CRISPR-Cas immune repertoires as an ecological record of bacterial interactions with mobile genetic elements in the human gut

Ekaterina Avershina[1,2,*], Einar E. Birkeland[1], Cecilie Bucher-Johannessen[1,2], and Trine Ballestad Rounge[1,2,3,*]

1 Department of Pharmacy, University of Oslo, Oslo, Norway

2 Department of Tumor Biology, Oslo University Hospital, Oslo, Norway

3 Cancer Registry of Norway, Norwegian Institute of Public Health, Oslo, Norway

*Corresponding authors: Ekaterina Avershina & Trine B. Rounge, Dept. of Pharmacy, University of Oslo; PO box 1068, Blindern, 0316 Oslo, Norway; Emails: ekateria@uio.no & t.b.rounge@uio.no

Abstract

Bacteria in the human gut influence host physiology and disease risk, but their ecology is strongly shaped by mobile genetic elements (MGEs) such as phages and plasmids. Past interactions between bacteria and MGEs can be inferred from CRISPR-Cas cassettes, which contain short DNA fragments derived from invading elements. To lay the groundwork for research on the impact of such interactions on the human host, we examined bacteria, MGEs, and CRISPR-Cas in the gut microbiome. Using fecal shotgun metagenomes from 1034 Norwegians, we constructed an extended microbiome resource comprising 1.7K prokaryotic mOTUs, 19.5K viral vOTUs and 24.2K plasmid PTUs. We also recovered 74.2K unique CRISPR-Cas cassettes to map past bacteria-MGE interactions and assessed their associations with human diet and lifestyle factors.

CRISPR-Cas spacers, and which viruses and plasmids they targeted, varied substantially within bacterial species, but were predominantly directed towards cohort-specific MGEs. Moreover, bacteria were more likely to target MGEs present in the same sample, consistent with local exposure. Bacteria also shared more targets within taxonomic families than across families, where mobilizable plasmids were more frequent among the targets. We did not find evidence that CRISPR-Cas spacers were related to characteristics of the human host, beyond the host-bacteria associations.

Together, this research provides a large-scale resource and a structured analysis of bacteria-MGE interactions in the gut microbiome, and their contribution to microbial ecosystem dynamics.

Keywords: gut microbiome, gut mobilome, virus, plasmid, CRISPR-Cas, CRISPR spacer, mobile genetic elements

Introduction

Human gut microbiomes are one of the most densely populated microbial environments, under intense ecological pressure and with frequent interactions between its inhabitants. To survive in such a competitive environment, bacteria must continuously adapt to selective pressures imposed by host immunity, nutrient availability, and other microbes. To respond to these fluctuating pressures, bacteria commonly exploit horizontal gene transfer, especially via plasmids and phages.[1,2]

At first glance, plasmids and phages seem to play opposite roles in the bacterial lifecycle, - plasmids often increase bacterial fitness, whereas phages exploit bacterial machinery for self-replication. Each, however, represent a mobile genetic element (MGEs) that carries both opportunity and risk for the bacterial host.[3,4] On one hand, they drive adaptivity by disseminating genes involved in biofilm formation, virulence, toxin-antitoxin defense systems, and antibiotic resistance.[5–8] On the other hand, they exert strong selective pressure through metabolic burden or cell death. To mitigate these antagonistic interactions, bacteria employ CRISPR-Cas immune system that targets invading and potentially deleterious DNA.[9,10]

The CRISPR-Cas immune system captures short DNA fragments from invading MGEs and retains them as spacers within the repeat-spacer-repeat arrays to rapidly recognize and fight off future invasions.[9] These records can provide a window into the microecological context in which bacteria persist, including MGE populations they interact with.[11] Variation in such contexts can translate into functional differences among closely related bacteria, including altered metabolic and physiological activity.

A substantial body of evidence links the composition and functional potential of the gut microbiome with human health outcomes, including gastrointestinal disorders, colorectal cancer, and a range of extraintestinal diseases.[12–15] Although studies linking pathogenicity, CRISPR-Cas immunity, and MGEs do exist, direct mechanistic links between these processes and human health remain very limited.[16]

Despite the conceptual relevance of CRISPR-Cas systems for studying microbial interactions, they have barely been applied to investigate gut bacteria-virus-plasmid relationships in an integrated manner. We retrieved only three reports in PubMed that covered CRISPR-Cas interactions between gut bacteria and viruses[17–19], and one that additionally touched upon plasmids[11]. Such paucity of studies reflects both methodological gaps and limited MGE databases, which hinder reliable spacer-target assignment and constrain inference of bacteria-MGE interactions.[20] Consistent with this limitation, the largest database of CRISPR-Cas spacers available to date, SpacerDB, identified putative targets for only 10% of the spacers.[20]

To lay the groundwork for understanding the ecology of gut microbiome beyond its single constituents and its impact on human health, we characterize bacteria, viruses, plasmids, and CRIPSR-Cas cassettes from stools of 1034 adult individuals (Figure 1). We identify 140K novel CRISPR-Cas spacers, 17.4K novel plasmids, add bacterial host inference for 6K unknown-host or novel plasmids and provide these as a resource for the scientific community. We use this database to examine bacteria-MGE interactions by utilizing a historical record of CRISPR-Cas spacer repertoire, and to explore their associations to human host factors.

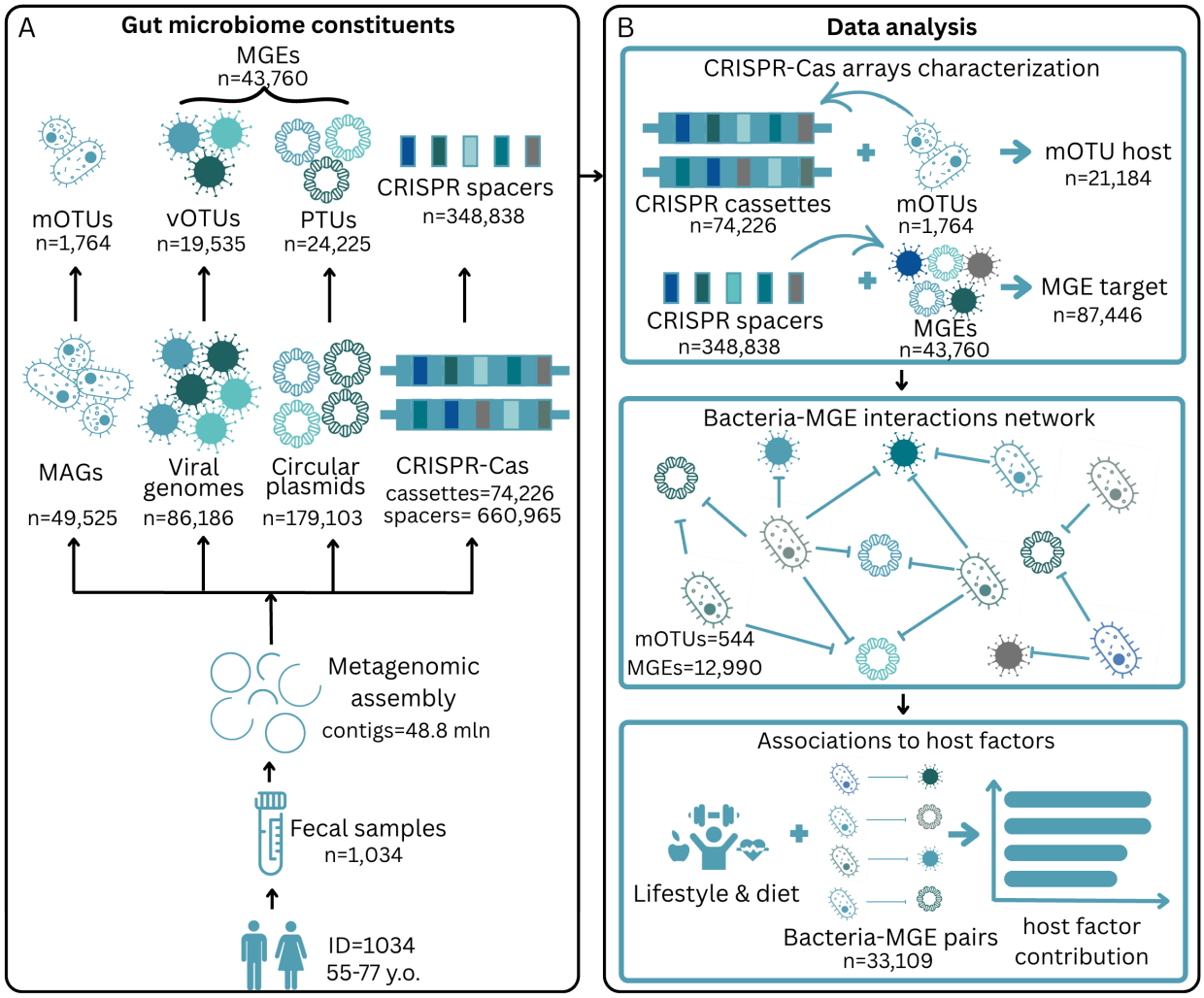

Figure 1. Overview of the study. A). Gut microbiome resource generation. The CRCbiome dataset comprises fecal samples from the Norwegian population (n=1034) aged 55-77 residing in South-East Norway. The samples were collected using a Fecal Immunochemical Test sampling kit. Shotgun metagenomic reads from these samples were assembled into 48.8 mln contigs. Prokaryotic metagenomic-assembled genomes and circular plasmids were generated from the contigs, and viral genomes and CRISPR-Cas cassettes were extracted. Prokaryotic genomes, viral genomes and circular plasmids were dereplicated to mOTUs, vOTUs and PTUs respectively based on sequence identity (mOTUs: 95%; vOTUs: 97.5%; PTUs: 90%). B). Data analysis. Bacteriome immunity towards vOTUs and PTUs was identified via CRISPR-Cas spacers' target identification. The bacterial host was assigned if the cassette was located on a contig belonging to mOTUs. mOTU-vOTU-PTU networks were constructed, and MGE targets and mOTU-MGE interaction pairs were compared between individuals across different demographic and lifestyle factors. MGE: mobile genetic elements

Materials and methods

All data analyses were performed using *pandas* v1.5.3, *numpy* v1.26.3, *statsmodels* v0.14.1, *scipy* v1.10.1 libraries, and visualized using *seaborn* v0.13.2 in Python v3.9.6 using Spyder IDE v5.3.0 unless stated otherwise.

Study cohort

The dataset was previously described in[14,21]. The study used samples from the CRCbiome study, which took place between 2017 and 2022 and recruited participants from the FIT-arm of the Bowel Cancer Screening in Norway trial. All individuals were between the age of 55-71 and had tested positive for occult blood in their FIT sample (Eiken Chemicals Ltd., Tokyo, Japan) and were subsequently invited for colonoscopy. FIT samples were collected at home and shipped to the testing facilities (delivery time estimated 3-10 days). The leftover FIT-buffer was subjected to DNA extraction using the QIAsymphony automated system (Qiagen, Hilden, Germany) according to the manufacturer's protocol with an additional off-board lysis procedure. Library preparation was performed with the Nextera DNA Flex Library Prep Reference Guide (Illumina, CA, USA) but scaling down to ¼ of the reference volume. Sequencing was performed on the Illumina NovaSeq X. In total, metagenomic profiles from 1034 individuals with a sequencing depth ≥1 Gbp were generated.

Additionally, participants filled in a Lifestyle and Demographic Questionnaire (LDQ) and a Food Frequency Questionnaire as described previously in Kværner et al.[22] Relevant for this study, the LDQ contained questions regarding demographic factors (national affiliation, education, employment status, and marital status), and lifestyle factors (antacids use, smoking and snus habits, BMI, and physical activity level). The FFQ assessed habitual dietary habits over the preceding year. Based on the LDQ and FFQ, a healthy lifestyle index was calculated as described in Kværner et al.[23] Prescription antibiotic use for each participant was obtained from the Norwegian Prescription Database (NorPD) and defined as having one or more prescriptions corresponding to any of the ATC codes J01, A07A, or P01A, four months prior to stool sampling.

Metagenome assembled genomes dataset generation

Sequencing data was processed using metagenome ATLAS[24] (v2.4.3) for assembly-based *de novo* generation of metagenome-assembled genomes (MAGs) and selection of representative genomes (mOTUs), as briefly described in[14]. Here, quality filtering was applied to paired-end sequencing reads using the BBMap[25] (v37.99). Assembly was carried out using metaSpades[26] (v3.13), using automatic k-mer size selection. Contigs were binned for the generation of MAGs using MetaBat[27] (v2.2) and MaxBin[28] (v2.14), using a minimum contig size of 1,500 and 1,000 bp, respectively. DAS Tool[29] (v1.1) was used to resolve a non-redundant set of bins, using a score threshold of 0.5. Microbial community types for the samples were extracted from Birkeland et al[14], where they were calculated using Dirichlet Multinomial Mixture models based on the bacterial relative abundance.

Viral dataset generation

Viruses were recovered from ATLAS-generated contigs as described in Istvan et al.,[21] with one notable modification. Unlike the previous publication, where only VirSorter2[30] was used for viral sequence identification, here viruses were recovered using both VirSorter2[30] (v2.2.2) and geNomad[31] (v1.7.0). Only viruses identified by both were included in the final dataset. In case a viral sequence was identified within a longer contig, or if its location did not correspond between the two tools, the shorter sequence was assigned as viral. Viral genome sequences assigned a medium quality or higher (at least 50% genome completeness) as determined by CheckV[32] (v0.8.1) were retained and dereplicated using Galah[33] (v0.3.1), selecting the highest quality genome of the shortest possible length as the representative viral operational taxonomic unit (vOTUs). Taxonomy was determined using vConTACT2[34] (v0.11.0) with the INPHARED database[35] (version 01.05.2023). This analysis is a modified version of the VirMake pipeline[36].

Plasmid dataset generation

Circular plasmids were assembled with Sequence Contents-Aware Plasmid Peeler (SCAPP[37] v 0.1.4) using minimal plasmid length of 1000 bp (default settings). Metagenome assembly graphs and BAM alignment files generated by read mapping to contigs were used as input. Assembled plasmids were then dereplicated into PTUs with *dedupe.sh* function from BBMap[38] (v38.96) on 90%, 95% and 99% identity levels. After inspection, a 90% dereplication level was retained. Sequences matching the resulting PTUs were identified in the PLSDB database[39] (version 10.02.2022) and the IMG/PR database[40] (version 08.08.2023) using BLAST (v2.12). The results were then filtered, keeping only BLAST hits with ≥90% identity over at least 20% of the PTU length. Metagenomic reads from each sample were then mapped to a dereplicated plasmid database for relative abundance estimation. The plasmid was regarded as detected if metagenomic reads were mapped onto at least 75% of its length with a minimum of 90% identity. PTU mobility was derived using MOB-suite[41] v3.1.9 typing, and antibiotic resistance genes were identified using RGI[42] v6.0.5, allowing only strict and perfect hits. The workflow was organized into a pipeline using Snakemake[43] v5.8.2.

CRISPR-Cas dataset generation

CRISPR-Cas cassettes and their spacers were identified using CCTyper[44] v1.8.0 with default settings in all metagenomic contigs and dereplicated PTU and vOTU datasets. CRISPR spacers were dereplicated at 99% identity level using cd-hit v4.8.1 and searched against local vOTU and PTU datasets, and against public databases of CRISPR spacers (CRISPR-Casdb[45], version 06.04.2022, and SpacersDB[20], version 10.05.2025), phages (INPHARED[35], version 01.05.2023) and plasmids (PLSDB[39], version 10.02.2022, and IMG/PR[40], version 08.08.2023) using *blastn-short* in BLAST[46] suite v2.12 with a minimum of 99% identity on at least 95% of the spacer's length, to detect their potential targets. To avoid self-mapping, vOTUs and PTUs that contain CRISPR-Cas cassettes were excluded. The workflow was organized into a pipeline using Snakemake[43] v5.8.2.

Rarefaction analysis for CRISPR-Cas spacer representativeness in cohort-specific and public databases was performed by randomly subsampling the databases with an increment of 1000 records until the size of the database was reached, 10 replicates per increment were generated. Individuality of spacers with regards to their targets was calculated as a median difference between the number of MGE targets present in an individual and the number of MGE targets present in ten other randomly selected individuals.

mOTU-MGE interactions

In case CRISPR-Cas spacer with a known vOTU or PTU target was detected on a contig belonging to a known mOTU, the interaction between this mOTU and the target was recorded. All observed encounters of interactions between mOTUs and MGEs were then organized into a network. The observed mOTU-MGE network was visualized in Cytoscape[47] v3.10.3. Pairwise mOTU connectedness was calculated based on presence of common CRISPR spacer targets between mOTUs.

To assess whether the observed bacteria-MGE interaction structure deviated from random expectations, we generated degree-preserving null networks using a bipartite configuration model. Weighted bacteria-MGE interaction counts were expanded into observation-level edges and randomized by double-edge swaps that preserved the total number of observations per mOTU and per MGE. The null model was initialized using 5 swaps per edge (burn-in phase) to ensure adequate randomization. The randomized edges were subsequently re-aggregated into weighted interaction pairs, and the number of unique

bacteria-MGE pairs, interaction multiplicity distributions, and edge concentration metrics (Gini coefficient and Herfindahl-Hirschman index) were compared between the observed network and the null distributions using 200 permutations.

