

Planetary-scale metagenomic analysis reveals CRISPR targeting dynamics across ecosystems

In the format provided by the
authors and unedited

5

10

Supplementary material for:

Planetary-scale metagenomic search reveals new patterns of CRISPR targeting

15 Simon Roux^{1*}, Uri Neri¹, Brian Bushnell¹, Brayon Fremin¹, Nikhil A. George², Uri Gophna³, Laura A. Hug⁴,
Antonio Pedro Camargo¹, Dongying Wu¹, Natalia Ivanova¹, Nikos Kyrpides¹, Emiley Eloë-Fadrosh^{1*}

¹DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

20 ²Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

³The Shmunis School of Biomedicine and Cancer Research, