likely) explanation is that locally adapted spacers near the trailing end
352 are a consequence of sustained selection for long-term immunity against endemic MGE. It is
353 widely accepted that intense selection and frequent spacer loss via homologous recombination
354 shape the composition of CRISPR arrays in natural populations, promoting the maintenance of
355 functional spacers (42). Therefore, if endemic MGE impose a sustained burden on the
356 population, it can be expected that older spacers providing with effective long-term immunity
357 against such MGE accumulate at the distal part of the array. This phenomenon has been
358 observed in *Leptospirillum* strains from an acid mine drainage where the same phage

359 population persisted for at least 5 years (23, 83). We propose a similar explanation in the case
360 of the adult human gut, where bacteria coexist with a very stable virome dominated by crass-
361 like phages (82, 84-86) and exposed to continuous (but low-intensity) induction of prophages
362 (87, 88). In support of the long-term memory role of distal CRISPR spacers, we found that the
363 fraction of spacers targeting lysogenic phages systematically increases from the leading to the
364 trailing end of the array (Figure 3d). Considering the high rates of lytic activity and CRISPR
365 spacer turnover that characterize the infant gut, and the possible contribution of early phage-
366 bacteria interactions to shaping the adult gut microbiome (13, 89-91), an intriguing possibility
367 might be that long-term memory spacers were selected during the consolidation of a
368 coexisting community of MGE and bacteria in the first years of age.

369 To provide efficient long-term immune memory, spacers near the trailing end of the array
370 must be expressed at sufficient levels to form effector complexes. Indeed, lack of expression
371 due to early termination of transcription limits the functionality of distal spacers in
372 experimental conditions (39-41). By analyzing metatranscriptomes, we found that distal
373 spacers are moderately expressed, with transcript abundances around 30-40% of those
374 observed for spacers near the leading end (Figure S2). Distal spacers may also contribute to
375 long-term immunity through primed adaptation, that is, by promoting the acquisition of
376 additional spacers from cognate MGE (92). Primed adaptation is widespread in type I and II
377 CRISPR systems (93, 94) and operates for spacers that do not exactly match their target (95) or
378 are expressed at low levels (96). Because spacers acquired through this mechanism are often
379 biased in terms of genomic location and strand specificity, indirect evidence for primed
380 adaptation can be obtained by testing for non-random patterns in the relative location of
381 protospacers within the target genome. Following that strategy, we identified several
382 instances of MGE targeted by >1 spacer from the same array, which could correspond to
383 primed adaptation. However, the number of instances observed was insufficient for statistical
384 confirmation.

385 CRISPR-Cas immunity modulates microbiome composition

386 The longitudinal nature of the HMP2-IBD metagenomic survey allowed us to investigate how
387 CRISPR immunity affects the dynamics of MGE and bacterial hosts in the gut microbiome. We
388 first studied if the appearance of a new spacer in a preexisting CRISPR lineage (possibly
389 implying a spacer acquisition event) was associated with changes in the prevalence and
390 abundance of the target MGE (Figure 4a). Consistent with the expectation that exposure to
391 MGE induces immune adaptation, we found that putative spacer acquisitions occur in samples
392 in which the abundance of the target MGE is higher than average. In sharp contrast,
393 subsequent observations of the spacer are associated with significantly decreased MGE
394 abundances. Notably, such decrease is not observed in populations that have seemingly lost
395 the spacer but in which the CRISPR lineage (and therefore the host) is still detected. As
396 expected, these trends are strongest in the case of lytic and lysogenic phages, but also
397 significant for plasmids. Despite its quantitative effect on MGE abundance, CRISPR-Cas
398 immunity does not generally lead to the complete elimination of its targets. Indeed, around
399 50% of lytic and lysogenic phages targeted by CRISPR-Cas can still be detected following the
400 acquisition of a cognate spacer in the population (Figure 4b). In the case of plasmids,
401 acquisition of cognate spacers does not produce any significant drop in prevalence. All in all,
402 these observations imply that CRISPR-Cas immunity modulates phage and plasmid abundances
403 in the human gut, while not compromising their long-term persistence. This conclusion is
404 consistent with previous observations from free-living communities and with the theoretical
405 prediction that CRISPR-Cas immunity can stabilize complex phage-host populations through
406 eco-evolutionary feedback (28, 97-99).

407 We then investigated if the interplay of CRISPR immunity and phage predation affects host
408 diversity at the population level. To that end, we divided CRISPR lineages into those that are
409 possibly involved in effective immune interactions (by harboring at least one locally adapted

410 spacer) and those that do not show any evidence of local adaptation. The former tend to be
411 more diverse, that is, they encompass more array variants per lineage, likely reflecting higher
412 rates of spacer acquisition and loss (Figure 4c). Moreover, the diversity of CRISPR arrays in
413 these lineages is lower in samples in which targeted phages are present (Figure 4d). This
414 observation underscores the occurrence of weak MGE-induced selective sweeps in the host
415 population, that transiently favor those hosts bearing CRISPR arrays with cognate spacers
416 (Figure 4e). The weakness of these selective sweeps could in part be explained by clonal
417 interference among multiple immune lineages, as previously suggested for hot spring
418 communities (27), and by the spatially structured nature of the gut microbiome.

419

420 Temporal dynamics of MGE and CRISPR arrays

421 To better understand the interplay between CRISPR-Cas immunity and the dynamics of MGE
422 and hosts, we selected 1,769 time series for single spacer-target pairs whose dynamics are
423 compatible with possible spacer acquisition events at the population level. Then, we applied
424 an unsupervised clustering algorithm to group similar trajectories based on the temporal
425 patterns of presence or absence of spacers, targets, and CRISPR lineages (the latter serving as
426 proxies for host subpopulations). Despite the potentially combinatorial diversity of individual
427 trajectories, we found that 40% of them belonged to one of the 10 largest clusters. The
428 inspection of such representative clusters reveals qualitative differences across MGE in their
429 temporal dynamics (Figure 5). Phages are preferentially associated with episodic dynamics (TC-
430 1, 3, 5, and 7), whereas plasmids display recurrency (TC-13 and 14). The most frequent
431 trajectory in the data set (TC-1) corresponds to a single observation of a CRISPR lineage
432 followed by a simultaneous observation of a new spacer and the cognate phage, in what might
433 represent a spacer acquisition event. Other representative trajectories involve variations of
434 the same theme, with loss of the phage following spacer acquisition (TC-5) or missing the

435 intermediate acquisition event (TC-2 and 3). In contrast, the most frequent trajectories
436 associated with plasmids represent two alternative scenarios, one dominated by the plasmid
437 (TC-13) and the other characterized by the stable coexistence of plasmid-bearing and plasmid-
438 resistant host subpopulations (TC-14). The latter is reminiscent of the global population
439 structure of *Pseudomonas aeruginosa*, in which spacers from active CRISPR-Cas systems target
440 MGE integrated in closely related genomes that lack CRISPR (100), and in hot spring
441 populations of *Hydrogenobacter sp.*, where the spacers of single-cell assembled genomes
442 target MGE found in other cells from the same population (101). Our results suggest that
443 crosstalk between CRISPR-Cas systems and plasmids that coexist in the host population is also
444 widespread in the human gut.