Statistical analyses

Kruskal-Wallis test and ordinary least squares linear regression adjusted for sequencing depth were used for continuous data statistical analysis. Chi-square test and binomial distribution test were employed for categorical data analysis. Relative abundance correlations between mOTUs and MGEs were assessed using Spearman correlation. Only taxa with ≥20% prevalence in a domain-specific dataset were included. In case two taxa within a domain had ≥99% correlation to each other, only one of them was retained.

Comparative analysis and visualizations of Jaccard distances were performed in R (v4.2.0). Inter-individual similarity in MGE targeting was visualized using principal coordinates analysis (PCoA) and plotted with ggplot2 (v3.3.5). Difference in Jaccard distances was assessed using PERMANOVA implemented in the *adonis2* function of the vegan package (v2.6-2), using 999 permutations and the *by = "margin"* option to evaluate the independent contribution of each variable. The test was adjusted for the sequencing depth. In addition, a set of lifestyle, diet, and demographic variables' contribution to mOTU sharing of MGEs was evaluated using individual PERMANOVA models. In this analysis, only individuals with at least one known mOTU-MGE pair were included (n=1031). Prior to analyses, missing values for host factors were imputed using the median for continuous variables and coded as missing for categorical values. Each model was adjusted for sequencing depth and microbial community type. The resulting R^2 and p-values were summarized and visualized in barplots.

Results

To assess bacterial immunity towards MGEs and its associations with human host factors, we constructed an extended microbiome resource of gut prokaryotes, viral and plasmid MGEs and CRISPR-Cas arrays from a CRC screening cohort[22] comprised of 582 men and 452 women with a mean age of 67.0 ± 5.9 years. Prescription drug registry records showed that 141 individuals (13.6%) were prescribed antibiotics within 4 months prior to stool sampling. The metagenome dataset had an average sequencing depth of 3.3 ± 0.9 Gbp (12.0M ± 3.7M QC reads) per sample and was assembled into 48.8 million contigs with a mean length of 3,516 bp (SD=11,783 bp; min=500 bp; max=1.9 Mbp).

We have previously published a detailed description of read-based bacterial and archaeal composition of these samples in Birkeland *et al.*[14], and of viral composition in Istvan *et al.*[21], whereas plasmids and CRISPR-Cas were not previously reported. Here, we recovered 49,525 prokaryotic metagenome-assembled genomes dereplicated into 1,764 mOTUs (Bacteria, n=1,752; Archaea, n=12), and updated virus recovery using a combination of two tools for identification of viral sequences, resulting in 19,535 vOTUs, derived from 86,186 viral genomes. we also recovered 179,103 circular plasmids using SCAPP[37] and dereplicated them into a set of 24,225 plasmid taxonomic units (PTUs). A detailed description of the prokaryome and virome can be found in Supplementary Text 1, and of the plasmidome in Supplementary Text 2. Additionally, pairwise diversity and relative abundance correlations between the microbiome constituents, as well as the overall microbial community composition assessments, are provided in the Supplementary Text 3.

208 *CRISPR-Cas cassettes exhibited high bacterial intraspecies variability*

209 A total of 74,226 CRISPR-Cas cassettes were detected across all individuals (71±22 cassettes per sample),
210 with the number of cassettes per sample largely determined by sequencing depth and assembly quality
211 (Supplementary Figure 1A). Each CRISPR-Cas cassette comprised on average 12 repeats (SD=12; min=3;
212 max=319) (Figure 2A), and the detection of the *cas* operon in addition to the repeat-spacer-repeat array
213 depended on the contig length (Supplementary Figure 1B).

214 Of the total identified CRISPR-Cas cassettes, 21,184 (28.5%) were located on contigs belonging to
215 recovered genomes assigned to mOTUs. Overall, 916 mOTUs (51.9%) had at least one genome with an
216 identified CRISPR-Cas cassette, and both the frequency of detection (Figure 2B) and the cassette length,
217 i.e. number of CRISPR-Cas repeats (Supplementary Figure 1C), varied between bacterial families. There
218 were 14 taxonomic families, with 1-3 mOTUs each, where no CRISPR cassettes were recovered
219 (Supplementary Table 1). Since CRISPR-Cas cassettes have been reported to be encoded in plasmids and
220 viruses[48,49], we additionally searched for CRISPR-Cas cassettes in local PTU and vOTU sets. We found only
221 0.05% of PTUs (n=134) and 0.05% of vOTUs (n=109) to contain CRISPR-Cas cassettes.

222 For stringent identification of CRISPR spacers, we proceeded only with spacers from cassettes with a
223 length of ≥10 repeats. This filtering excluded 91% of putative CRISPR-Cas cassettes and 24.3% of all
224 identified spacers. Of the 660,965 remaining spacers, dereplication resulted in 348,838 unique CRISPR
225 spacer clusters (average length 33±3 bp), 59.8% (n=208,446) of which were previously reported in the
226 SpacerDB database (Figure 2C). Most dereplicated CRISPR spacers were singletons (n = 264,014, 75.7%),
227 and most non-singleton spacers tended to be shared between individuals (n=81,792, 96.4%). Still, only 21
228 of those spacers were detected in >=10% population. Most non-singleton CRISPR spacers with known
229 mOTU host (n=55,088) were detected in MAG bins belonging to the same mOTUs (n=51,535; 93.5%), and
230 only 6.5% of them were detected in multiple mOTUs, mostly belonging to the same taxonomic family
231 (Supplementary Figure 1D).

232 To assess the conservation of CRISPR-Cas cassettes within prokaryotes, we searched for shared spacers
233 between MAGs within mOTUs with at least 10 CRISPR-Cas containing MAGs. Generally, mOTUs
234 demonstrated high degree of heterogeneity in CRISPR spacers (Figure 2D). The one striking exception was
235 mOTU1812 *Bifidobacterium animalis*, that shared most of its spacers in at least 30 of its 33 MAG genome
236 bins. Such high conservation of spacers likely points towards a common origin of this strain. *B. animalis* is
237 often added as a probiotic species in dairy products. Individuals with the *B. animalis* genome detected,
238 had higher daily consumption of local fermented milk products which have *B. animalis* as a supplement
239 (Supplementary Figure 2E, Kruskal-Wallis Hst=85.9, $p=1.9*10^{-20}$), corroborating the recovery of a likely
240 foodborne strain.