445

446 **Conclusions**

447 By analyzing a large metagenomic dataset encompassing 130 individual gut microbiomes
448 periodically sampled for up to 1 year, we found that the interaction of phages and CRISPR-Cas
449 immunity quantitatively affects the composition of the human gut microbiome. Phage and
450 plasmid abundances tend to decrease after the first appearance of cognate CRISPR spacers in
451 the microbiome, although such decrease does not generally lead to their complete elimination.
452 On the other side, in the presence of MGE, bacterial populations experience weak and
453 incomplete selective sweeps favoring immune lineages. The coexistence of phages and diverse
454 host lineages despite effective CRISPR-Cas immunity suggests that spatial heterogeneity, clonal
455 interference, and fluctuating selection could have a major role in explaining phage-bacteria
456 interaction dynamics in the gut. Disentangling the contribution of these mechanisms warrants
457 further investigation, as they may differentially affect the efficacy of phage therapy and other
458 intervention strategies aimed at controlling the gut microbiome.

459 A second major finding of this study was the high relative abundance of locally adapted
460 spacers located in the distal region of CRISPR arrays and targeting lysogenic phages. This
461 finding strongly suggests that selection promotes the maintenance of long-term memory
462 spacers against endemic MGE, and that periodic reactivation of prophages is an important
463 stressor for the human gut microbiome. Interestingly, leader and trailer regions of CRISPR
464 arrays appear to be specialized in short- and long-term memory, respectively. We expect that
465 selection for long-lasting immunity and polar allocation of long-term memory spacers are not
466 unique to the human gut microbiome, but general features of microbial communities with
467 stable viromes. Testing such prediction constitutes an interesting subject of future research.

468

469 **Acknowledgements**

470 A.L-B. is supported by the Agencia Estatal de Investigación of Spain (Grant No. PRE2020-
471 092935). J.B. is supported by the Maria Zambrano grant of the Spanish Ministry of Universities
472 (Grant No. UP2021-035), and the Severo Ochoa Program for Centres of Excellence in R&D of
473 the Agencia Estatal de Investigación of Spain (Grant No. CEX2020-000999-S (2022–2025) to the
474 CBGP). J.I is supported by the Ramón y Cajal Programme of the Spanish Ministry of Science
475 (Grant No. RYC-2017–22524); the Agencia Estatal de Investigación of Spain (Grant No.
476 PID2019-106618GA-I00), the Severo Ochoa Programme for Centres of Excellence in R&D of the
477 Agencia Estatal de Investigación of Spain (Grant No. SEV-2016–0672 (2017–2021) to the
478 CBGP); and the Comunidad de Madrid (through the call Research Grants for Young
479 Investigators from Universidad Politécnica de Madrid, Grant No. M190020074JIIS).

480

481 **Data availability**

482 All relevant data produced in this study have been deposited in the Zenodo repository (DOI:

483 10.5281/zenodo.10555110) and are freely available.

484 **References**

485

486 1. Diaz-Munoz SL, Koskella B. Bacteria-phage interactions in natural environments. Adv
487 Appl Microbiol. 2014;89:135-83.

488 2. Fernandez L, Rodriguez A, Garcia P. Phage or foe: an insight into the impact of viral
489 predation on microbial communities. The ISME journal. 2018;12(5):1171-9.

490 3. Chevallereau A, Pons BJ, van Houte S, Westra ER. Interactions between bacterial and
491 phage communities in natural environments. Nat Rev Microbiol. 2022;20(1):49-62.

492 4. Brown TL, Charity OJ, Adriaenssens EM. Ecological and functional roles of
493 bacteriophages in contrasting environments: marine, terrestrial and human gut. Curr Opin
494 Microbiol. 2022;70:102229.

495 5. Koskella B, Brockhurst MA. Bacteria-phage coevolution as a driver of ecological and
496 evolutionary processes in microbial communities. FEMS Microbiol Rev. 2014;38(5):916-31.

497 6. Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, Breitbart M, et al. Viral and
498 microbial community dynamics in four aquatic environments. The ISME journal. 2010;4(6):739-
499 51.

500 7. Faruque SM, Islam MJ, Ahmad QS, Faruque AS, Sack DA, Nair GB, et al. Self-limiting
501 nature of seasonal cholera epidemics: Role of host-mediated amplification of phage.
502 Proceedings of the National Academy of Sciences of the United States of America.
503 2005;102(17):6119-24.

504 8. Suttle CA. Marine viruses--major players in the global ecosystem. Nat Rev Microbiol.
505 2007;5(10):801-12.

506 9. Braga LPP, Spor A, Kot W, Breuil MC, Hansen LH, Setubal JC, et al. Impact of phages on
507 soil bacterial communities and nitrogen availability under different assembly scenarios.
508 Microbiome. 2020;8(1):52.

509 10. Koskella B. Phage-mediated selection on microbiota of a long-lived host. Curr Biol.
510 2013;23(13):1256-60.
511 11. Manrique P, Dills M, Young MJ. The Human Gut Phage Community and Its Implications
512 for Health and Disease. Viruses. 2017;9(6).
513 12. Gogokhia L, Buhrke K, Bell R, Hoffman B, Brown DG, Hanke-Gogokhia C, et al.
514 Expansion of Bacteriophages Is Linked to Aggravated Intestinal Inflammation and Colitis. Cell
515 Host Microbe. 2019;25(2):285-99 e8.
516 13. Shamash M, Maurice CF. Phages in the infant gut: a framework for virome
517 development during early life. The ISME journal. 2022;16(2):323-30.
518 14. Millman A, Melamed S, Leavitt A, Doron S, Bernheim A, Hor J, et al. An expanded
519 arsenal of immune systems that protect bacteria from phages. Cell Host Microbe.
520 2022;30(11):1556-69 e5.
521 15. Georjon H, Bernheim A. The highly diverse antiphage defence systems of bacteria. Nat
522 Rev Microbiol. 2023;21(10):686-700.
523 16. Watson BNJ, Steens JA, Staals RHJ, Westra ER, van Houte S. Coevolution between
524 bacterial CRISPR-Cas systems and their bacteriophages. Cell Host Microbe. 2021;29(5):715-25.
525 17. Meaden S, Biswas A, Arkhipova K, Morales SE, Dutilh BE, Westra ER, et al. High viral
526 abundance and low diversity are associated with increased CRISPR-Cas prevalence across
527 microbial ecosystems. Curr Biol. 2022;32(1):220-7 e5.
528 18. Chabas H, Muller V, Bonhoeffer S, Regoes RR. Epidemiological and evolutionary
529 consequences of different types of CRISPR-Cas systems. PLoS computational biology.
530 2022;18(7):e1010329.
531 19. Martinez Arbas S, Narayanasamy S, Herold M, Lebrun LA, Hoopmann MR, Li S, et al.
532 Roles of bacteriophages, plasmids and CRISPR immunity in microbial community dynamics
533 revealed using time-series integrated meta-omics. Nature microbiology. 2021;6(1):123-35.

534 20. Cornuault JK, Moncaut E, Loux V, Mathieu A, Sokol H, Petit MA, et al. The enemy from
535 within: a prophage of Roseburia intestinalis systematically turns lytic in the mouse gut, driving
536 bacterial adaptation by CRISPR spacer acquisition. The ISME journal. 2020;14(3):771-87.
537 21. Westra ER, van Houte S, Gandon S, Whitaker R. The ecology and evolution of microbial
538 CRISPR-Cas adaptive immune systems. Philos Trans R Soc Lond B Biol Sci.
539 2019;374(1772):20190101.
540 22. Emerson JB, Andrade K, Thomas BC, Norman A, Allen EE, Heidelberg KB, et al. Virus-
541 host and CRISPR dynamics in Archaea-dominated hypersaline Lake Tyrrell, Victoria, Australia.
542 Archaea. 2013;2013:370871.
543 23. Sun CL, Thomas BC, Barrangou R, Banfield JF. Metagenomic reconstructions of
544 bacterial CRISPR loci constrain population histories. The ISME journal. 2016;10(4):858-70.
545 24. Pauly MD, Bautista MA, Black JA, Whitaker RJ. Diversified local CRISPR-Cas immunity to
546 viruses of Sulfolobus islandicus. Philos Trans R Soc Lond B Biol Sci. 2019;374(1772):20180093.
547 25. Medvedeva S, Liu Y, Koonin EV, Severinov K, Prangishvili D, Krupovic M. Virus-borne
548 mini-CRISPR arrays are involved in interviral conflicts. Nature communications.
549 2019;10(1):5204.
550 26. Guerrero LD, Perez MV, Orellana E, Piuri M, Quiroga C, Erijman L. Long-run bacteria-
551 phage coexistence dynamics under natural habitat conditions in an environmental
552 biotechnology system. The ISME journal. 2021;15(3):636-48.
553 27. Held NL, Herrera A, Cadillo-Quiroz H, Whitaker RJ. CRISPR associated diversity within a
554 population of Sulfolobus islandicus. PLoS One. 2010;5(9).
555 28. Iranzo J, Lobkovsky AE, Wolf YI, Koonin EV. Evolutionary dynamics of the prokaryotic
556 adaptive immunity system CRISPR-Cas in an explicit ecological context. Journal of bacteriology.
557 2013;195(17):3834-44.

558 29. Iranzo J, Lobkovsky AE, Wolf YI, Koonin EV. Immunity, suicide or both? Ecological
559 determinants for the combined evolution of anti-pathogen defense systems. BMC evolutionary
560 biology. 2015;15:43.
561 30. Payne P, Geyrhofer L, Barton NH, Bollback JP. CRISPR-based herd immunity can limit
562 phage epidemics in bacterial populations. Elife. 2018;7.
563 31. Iranzo J, Faure G, Wolf YI, Koonin EV. Game-Theoretical Modeling of Interviral Conflicts
564 Mediated by Mini-CRISPR Arrays. Front Microbiol. 2020;11:381.
565 32. Andersson AF, Banfield JF. Virus population dynamics and acquired virus resistance in
566 natural microbial communities. Science. 2008;320(5879):1047-50.
567 33. Kosmopoulos JC, Campbell DE, Whitaker RJ, Wilbanks EG. Horizontal Gene Transfer
568 and CRISPR Targeting Drive Phage-Bacterial Host Interactions and Coevolution in "Pink Berry"
569 Marine Microbial Aggregates. Applied and environmental microbiology. 2023;89(7):e0017723.
570 34. England WE, Kim T, Whitaker RJ. Metapopulation Structure of CRISPR-Cas Immunity in
571 Pseudomonas aeruginosa and Its Viruses. mSystems. 2018;3(5).
572 35. Devoto AE, Santini JM, Olm MR, Anantharaman K, Munk P, Tung J, et al. Megaphages
573 infect Prevotella and variants are widespread in gut microbiomes. Nature microbiology.
574 2019;4(4):693-700.
575 36. Bradde S, Nourmohammad A, Goyal S, Balasubramanian V. The size of the immune
576 repertoire of bacteria. Proceedings of the National Academy of Sciences of the United States of
577 America. 2020;117(10):5144-51.
578 37. Weissman JL, Fagan WF, Johnson PLF. Selective Maintenance of Multiple CRISPR Arrays
579 Across Prokaryotes. CRISPR J. 2018;1(6):405-13.
580 38. Ou Y, McInerney JO. High Frequency of Dynamic Rearrangements In Crispr loci.
581 bioRxiv. 2022:2022.05.19.492656.

582 39. Richter H, Zoephel J, Schermuly J, Maticzka D, Backofen R, Randau L. Characterization
583 of CRISPR RNA processing in Clostridium thermocellum and Methanococcus maripaludis.
584 Nucleic acids research. 2012;40(19):9887-96.
585 40. Elmore JR, Yokooji Y, Sato T, Olson S, Glover CV, 3rd, Graveley BR, et al. Programmable
586 plasmid interference by the CRISPR-Cas system in Thermococcus kodakarensis. RNA Biol.
587 2013;10(5):828-40.
588 41. Nickel L, Weidenbach K, Jager D, Backofen R, Lange SJ, Heidrich N, et al. Two CRISPR-
589 Cas systems in Methanosarcina mazei strain Go1 display common processing features despite
590 belonging to different types I and III. RNA Biol. 2013;10(5):779-91.
591 42. Garrett SC. Pruning and Tending Immune Memories: Spacer Dynamics in the CRISPR
592 Array. Front Microbiol. 2021;12:664299.
593 43. Stern A, Mick E, Tirosh I, Sagy O, Sorek R. CRISPR targeting reveals a reservoir of
594 common phages associated with the human gut microbiome. Genome research.
595 2012;22(10):1985-94.
596 44. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et
597 al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature.
598 2019;569(7758):655-62.
599 45. Podlevsky JD, Hudson CM, Timlin JA, Williams KP. CasCollect: targeted assembly of
600 CRISPR-associated operons from high-throughput sequencing data. NAR genomics and
601 bioinformatics. 2020;2(3):lqaa063.
602 46. Rognes T, Flouri T, Nichols B, Quince C, Mahe F. VSEARCH: a versatile open source tool
603 for metagenomics. PeerJ. 2016;4:e2584.
604 47. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes De Novo
605 Assembler. Current protocols in bioinformatics. 2020;70(1):e102.
606 48. Grissa I, Vergnaud G, Pourcel C. The CRISPRdb database and tools to display CRISPRs
607 and to generate dictionaries of spacers and repeats. BMC Bioinformatics. 2007;8:172.