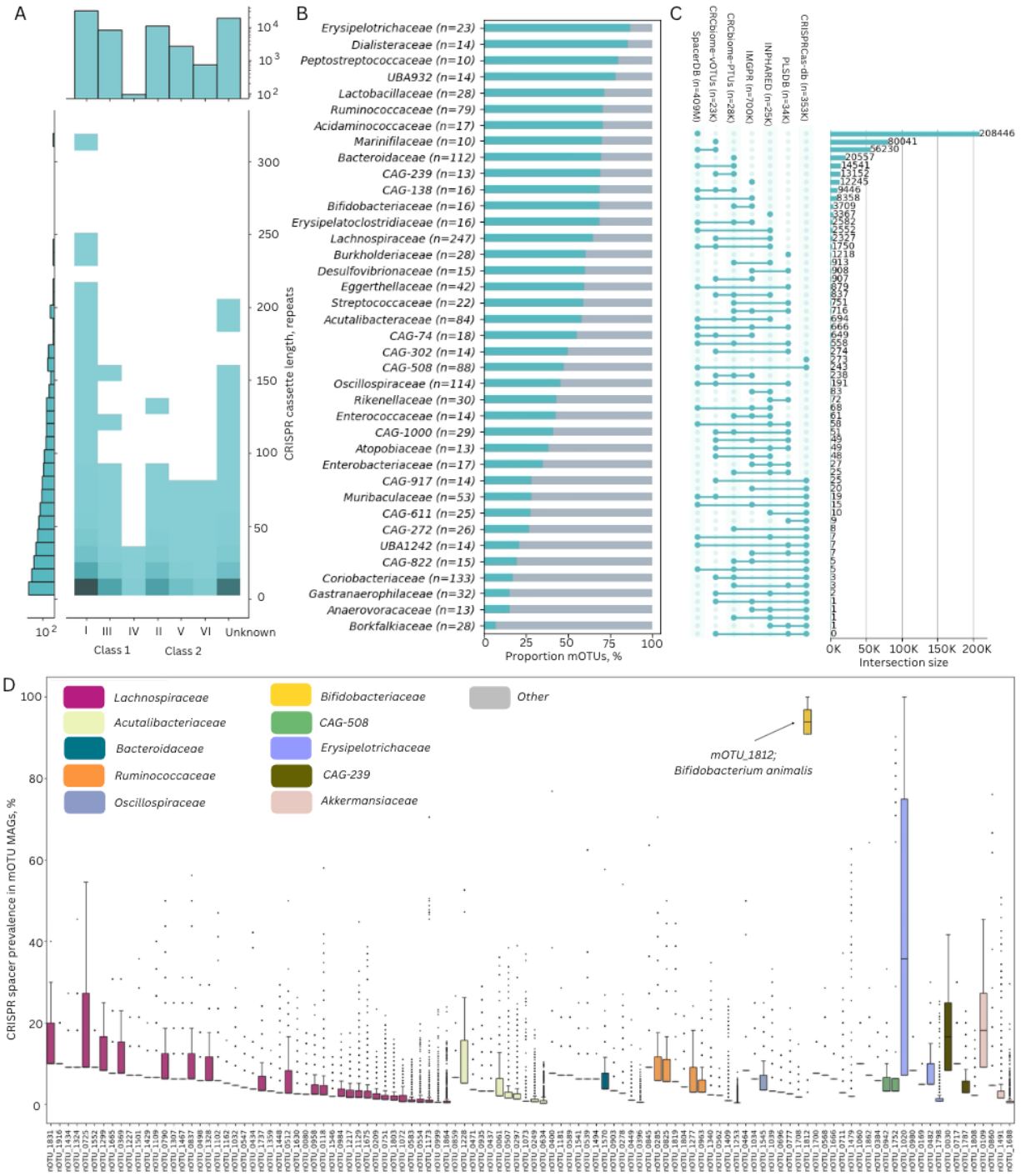

Figure 2. CRISPR-Cas dataset overview. A) Frequency of detection of CRISPR-Cas cassettes of each class and type, and number of repeats in these cassettes. B) Proportion of mOTUs belonging to a given taxonomy class where CRISPR-Cas cassettes were detected in at least one MAG (blue) or not detected in any MAG (grey). Only families with ≥10 mOTU representatives are depicted C) Number of CRISPR-Cas spacer clusters with ≥99% homology to records in public databases and to vOTUs and PTUs detected in the study. D) Heterogenity of CRISPR-Cas spacers in mOTUs. For each mOTU, the proportion of mOTU-MAGs that contained each CRISPR-Cas spacer detected in this mOTU, is depicted. All mOTUs with at least 10 MAGs, where CRISPR-Cas cassettes were detected, are included in the analysis.

249 *CRISPR-Cas spacers targeted individual and population-specific MGEs, varying between microbial*
250 *community types*

251 To identify which MGEs CRISPR-Cas spacers targeted, we searched for homologous protospacers in public
252 virus and plasmid databases, and the CRCbiome dataset-specific MGEs (Figures 2C, 3A).

253 A CRCbiome vOTU or PTU target was identified for 25.1% of CRISPR non-redundant spacers (n = 87,446),
254 with a mean of 10 CRCbiome-MGE targets per spacer (SD=9, min=1, max=84). Spacers were more likely
255 to target vOTUs or PTUs present in the same individual than those present in others (Kruskal-Wallis,
256 $p<1*10^{-10}$, Figure 3B), likely reflecting adaptation to an intra-individual gut environment. To assess
257 whether there was an enrichment of targets in the CRCbiome dataset, we performed a rarefaction analysis
258 excluding viruses and plasmids detected in the same individual from the matching dataset. MGEs present
259 in other samples from the CRCbiome dataset were more commonly targeted than MGEs in unrelated
260 public databases, potentially suggesting population specificity in addition to individual specificity (Figure
261 3A).

262 As with spacers, we also observed heterogeneity in MGE targeting at the mOTU level, assessing all mOTUs
263 with at least 10 CRISPR-Cas containing MAGs. Each MGE was on average targeted by 9.8% of MAGs within
264 each mOTU (SD=12.0%), and none were consistently targeted by all MAGs within any mOTU. At most,
265 41/47 (87.2%) CRISPR-containing genomes of the uncharacterized species UBA sp900544375 of the
266 *Acutalibacteriaceae* family targeted the unknown virus vOTU-CB-12066. In addition, a set of viruses
267 classified as *Akkermansia phage DTMo-2021a* were targeted by spacers detected in ~85% of *Akkermansia*
268 sp001580195 and *Akkermansia muciniphila* B genomes.

269 Overall, vOTUs where at least one, but not all, viral genomes were integrated into bacterial host genomes,
270 were more frequently targeted than those vOTUs where viral genomes were either never or always
271 integrated (Figure 3C). Similarly, PTUs with no genes enabling plasmid mobility or conjugation (non-
272 mobilizable PTUs) were more frequently encountered among CRISPR spacer targets than those that had
273 such genes (Figure 3D).

274 CRISPR-Cas targeting of MGEs reflects prior exposure to such MGEs, and the collective record of targeted
275 elements within a sample could be seen as a reflection of long-term exposure. We hypothesized that the
276 collective record of MGE exposure encoded by CRISPR spacers would be determined by which bacteria
277 harbor them. We therefore evaluated to what extent microbial community states derived from
278 prokaryote abundances (defined in Birkeland et al.[14]) were associated with differences in the overall
279 pattern of MGE targeting in a sample. Although overall MGE targeting by CRISPR-Cas did not significantly
280 reflect the community states, individuals enriched in either *Clostridia* or *Methanobrevibacter smithii,*
281 resembled other members of the same group more than individuals from other community states (Figures
282 3E & 3F). Interestingly, individuals with the *Bacteroides* spp.-enriched microbial community state showed
283 an opposite pattern, potentially reflecting higher diversity among the dominant bacterial species.