608 49. Makarova KS, Wolf YI, Iranzo J, Shmakov SA, Alkhnbashi OS, Brouns SJJ, et al.
609 Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. Nat
610 Rev Microbiol. 2020;18(2):67-83.
611 50. Russel J, Pinilla-Redondo R, Mayo-Munoz D, Shah SA, Sorensen SJ. CRISPRCasTyper:
612 Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. The CRISPR
613 journal. 2020;3(6):462-9.
614 51. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation
615 sequencing data. Bioinformatics. 2012;28(23):3150-2.
616 52. Milicevic O, Repac J, Bozic B, Djordjevic M, Djordjevic M. A Simple Criterion for
617 Inferring CRISPR Array Direction. Frontiers in microbiology. 2019;10:2054.
618 53. Alkhnbashi OS, Mitrofanov A, Bonidia R, Raden M, Tran VD, Eggenhofer F, et al.
619 CRISPRloci: comprehensive and accurate annotation of CRISPR-Cas systems. Nucleic acids
620 research. 2021;49(W1):W125-W30.
621 54. Biswas A, Fineran PC, Brown CM. Accurate computational prediction of the transcribed
622 strand of CRISPR non-coding RNAs. Bioinformatics. 2014;30(13):1805-13.
623 55. Alstott J, Bullmore E, Plenz D. Powerlaw: a Python package for analysis of heavy-tailed
624 distributions. PLoS One. 2014;9(1):e85777.
625 56. Benler S, Yutin N, Antipov D, Rayko M, Shmakov S, Gussow AB, et al. Thousands of
626 previously unknown phages discovered in whole-community human gut metagenomes.
627 Microbiome. 2021;9(1):78.
628 57. Botelho J. Defense systems are pervasive across chromosomally integrated mobile
629 genetic elements and are inversely correlated to virulence and antimicrobial resistance.
630 Nucleic acids research. 2023;51(9):4385-97.
631 58. Lai S, Jia L, Subramanian B, Pan S, Zhang J, Dong Y, et al. mMGE: a database for human
632 metagenomic extrachromosomal mobile genetic elements. Nucleic acids research.
633 2021;49(D1):D783-D91.

634 59. Jiang X, Hall AB, Xavier RJ, Alm EJ. Comprehensive analysis of chromosomal mobile
635 genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. PLoS
636 One. 2019;14(12):e0223680.
637 60. Wang M, Liu G, Liu M, Tai C, Deng Z, Song J, et al. ICEberg 3.0: functional categorization
638 and analysis of the integrative and conjugative elements in bacteria. Nucleic acids research.
639 2024;52(D1):D732-D7.
640 61. Camargo AP, Call L, Roux S, Nayfach S, Huntemann M, Palaniappan K, et al. IMG/PR: a
641 database of plasmids from genomes and metagenomes with rich annotations and metadata.
642 Nucleic acids research. 2024;52(D1):D164-D73.
643 62. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic
644 gene recognition and translation initiation site identification. BMC Bioinformatics.
645 2010;11:119.
646 63. Kunin V, Sorek R, Hugenholtz P. Evolutionary conservation of sequence and secondary
647 structures in CRISPR repeats. Genome biology. 2007;8(4):R61.
648 64. von Meijenfeldt FAB, Arkhipova K, Cambuy DD, Coutinho FH, Dutilh BE. Robust
649 taxonomic classification of uncharted microbial sequences and bins with CAT and BAT.
650 Genome biology. 2019;20(1):217.
651 65. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics.
652 2018;34(18):3094-100.
653 66. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The
654 Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA
655 sequencing data. Genome research. 2010;20(9):1297-303.
656 67. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic
657 features. Bioinformatics. 2010;26(6):841-2.

658 68. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely
659 available Python tools for computational molecular biology and bioinformatics. Bioinformatics.
660 2009;25(11):1422-3.
661 69. Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal
662 community structure. Proceedings of the National Academy of Sciences of the United States of
663 America. 2008;105(4):1118-23.
664 70. Pavlova YS, Paez-Espino D, Morozov AY, Belalov IS. Searching for fat tails in CRISPR-Cas
665 systems: Data analysis and mathematical modeling. PLoS computational biology.
666 2021;17(3):e1008841.
667 71. Toms A, Barrangou R. On the global CRISPR array behavior in class I systems. Biology
668 direct. 2017;12(1):20.
669 72. Kostic AD, Xavier RJ, Gevers D. The microbiome in inflammatory bowel disease: current
670 status and the future ahead. Gastroenterology. 2014;146(6):1489-99.
671 73. Pinilla-Redondo R, Mayo-Munoz D, Russel J, Garrett RA, Randau L, Sorensen SJ, et al.
672 Type IV CRISPR-Cas systems are highly diverse and involved in competition between plasmids.
673 Nucleic acids research. 2020;48(4):2000-12.
674 74. Shmakov SA, Sitnik V, Makarova KS, Wolf YI, Severinov KV, Koonin EV. The CRISPR
675 Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes. MBio. 2017;8(5).
676 75. Shmakov SA, Wolf YI, Savitskaya E, Severinov KV, Koonin EV. Mapping CRISPR
677 spaceromes reveals vast host-specific viromes of prokaryotes. Commun Biol. 2020;3(1):321.
678 76. Soto-Perez P, Bisanz JE, Berry JD, Lam KN, Bondy-Denomy J, Turnbaugh PJ. CRISPR-Cas
679 System of a Prevalent Human Gut Bacterium Reveals Hyper-targeting against Phages in a
680 Human Virome Catalog. Cell Host Microbe. 2019;26(3):325-35 e5.
681 77. McGinn J, Marraffini LA. Molecular mechanisms of CRISPR-Cas spacer acquisition. Nat
682 Rev Microbiol. 2019;17(1):7-12.