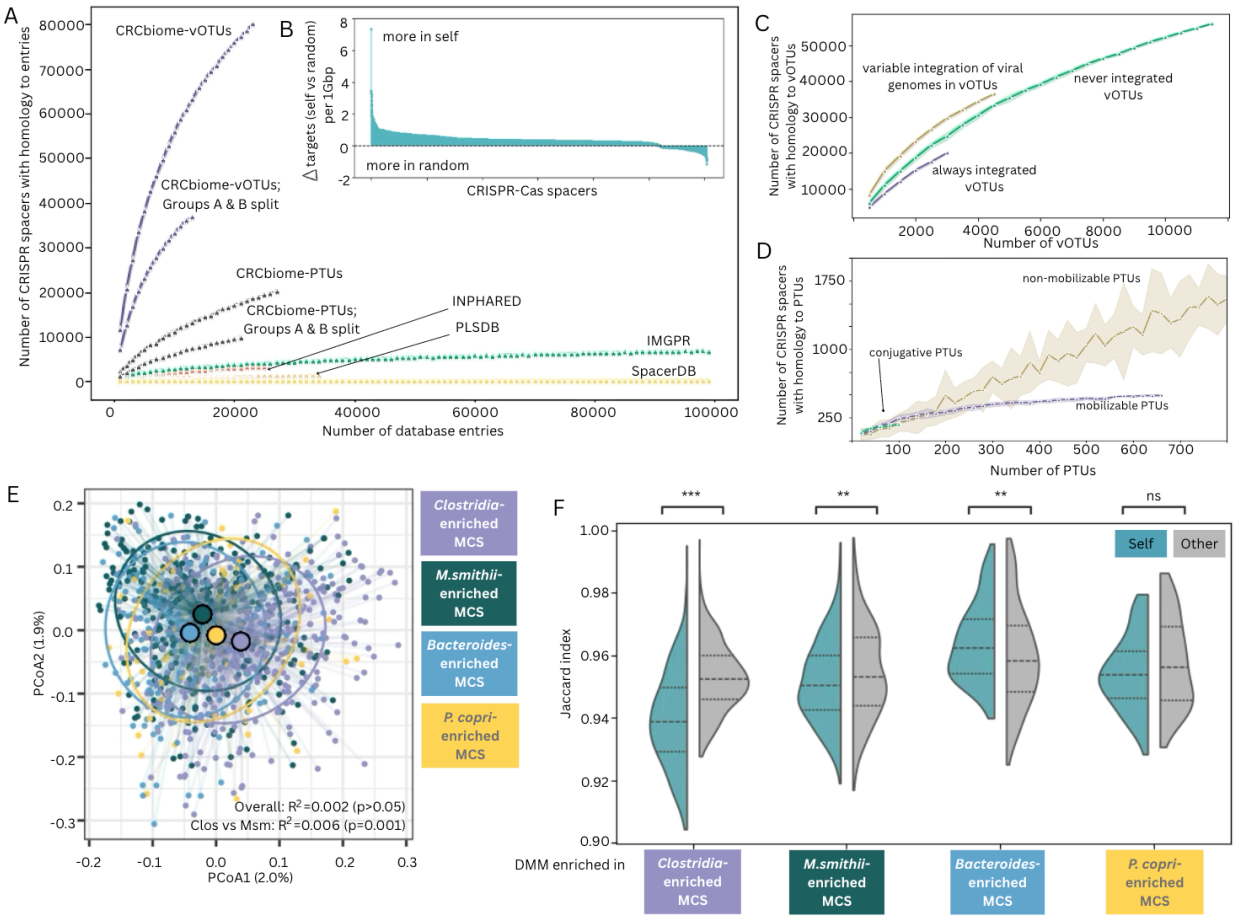

*Figure 3. Targeting of MGEs by CRISPR-Cas spacers. A) CRISPR-Cas spacer clusters are better represented in the CRCbiome viral and plasmid datasets than in public viral, plasmid, and spacer databases. Rarefaction curves were calculated as the number of spacers with a homology to a sequence in a random subset of database sequences, using the mean of 10 randomizations and 95% confidence interval (shading), at increments of 1,000 database entries. To assess population specificity in MGE targeting, the CRCbiome individuals were randomly split in two equally sized groups, Groups A and B. Spacers detected in samples from Group A were then matched against the vOTUs and PTUs detected by read mapping in samples from Group B. The splitting was performed ten times for the calculation of the mean and of confidence intervals. B) Individuality of CRISPR-Cas targeting was assessed using the average difference in the number of MGE spacer targets detected in self vs other 10 randomly selected individuals. To account for sequencing depth, the number of detected targets in each sample was normalized to targets per 1 GBp sequencing data prior to difference calculation. For visualization purposes, only spacer clusters with targets detected either in self or in random individuals are included (n = 9,534, 10.9%); C) Enrichment of vOTUs by their host genome integration ability based on rarefaction analysis with 500 vOTUs increment. The average per 10 randomizations and 95% confidence interval (shading) are presented. D) C) Enrichment of PTUs by their mobility status based on rarefaction analysis with 25 PTUs increment. The average per 10 randomizations and 95% confidence interval (shading) are presented. (E) PCoA analysis of MGE targeting similarities between individuals contrasted by microbial community state defined in Birkeland et al.[14] The similarity was assessed using Jaccard distance based on the fraction of spacers that shared common MGE targets. R^2 and p values are calculated using PERMANOVA analysis. MCS: microbial community state; Clos: Clostridia-enriched MCS; Msm: M.smithii-enriched MCS. (F) Difference in MGE targeting similarities between individuals within the same (blue) or different (grey) microbial community states defined in Birkeland et al[14] The similarity was assessed using Jaccard distance based on the presence of spacers with common MGE targets, and varies from 0 (same targeted MGEs) to 1 (no overlap in targeted MGEs). MCS: microbial community state. ***p≤0.001; **p≤0.01; ns: not significant, p>0.05*

307 *Bacteria tended to share CRISPR-Cas MGE targets with other bacteria belonging to the same taxonomic*
308 *families*

309 Both plasmids and phages show non-random bacteria host association, and host range spans a continuum
310 from highly restricted strain-specific phages to relatively generalist broad-host-range plasmids or
311 polyvalent phages.[50–52] CRISPR-mediated targeting of MGEs provides evidence of prior infections or
312 contact between bacteria and MGE, and we therefore set out to elucidate mOTU-MGE networks based
313 on these patterns of recognition.

314 There were 146,256 non-redundant CRISPR-Cas spacers with at least one representative located on
315 contigs belonging to recovered mOTUs. Out of these, 35,887 (24.5%) had targets identified among
316 CRCbiome vOTUs and/or PTUs (excluding those with CRISPR-Cas systems). Overall, we recovered 27,062
317 unique mOTU-vOTU and 6,047 mOTU-PTU target pairs with 544 mOTUs, 10,519 vOTUs and 2,471 PTUs.
318 This observed number of mOTU-MGE pairs is significantly lower than what would have been expected
319 under the pretense of random pairing of these constituents (expected=116,838±103, p=0.005). Moreover,
320 most of the observed mOTU-MGE pairs were encountered more often than expected (mean multiplicity
321 observed = 4.7±11.7 vs expected = 1.3±0.01; p=0.005), and with a high degree of inequality between the
322 pairs' detection (Gini inequality coefficient observed = 0.65 vs expected = 0.22±0.00; p=0.005), strongly
323 supporting selectivity in bacterial immunity towards specific MGEs.

324 Interactions defined by mOTU-MGE pairs tended to form subnetworks comprising taxonomically related
325 mOTUs sharing multiple vOTUs/PTUs (Figure 4). However, some bacteria, such as mOTUs belonging to
326 *Acutalibacteriaceae*, had similar connectivity to mOTUs from other families (Supplementary Figure 2).
327 vOTUs and PTUs were characterized using their integrative status and mobility traits. Viruses that were
328 found integrated in the host genome were targeted by a narrower range of bacteria than those without
329 evidence of integration (χ^2=73.7; $p<1*10^{-10}$). For plasmids, those that were mobilizable were targeted by
330 a broader range of bacteria compared to non-mobilizable or conjugative ones (χ^2=84.6; $p<1*10^{-10}$).

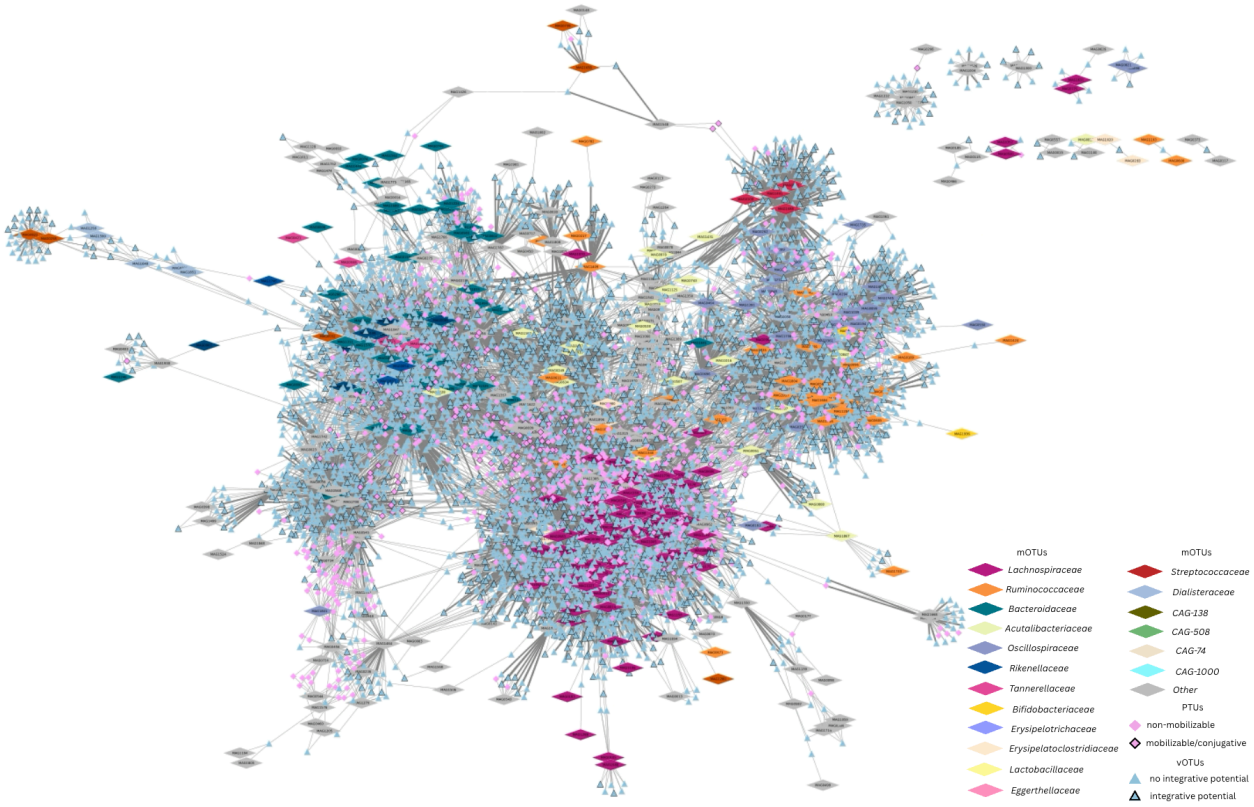

Figure 4. mOTU-MGE interaction network. Each diamond represents an mOTU colored by a taxonomic family. Blue triangle codes vOTUs, pink diamonds - PTUs. Each line represents a connection via a spacer, with the thickness depicting whether the connection was detected in ≥5 individuals (thick line) or <5 individuals (thin line). For visualization purposes, only interactions that connect at least 2 mOTUs, are depicted.

Prokaryotic immunity mirrors associations between microbial constituents and host factors

CRISPR-Cas spacers provide a record of past MGE invasions, thus extending the information obtained from a single microbiome snapshot to include recent history. In fact, the targets of most CRISPR-Cas spacers with a known MGE target were below the detection limit (n=78,482, 89.7%). Incorporating this signal could reveal gut microbiome-host factor associations that are otherwise overlooked. Furthermore, the presence of a shared target within the same mOTU may reflect similarities in the microenvironments occupied by its members. For two bacteria within the same mOTU to share a target, both must have been exposed to the same invading MGE, implying physical and ecological proximity.

To evaluate the environmental influence on mOTU sharing of MGE targets across the gut microbiome, we constructed a distance matrix by performing a pairwise comparison of the fraction of mOTU-MGE target pairs that were shared for each pair of participants. We tested whether mOTU-specific sharing of targets was associated with host demographic, lifestyle, or dietary variables and compared these patterns to associations with beta-diversity based on the presence or absence of mOTUs, vOTUs, and pOTUs. mOTU-MGE targeting-derived associations seemed to mirror individual microbial constituents' associations, especially that of prokaryotes, also when adjusting for the microbiome community state and sequencing depth (Figure 5). In line with this, no statistically significant associations were observed for any commonly targeted MGEs (MGEs targeted by at least 50 genomes within an mOTU, Supplementary Table 2), which suggests that prokaryote immunity *per se* is not influenced to a significant extent by lifestyle factors.

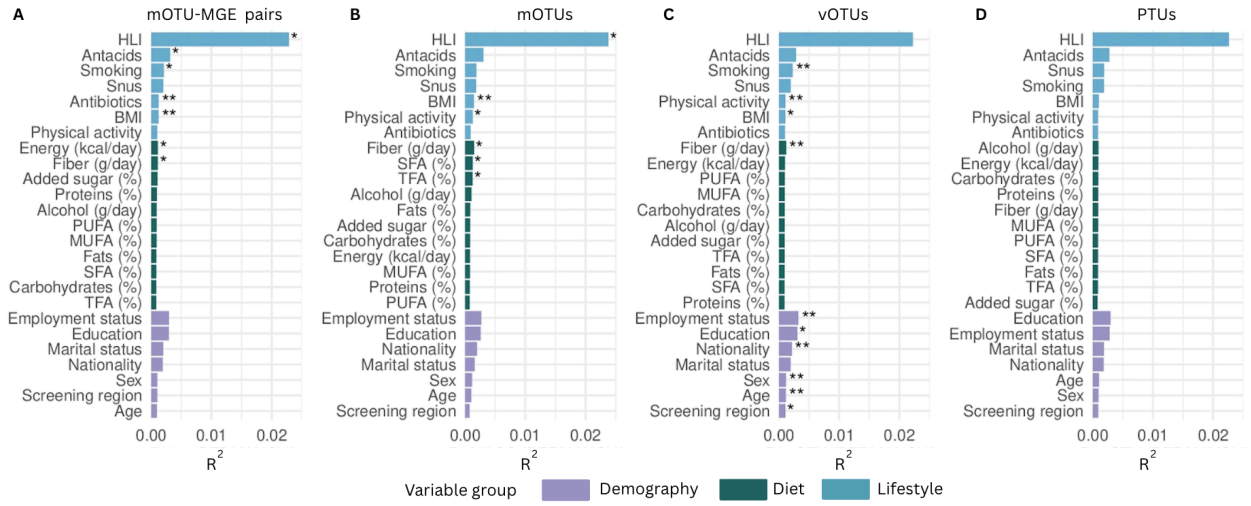

*Figure 5. Associations between individuals' microbial community similarities and human host factors. Similarities were assessed using Jaccard index calculated based on the presence/absence of (A) mOTU-MGE pairs (presence/absence of genomes within the same mOTUs with a CRISPR-Cas immunity towards the same MGEs); (B) mOTUs; (C) vOTUs and (D) PTUs. *PERMANOVA p<0.05; **PERMANOVA p<0.01. PERMANOVA analysis was adjusted for the sequencing depth and microbial community states. HLI: Healthy Lifestyle Index; BMI: Body Mass Index; SFA: Short-chain fatty acids; TFA: Trans Fatty Acids; MUFA: Monounsaturated fatty acids; PUFA: Polyunsaturated fatty acids; (%): Percentage of total energy intake per day.*

Discussion

In this work, we identified and characterized the prokaryome, virome, plasmidome, and CRISPR-Cas repertoires from 1,034 shotgun metagenomes, and present this as a comprehensive microbiome resource for the community to enable comparative analyses, identification of novel taxa and functions, and exploration of community-level and cross-entity interactions. Using this resource, we characterized bacteria-MGE interactions at a population and individual scale, and revealed a highly individualized landscape of bacterial immunity, which appears to reflect microbial ecology within the gut rather than host-associated factors.

Despite the rapid growth of public metagenomic sequencing, a substantial fraction of gut microbial diversity remains uncharacterized.[53] This "dark matter" is particularly pronounced for MGEs, where many sequences lack confident taxonomy and host annotation. Large-scale genome- and virome-catalog efforts continue to expand available reference space, yet they consistently report extensive novelty and limited host resolution for many recovered viral and plasmid sequences.[40,54] By leveraging CRISPR-Cas spacers as records of prior encounters between bacteria and MGEs alongside bacterial, viral and plasmid genomes, we provide empirical links between bacterial hosts and their associated viruses and plasmids, thereby offering host-range hypotheses at the species-, genus-, and family-level.

For a CRISPR-Cas spacer to be incorporated into a CRISPR array, several sequential events must occur: (i) an MGE must invade the bacterial cell, (ii) the CRISPR adaptation machinery must be triggered, and (iii) a segment of the invader's DNA must be sequentially integrated as a new spacer into the host CRISPR locus, where multiple regions of an invading genome can serve as potential protospacer templates.[55,56] As a consequence, each individual's CRISPR spacer repertoire reflects not only bacteriome variation but also personal history of exposures to distinct viral and plasmid MGEs.[57,58] Consistent with this, we observed marked interindividual variation in CRISPR spacer content, with most spacers being detected in only one or a few individuals. This high degree of personalization is likely a result of bacterial adaptation to local MGE exposure and could serve as a fingerprint of microevolutionary dynamics within the gut. A notable

386 deviation was the relatively conserved spacer repertoires in *Bifidobacterium animalis* genomes recovered
387 from 33 individuals. These individuals reported higher consumption of dairy products typically containing
388 *B. animalis*, consistent with diet-driven species acquisition.[59,60] Although spacer variability within this
389 species was markedly lower than that observed for other taxa, likely reflecting a common lineage,
390 detectable differences in CRISPR content were still present. This indicates that even shared, diet-
391 associated strains undergo individual-specific diversification following colonization. Apart from this
392 observation, we did not observe other host lifestyle-related patterns in MGE targeting by bacteria.

393 The pronounced bacterial adaptivity to local MGE exposure was further supported by the higher detection
394 rate of MGE targets in the cohort-specific database compared to public repositories. Cross-kingdom
395 network analyses are typically inferred from co-occurrence patterns derived from relative abundances,
396 reflecting predominantly indirect or environmentally driven associations rather than biological
397 interactions.[61] In contrast, CRISPR-Cas spacers provide records of prior encounters between bacteria and
398 MGE, and enable the network reconstruction grounded in direct molecular evidence. We found that
399 bacteria exhibited strong selectivity in incorporating CRISPR-Cas spacers targeting specific MGEs, with
400 bacteria within the same taxonomic families tending to share overlapping MGE pools.

401 By contrast, mobilizable plasmids and non-integrated viruses, exhibited interactions indicative of a
402 broader bacterial host range. These genetic elements are intrinsically more 'selfish' and likely selected for
403 maximizing horizontal transmission across diverse hosts, rather than long-term alignment with any single
404 host lineage, whereas prophages, and possibly non-mobilizable plasmids rely more strongly on their hosts'
405 fitness.[62,63] Consistent with this, CRISPR-Cas targeting varied by MGE lifestyle and mobility. Phages with
406 intermittent integration into bacterial genomes were more frequently targeted than those that were
407 always integrated or never integrated. This observation aligns with the report that bacteria more readily
408 acquire spacers against viruses entering a lysogenic state upon infection, as a purely lytic phage infection
409 may leave insufficient time for spacer acquisition, while targeting a stable prophage carries the risk of
410 autoimmunity.[64] Similarly, non-mobilizable plasmids showed higher targeting rates than mobilizable or
411 conjugative ones.

412 Our study has several strengths. In addition to standard metagenomic assembly, we systematically
413 recovered viral genomes and independently assembled bacterial plasmids, enabling a high-resolution
414 characterization of MGEs. We further integrated these with CRISPR-Cas spacer profiles to investigate
415 bacterial adaptation to individual MGEs. Lastly, the inclusion of participants with extensively validated
416 dietary, lifestyle, and demographic data allowed us to explore associations between CRISPR-Cas immunity
417 and host-related factors. Several limitations should be acknowledged. First, FIT samples contain ~10 mg
418 of stool. While suitable for microbiome profiling, the limited input material, combined with the modest
419 sequencing depth, has likely reduced the recovery of genomes from low-abundance taxa.[21,65,66] Second,
420 we did not account for spacer chronology to distinguish between recent and more ancient bacteria-MGE
421 interaction events. Finally, our analysis focused on viruses and plasmids and did not include other classes
422 of MGEs that may also be targeted by CRISPR-Cas systems.