683 78. McGinn J, Marraffini LA. CRISPR-Cas Systems Optimize Their Immune Response by
684 Specifying the Site of Spacer Integration. Mol Cell. 2016;64(3):616-23.
685 79. Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in
686 staphylococci by targeting DNA. Science. 2008;322(5909):1843-5.
687 80. van Houte S, Ekroth AK, Broniewski JM, Chabas H, Ashby B, Bondy-Denomy J, et al. The
688 diversity-generating benefits of a prokaryotic adaptive immune system. Nature.
689 2016;532(7599):385-8.
690 81. Watson BNJ, Capria L, Alseth EO, Pons B, Biswas A, Lenzi L, et al. CRISPR-Cas in
691 Pseudomonas aeruginosa provides transient population-level immunity against high phage
692 exposures. The ISME journal. 2024.
693 82. Shkoporov AN, Clooney AG, Sutton TDS, Ryan FJ, Daly KM, Nolan JA, et al. The Human
694 Gut Virome Is Highly Diverse, Stable, and Individual Specific. Cell Host Microbe.
695 2019;26(4):527-41 e5.
696 83. Weinberger AD, Sun CL, Plucinski MM, Denef VJ, Thomas BC, Horvath P, et al.
697 Persisting viral sequences shape microbial CRISPR-based immunity. PLoS computational
698 biology. 2012;8(4):e1002475.
699 84. Manrique P, Bolduc B, Walk ST, van der Oost J, de Vos WM, Young MJ. Healthy human
700 gut phageome. Proceedings of the National Academy of Sciences of the United States of
701 America. 2016;113(37):10400-5.
702 85. Koonin EV, Dolja VV, Krupovic M. The healthy human virome: from virus-host
703 symbiosis to disease. Curr Opin Virol. 2021;47:86-94.
704 86. Smith L, Goldobina E, Govi B, Shkoporov AN. Bacteriophages of the Order Crassvirales:
705 What Do We Currently Know about This Keystone Component of the Human Gut Virome?
706 Biomolecules. 2023;13(4).
707 87. Sutcliffe SG, Reyes A, Maurice CF. Bacteriophages playing nice: Lysogenic
708 bacteriophage replication stable in the human gut microbiota. iScience. 2023;26(2):106007.

709 88. Dion MB, Shah SA, Deng L, Thorsen J, Stokholm J, Krogfelt KA, et al. Escherichia coli
710 CRISPR arrays from early life fecal samples preferentially target prophages. The ISME journal.
711 2024.
712 89. Shah SA, Deng L, Thorsen J, Pedersen AG, Dion MB, Castro-Mejia JL, et al. Expanding
713 known viral diversity in the healthy infant gut. Nature microbiology. 2023;8(5):986-98.
714 90. Wu J, Zhang H, Gan R, Xia Y, Zhang F, Wang D, et al. CRISPR dynamics during the
715 interaction between bacteria and phage in the first year of life. Microb Genom. 2023;9(7).
716 91. Redgwell TA, Thorsen J, Petit M-A, Deng L, Vestergaard GA, Russel J, et al. Prophages in
717 the infant gut are largely induced, and may be functionally relevant to their hosts. bioRxiv.
718 2021:2021.06.25.449885.
719 92. Sternberg SH, Richter H, Charpentier E, Qimron U. Adaptation in CRISPR-Cas Systems.
720 Mol Cell. 2016;61(6):797-808.
721 93. Nussenzweig PM, Marraffini LA. Molecular Mechanisms of CRISPR-Cas Immunity in
722 Bacteria. Annu Rev Genet. 2020;54:93-120.
723 94. Nicholson TJ, Jackson SA, Croft BI, Staals RHJ, Fineran PC, Brown CM. Bioinformatic
724 evidence of widespread priming in type I and II CRISPR-Cas systems. RNA Biol. 2019;16(4):566-
725 76.
726 95. Datsenko KA, Pougach K, Tikhonov A, Wanner BL, Severinov K, Semenova E. Molecular
727 memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system.
728 Nature communications. 2012;3:945.
729 96. Deecker SR, Ensminger AW. Type I-F CRISPR-Cas Distribution and Array Dynamics in
730 Legionella pneumophila. G3. 2020;10(3):1039-50.
731 97. Haerter JO, Sneppen K. Spatial structure and Lamarckian adaptation explain extreme
732 genetic diversity at CRISPR locus. mBio. 2012;3(4):e00126-12.
733 98. Koonin EV, Wolf YI. Evolution of the CRISPR-Cas adaptive immunity systems in
734 prokaryotes: models and observations on virus-host coevolution. Mol Biosyst. 2015;11(1):20-7.

735 99. Weissman JL, Holmes R, Barrangou R, Moineau S, Fagan WF, Levin B, et al. Immune
736 loss as a driver of coexistence during host-phage coevolution. The ISME journal.
737 2018;12(2):585-97.
738 100. Wheatley RM, MacLean RC. CRISPR-Cas systems restrict horizontal gene transfer in
739 Pseudomonas aeruginosa. The ISME journal. 2021;15(5):1420-33.
740 101. Bowers RM, Nayfach S, Schulz F, Jungbluth SP, Ruhl IA, Sheremet A, et al. Dissecting
741 the dominant hot spring microbial populations based on community-wide sampling at single-
742 cell genomic resolution. The ISME journal. 2022;16(5):1337-47.

743

744

745 **Figure legends**

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

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

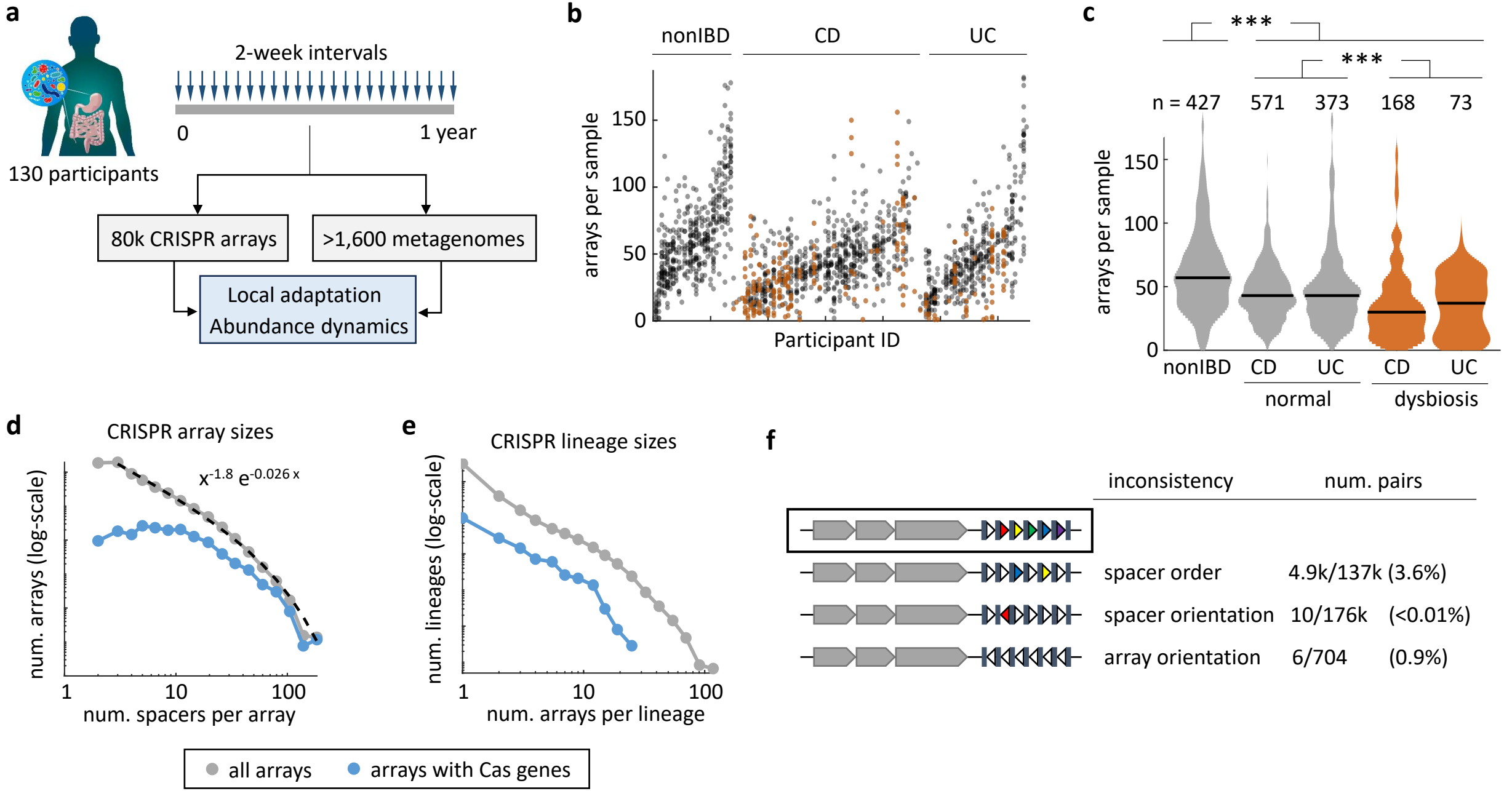
766 **Figure 3:** Targets of CRISPR-Cas immunity in the gut microbiome. (a) Relative representation of
767 different classes of mobile genetic elements among the targets of CRISPR-Cas systems. Left pie
768 chart: hits to a merged database of known mobile genetic elements (MGE db). Right:
769 protospacers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