423 **Conclusion**

424 Integrating prokaryotic, viral, plasmid, and CRISPR-Cas data from >1000 individuals, we provide a
425 comprehensive population and individual-scale resource for investigating MGEs and their interactions in
426 the gut. This includes spacer-based bacterial host ranges for isolated MGEs. Our analyses reveal a highly
427 individualized and cohort-specific landscape of CRISPR-based immunity that reflects bacterial host
428 microenvironment more strongly than human host characteristics. CRISPR spacer-informed network
429 reconstruction uncovered selective MGE targeting and shared MGE pools within bacterial families that

430 would not be detectable through co-occurrence networks. Variation in host range based on MGE lifestyle
431 and mobility further suggests evolutionary trade-offs between MGEs' horizontal dissemination and long-
432 term host alignment. This work provides a valuable foundation for future investigations linking microbial
433 evolutionary dynamics and ecosystem stability.

434 **Data availability**

435 DNA sequencing data generated in this study are deposited in Federated EGA under accession code
436 EGAS50000000170. Per participant consent, submitted FASTQ files exclude reads mapping to the human
437 genome. Processing of data from this study must comply with the General Data Protection Regulation
438 (GDPR). Access can be obtained by following the procedure described here:
439 https://www.mn.uio.no/bils/english/groups/rounge-group/crcbiome/.

440 DNA sequences of the GM constituents detected in at least 5 individuals, and all related metadata are
441 provided in FigShare (doi:10.6084/m9.figshare.31707688), and scripts are available at
442 https://github.com/Rounge-lab/CRISPR_interactions. The associated metadata include genome length,
443 completeness, and the number of genomes used to construct each taxonomic unit. For viruses, we report
444 detected auxiliary metabolic genes, whereas for plasmids we provide information on mobility, antibiotic
445 resistance genes, and predicted bacterial hosts based on reference databases or relative abundance
446 patterns. We indicate whether CRISPR-Cas systems were identified in each mOTU. For MGEs, we report
447 whether they were targeted by recovered CRISPR-Cas spacers, whether they encoded CRISPR-Cas systems
448 themselves, and their spacer-inferred potential hosts. Additionally, we provide all detected mOTU-MGE
449 target pairs along with their frequency of occurrence.

450 **Author contributions**

451 EA, EB, TBR - conceptualization; EB, CBJ - data generation; EA, EB, CBJ - data curation, formal analysis &
452 visualization; EA - original draft writing; EA, EB, CBJ, TBR - draft review & editing; TBR- funding acquisition;
453 TBR - supervision & project administration

454 **Acknowledgements**

455 We would like to express our gratitude to Jan Inge Nordby from the Department of Medical Biochemistry
456 at Oslo University Hospital, Erik Natvig and Anita Jørgensen from the Department of Colorectal Cancer
457 Screening at the Cancer Registry of Norway (CRN), National Institute of Public Health (NIPH), and Vahid
458 Bemanian from the Pathology Department at the Akershus University Hospital for their contributions to
459 biobanking, laboratory work, and data management. We also wish to thank Harri Kangas and Pekka
460 Ellonen at the Finnish Institute of Molecular Medicine (FIMM) for their assistance in preparing sequencing
461 data; library preparation and sequencing were conducted at the FIMM Technology Centre, supported by
462 HiLIFE and Biocenter Finland. Additionally, we acknowledge the support of current and former group
463 members for their administrative and scientific contributions, including Maja Sigerseth Jacobsen, Paula
464 Istvan and Paula Berstad. The CRCbiome study used in this publication, was supported by the Norwegian
465 Cancer Society, projects 190179 (TBR), 297310 (TBR), and 198048 (Paula Berstad, Cancer Registry of
466 Norway, Norwegian Institute of Public Health, Oslo, Norway), and the South-East Norway Regional Health
467 Authority projects 2022067 (TBR) and 2020056 (TBR).

468 In this work, ChatGPT LLM was used for assistance with scripting, and to improve language and readability.
469 All AI-assisted text revisions and code suggestions were reviewed, evaluated and finalized by the authors.
470 Figure panels were assembled in *Canva*, and explanatory text and annotations were added without
471 altering the original graphs.

472 **Abbreviations**

473 CRISPR-Cas: Clustered Regularly Interspaced Palindromic Repeats (CRISPR) and CRISPR-associated
474 proteins (Cas)

475 IMG/PR: Plasmid database derived from the Integrated Microbial Genomes & Microbiomes

476 IQR: interquartile range

477 MAG: Metagenome-Assembled Genome

478 MCS: microbial community state

479 MGE: Mobile Genetic Elements

480 mOTU: metagenomic Operational Taxonomic Unit

481 PCoA: Principal Coordinate Analysis

482 PLSDB: The Plasmid Database by NCBI

483 PTU: Plasmid Taxonomic Unit

484 SD: standard deviation

485 vOTU: viral Operational Taxonomic Unit

486 **References**

487 1. Smillie, C. S. *et al.* Ecology drives a global network of gene exchange connecting the human
488 microbiome. *Nature* **480**, 241–244 (2011).

489 2. Hehemann, J.-H. *et al.* Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut
490 microbiota. *Nature* **464**, 908–912 (2010).

491 3. Rajer, F. & Sandegren, L. The Role of Antibiotic Resistance Genes in the Fitness Cost of Multiresistance
492 Plasmids. *mBio* **13**, e03552-21 (2022).

493 4. Borodovich, T., Shkoporov, A. N., Ross, R. P. & Hill, C. Phage-mediated horizontal gene transfer and its
494 implications for the human gut microbiome. *Gastroenterol. Rep.* **10**, goac012 (2022).

495 5. Tang, X. *et al.* Lysogenic bacteriophages encoding arsenic resistance determinants promote bacterial
496 community adaptation to arsenic toxicity. *ISME J.* **17**, 1104–1115 (2023).

497 6. Davies, E. V. *et al.* Temperate phages both mediate and drive adaptive evolution in pathogen biofilms.
498 *Proc. Natl. Acad. Sci.* **113**, 8266–8271 (2016).

499 7. Heo, S. *et al.* Staphylococcus equorum plasmid pKS1030-3 encodes auxiliary biofilm formation and
500 trans-acting gene mobilization systems. *Sci. Rep.* **13**, 11108 (2023).

8. Arredondo-Alonso, S. *et al.* Plasmid-driven strategies for clone success in Escherichia coli. *Nat. Commun.* **16**, 2921 (2025).

9. Wang, J. Y., Pausch, P. & Doudna, J. A. Structural biology of CRISPR–Cas immunity and genome editing enzymes. *Nat. Rev. Microbiol.* **20**, 641–656 (2022).

10. Upreti, C., Kumar, P., Durso, L. M. & Palmer, K. L. CRISPR-Cas inhibits plasmid transfer and immunizes bacteria against antibiotic resistance acquisition in manure. *Appl. Environ. Microbiol.* **90**, e00876-24 (2024).

11. López-Beltrán, A., Botelho, J. & Iranzo, J. Dynamics of CRISPR-mediated virus–host interactions in the human gut microbiome. *ISME J.* **18**, wrae134 (2024).

12. Hays, K. E., Pfaffinger, J. M. & Ryznar, R. The interplay between gut microbiota, short-chain fatty acids, and implications for host health and disease. *Gut Microbes* **16**, 2393270 (2024).

13. Fabi, J. P. The connection between gut microbiota and its metabolites with neurodegenerative diseases in humans. *Metab. Brain Dis.* **39**, 967–984 (2024).

14. Birkeland, E. E. *et al.* Microbiome signatures for detection of colorectal lesions in population-based FIT screening. 2025.10.06.25336873 Preprint at https://doi.org/10.1101/2025.10.06.25336873 (2025).

15. Wang, T. *et al.* Divergent age-associated and metabolism-associated gut microbiome signatures modulate cardiovascular disease risk. *Nat. Med.* **30**, 1722–1731 (2024).

16. Nozawa, T. *et al.* CRISPR Inhibition of Prophage Acquisition in Streptococcus pyogenes. *PLOS ONE* **6**, e19543 (2011).

17. Zolfo, M. *et al.* Discovering and exploring the hidden diversity of human gut viruses using highly enriched virome samples. 2024.02.19.580813 Preprint at https://doi.org/10.1101/2024.02.19.580813 (2024).

18. Monaghan, T. M. *et al.* Metagenomics reveals impact of geography and acute diarrheal disease on the Central Indian human gut microbiome. *Gut Microbes* **12**, 1752605 (2020).

19. Nayfach, S. *et al.* Metagenomic compendium of 189,680 DNA viruses from the human gut microbiome. *Nat. Microbiol.* **6**, 960–970 (2021).