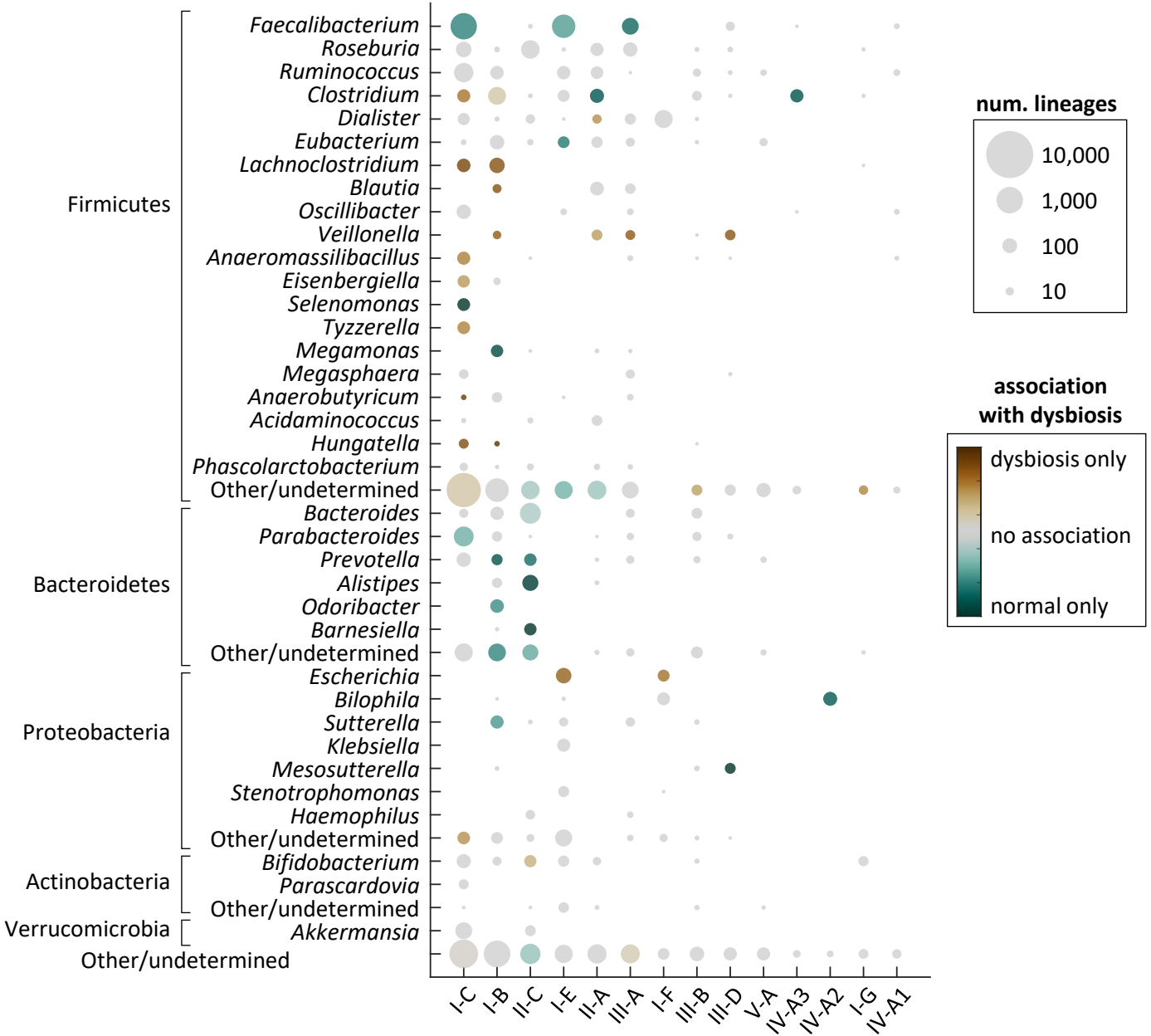
770 from the same participant. The segment for lytic viruses also includes chronic (non-lysogenic)
771 viruses (b) Distribution and functional profile of CRISPR targets across CRISPR-Cas subtypes.
772 Locally adapted spacers (bar plot on the left) are those that target sequences from the local
773 microbiome. (c) Relative position of locally adapted spacers within the CRISPR array. L: leading
774 end; T: trailing end. (d) Fraction of spacers targeting lysogenic phages (with respect to all that
775 target phages) in the leading, middle, and distal sections of the array. Whiskers represent 95%
776 confidence intervals calculated with the Clopper-Pearson method. In (b), (c), and (d), only
777 those subtypes with >100 locally adapted spacers are shown. Subtype II-C was additionally
778 excluded from (c) and (d) due to the unavailability of accurate predictions for the orientation
779 of those arrays.