20. Roux, S. *et al.* Planetary-scale metagenomic search reveals new patterns of CRISPR targeting. 2025.06.12.659409 Preprint at https://doi.org/10.1101/2025.06.12.659409 (2025).

21. Istvan, P. *et al.* Exploring the gut DNA virome in fecal immunochemical test stool samples reveals associations with lifestyle in a large population-based study. *Nat. Commun.* **15**, 1791 (2024).

22. Kværner, A. S. *et al.* The CRCbiome study: a large prospective cohort study examining the role of lifestyle and the gut microbiome in colorectal cancer screening participants. *BMC Cancer* **21**, 930 (2021).

23. Kværner, A. S. *et al.* Associations of the 2018 World Cancer Research Fund/American Institute of Cancer Research (WCRF/AICR) cancer prevention recommendations with stages of colorectal carcinogenesis. *Cancer Med.* **12**, 14806–14819 (2023).

24. Kieser, S., Brown, J., Zdobnov, E. M., Trajkovski, M. & McCue, L. A. ATLAS: a Snakemake workflow for assembly, annotation, and genomic binning of metagenome sequence data. *BMC Bioinformatics* **21**, 257 (2020).

541 25. Bushnell, B. BBMap: BBMap short read aligner, and other bioinformatic tools. *SourceForge*
542 https://sourceforge.net/projects/bbmap/ (2022).

543 26. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile
544 metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).

545 27. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing
546 single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).

547 28. Wu, Y.-W., Simmons, B. A. & Singer, S. W. MaxBin 2.0: an automated binning algorithm to recover
548 genomes from multiple metagenomic datasets. *Bioinformatics* **32**, 605–607 (2016).

549 29. Sieber, C. M. K. *et al.* Recovery of genomes from metagenomes via a dereplication, aggregation
550 and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).

551 30. Guo, J. *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA
552 viruses. *Microbiome* **9**, 37 (2021).

553 31. Camargo, A. P. *et al.* Identification of mobile genetic elements with geNomad. *Nat. Biotechnol.*
554 **42**, 1303–1312 (2024).

555 32. Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral
556 genomes. *Nat. Biotechnol.* **39**, 578–585 (2021).

557 33. Aroney, S. T. N., Camargo, A. P., Tyson, G. W. & Woodcroft, B. J. Galah: More scalable dereplication
558 for metagenome assembled genomes. https://doi.org/10.5281/zenodo.10526086 (2024).

559 34. Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by
560 gene-sharing networks. *Nat. Biotechnol.* **37**, 632–639 (2019).

561 35. Cook, R. *et al.* INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases
562 in the Current Collection of Cultured Phage Genomes. *PHAGE New Rochelle N* **2**, 214–223 (2021).

563 36. Birkeland, E., Istvan, P., Avershina, E., Rognes, T. & Rounge, T. B. VirMake: a flexible and user-
564 friendly pipeline for viral taxonomic and functional characterisation from shotgun metagenomic
565 sequencing data. 2025.02.07.637044 Preprint at https://doi.org/10.1101/2025.02.07.637044 (2025).

566 37. Pellow, D. *et al.* SCAPP: an algorithm for improved plasmid assembly in metagenomes.
567 *Microbiome* **9**, 144 (2021).

568 38. bio.tools · Bioinformatics Tools and Services Discovery Portal. https://bio.tools/bbtools.

569 39. Galata, V., Fehlmann, T., Backes, C. & Keller, A. PLSDB: a resource of complete bacterial plasmids.
570 *Nucleic Acids Res.* **47**, D195–D202 (2019).

571 40. Camargo, A. P. *et al.* IMG/PR: a database of plasmids from genomes and metagenomes with rich
572 annotations and metadata. *Nucleic Acids Res.* **52**, D164–D173 (2024).

573 41. Robertson, J., Bessonov, K., Schonfeld, J. & Nash, J. H. E. Universal whole-sequence-based plasmid
574 typing and its utility to prediction of host range and epidemiological surveillance. *Microb. Genomics* **6**,
575 mgen000435 (2020).

576 42. Alcock, B. P. *et al.* CARD 2023: expanded curation, support for machine learning, and resistome
577 prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.* **51**, D690–D699
578 (2023).

579 43. Mölder, F. *et al.* Sustainable data analysis with Snakemake. Preprint at
580 https://doi.org/10.12688/f1000research.29032.3 (2025).

581 44. Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S. A. & Sørensen, S. J. CRISPRCasTyper:
582 Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. *CRISPR J.* **3**, 462–469
583 (2020).

584 45. Pourcel, C. *et al.* CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes
585 from complete genome sequences, and tools to download and query lists of repeats and spacers.
586 *Nucleic Acids Res.* **48**, D535–D544 (2020).

587 46. Camacho, C. & Madden, T. BLAST+ Release Notes. in *BLAST® Help [Internet]* (National Center for
588 Biotechnology Information (US), 2025).

589 47. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular
590 interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

591 48. Al-Shayeb, B. *et al.* Diverse virus-encoded CRISPR-Cas systems include streamlined genome
592 editors. *Cell* **185**, 4574-4586.e16 (2022).

593 49. Pinilla-Redondo, R. *et al.* CRISPR-Cas systems are widespread accessory elements across bacterial
594 and archaeal plasmids. *Nucleic Acids Res.* **50**, 4315–4328 (2022).

595 50. Ono, R., Konno, N., Nishimura, Y. & Furusawa, C. Host range and antibiotic resistance
596 dissemination are shaped by distinct survival strategies of conjugative plasmids. *Nucleic Acids Res.* **54**,
597 gkaf1479 (2026).

598 51. Beamud, B. *et al.* Genetic determinants of host tropism in Klebsiella phages. *Cell Rep.* **42**, 112048
599 (2023).

600 52. Bignaud, A. *et al.* Phages with a broad host range are common across ecosystems. *Nat. Microbiol.*
601 **10**, 2537–2549 (2025).

602 53. Andreu-Sánchez, S. *et al.* Global genetic diversity of human gut microbiome species is related to
603 geographic location and host health. *Cell* **188**, 3942-3959.e9 (2025).

604 54. Camargo, A. P. *et al.* IMG/VR v4: an expanded database of uncultivated virus genomes within a
605 framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Res.* **51**, D733–
606 D743 (2023).

607 55. Kim, S. *et al.* Selective loading and processing of prespacers for precise CRISPR adaptation. *Nature*
608 **579**, 141–145 (2020).

609 56. Lee, H. & Sashital, D. G. Creating memories: molecular mechanisms of CRISPR adaptation. *Trends*
610 *Biochem. Sci.* **47**, 464–476 (2022).

611 57. McKenzie, R. E. *et al.* Single cell variability of CRISPR-Cas interference and adaptation. *Mol. Syst.*
612 *Biol.* **18**, MSB202110680 (2022).

613 58. Zhao, S. *et al.* Adaptive Evolution within Gut Microbiomes of Healthy People. *Cell Host Microbe*
614 **25**, 656-667.e8 (2019).

615 59. Dommels, Y. E. M. *et al.* Survival of Lactobacillus reuteri DSM 17938 and Lactobacillus rhamnosus
616 GG in the human gastrointestinal tract with daily consumption of a low-fat probiotic spread. *Appl.*
617 *Environ. Microbiol.* **75**, 6198–6204 (2009).

618 60. Poutsiaka, D. D. *et al.* 1681Randomized Clinical Trial of the Recovery of Probiotics Lactobacillus
619 rhamnosus GG (LGG) and Bifidobacterium animalis subspecies lactis (BB-12) from the Gastrointestinal
620 (GI) Tract of Healthy Volunteers. *Open Forum Infect. Dis.* **1**, S449 (2014).

621 61. Matchado, M. S. *et al.* Network analysis methods for studying microbial communities: A mini
622 review. *Comput. Struct. Biotechnol. J.* **19**, 2687–2698 (2021).

623 62. Bruce, J. B. *et al.* Temperate phage evolve to integrate host stress and quorum signals in lysis–
624 lysogeny decisions. *PLOS Biol.* **24**, e3003567 (2026).

625 63. Lopez, J. G., Donia, M. S. & Wingreen, N. S. Modeling the ecology of parasitic plasmids. *ISME J.* **15**,
626 2843–2852 (2021).

627 64. Keith, N. C., Snyder, R. A., Euler, C. W. & Modell, J. W. Bacteria exploit viral dormancy to establish
628 CRISPR-Cas immunity. *Cell Host Microbe* **33**, 330-340.e6 (2025).

629 65. Rounge, T. B. *et al.* Evaluating gut microbiota profiles from archived fecal samples. *BMC*
630 *Gastroenterol.* **18**, 171 (2018).

631 66. Jin, H. *et al.* Hybrid, ultra-deep metagenomic sequencing enables genomic and functional
632 characterization of low-abundance species in the human gut microbiome. *Gut Microbes* **14**, 2021790.

633