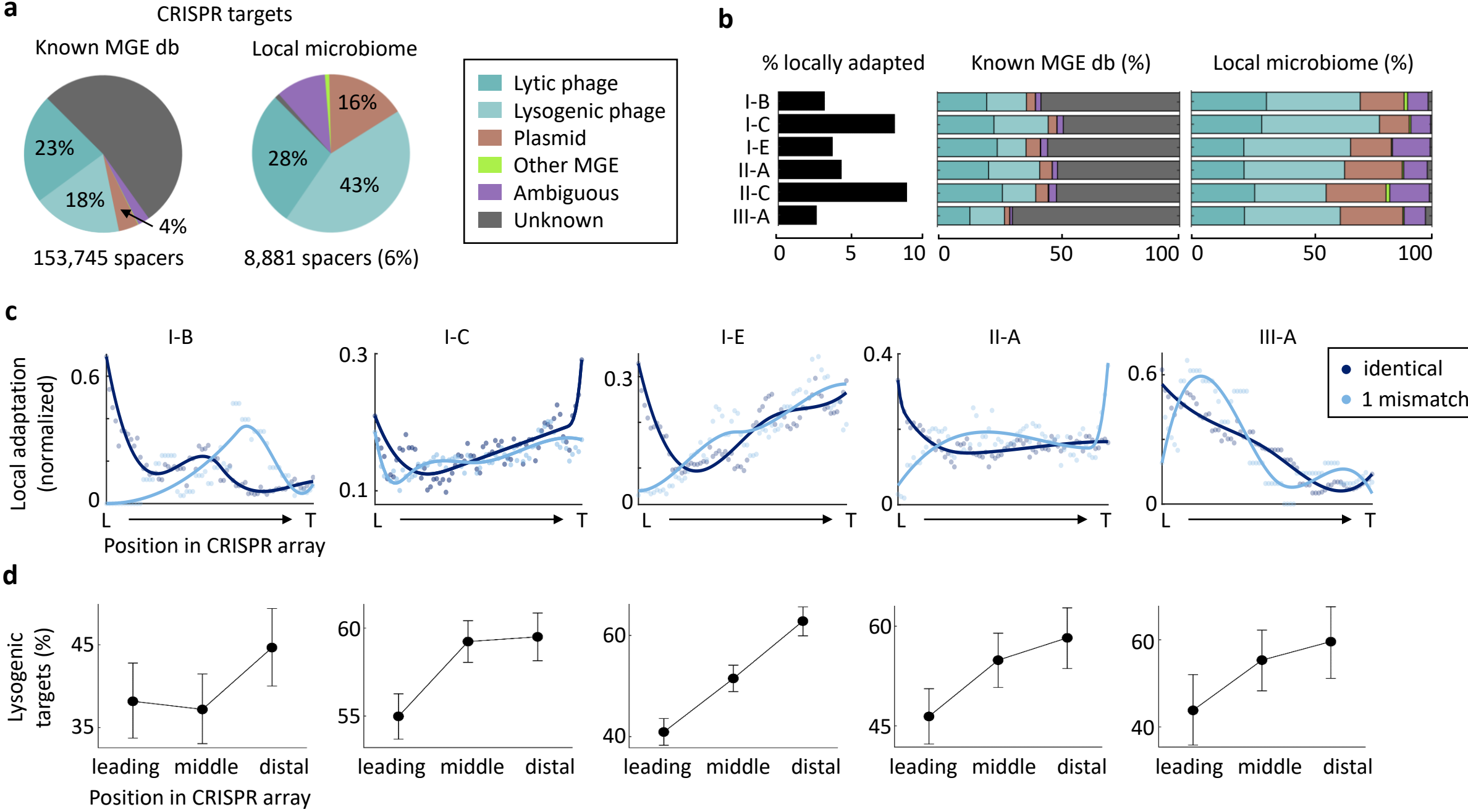
780 **Figure 4:** Effect of CRISPR-Cas immunity on MGE abundance and host diversity. (a) Median
781 normalized abundance of targeted MGE (plasmids, lytic/non-lysogenic phages, and lysogenic
782 phages) before the spacer was first observed in the host population ("Before"), in the sample
783 in which the spacer was observed for the first time ("Acquisition"), in subsequent samples in
784 which the spacer was present ("After – With spacer"), and in subsequent samples that
785 contained the same CRISPR lineage but not the spacer ("After – No spacer"). Error bars
786 indicate the 95% confidence intervals for the median, based on 2,000 bootstraps. All
787 comparisons shown above the plot are statistically significant with $p < 10^{-4}$ (Mann-Whitney
788 test). (b) Prevalence of targeted MGE, calculated as the fraction of samples in which the target
789 is detected. Error bars indicate the 95% confidence intervals (Clopper-Pearson method).
790 Statistical significance based on Fisher's exact test: $p < 0.05$ (*), $p < 0.01$ (**), $p < 0.001$ (***).
791 (c) Diversity distribution (number of distinct arrays) of CRISPR lineages that target at least one
792 MGE from the local microbiome ("active" lineages) and those that do not show any sign of
793 local adaptation ("inactive" lineages). (d) Number of arrays per lineage in the presence and
794 absence of MGE. For a given active lineage, samples with MGE are those that contain at least
795 one MGE targeted by at least one spacer from that lineage. (e) Percentage of arrays per

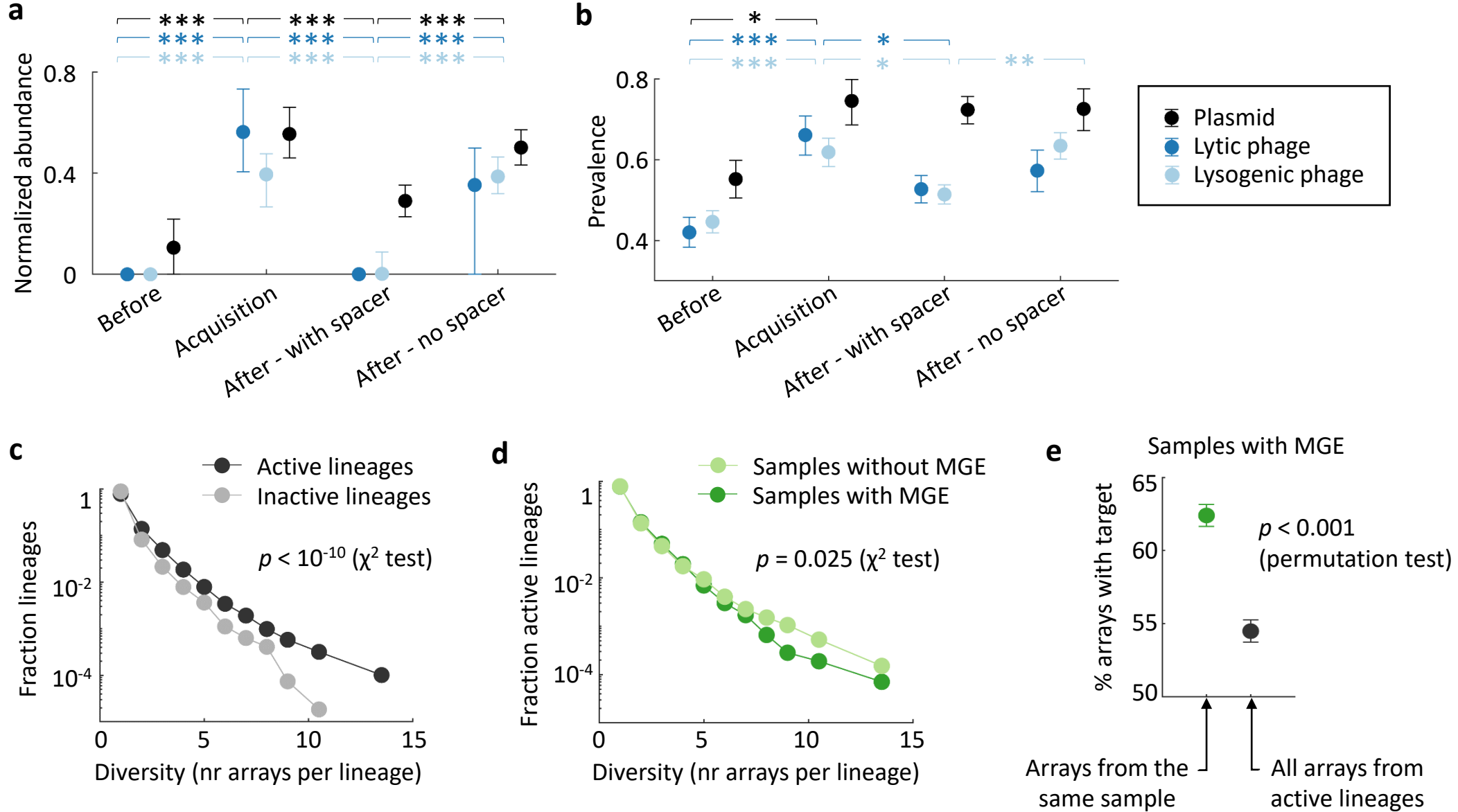
796 lineage that target MGE from the same sample, compared to the expected value for randomly
797 chosen arrays. The randomized set only includes active lineages and samples with MGE
798 targeted by those lineages. Error bars indicate the 95% confidence intervals based on 2,000
799 bootstraps.

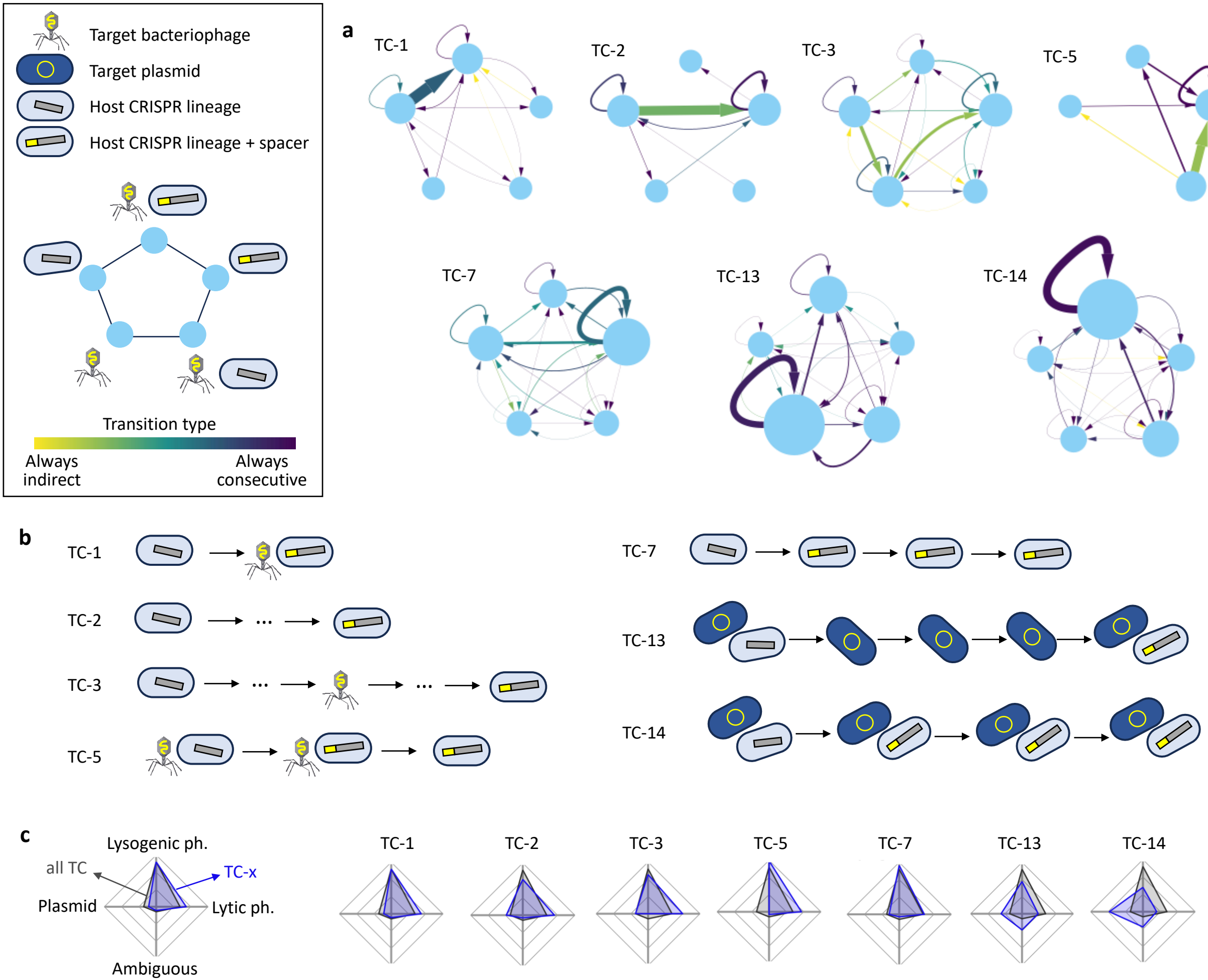
800 **Figure 5:** Temporal dynamics of MGE and CRISPR arrays. (a) Network representation of 7 of the
801 most frequent trajectory clusters (TC) representing potential immune adaptation in the host
802 population. Each node represents a state, defined by the presence or absence of a spacer, its
803 cognate MGE, and the CRISPR lineage in which the spacer was acquired (legend in the upper
804 left box). Node sizes are proportional to the number of times that a state is observed in a
805 trajectory. Arrow widths represent transition probabilities between states and the color scale
806 indicates the fraction of direct vs indirect transitions (direct transitions correspond to two
807 consecutive samples, indirect transitions involve intermediate samples in which neither the
808 MGE nor the CRISPR lineage were observed). (b) Simplified representation of typical
809 trajectories associated with the TC in (a). (c) Relative association of lytic (or non-lysogenic)
810 phages, lysogenic phages, 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
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
Lysogenic targets (%)
45
35
60
55
60
40
60
45
60
40
leading
middle
distal
Position in CRISPR array


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
TC-2
TC-3
TC-5
TC-7
TC-13
TC-14
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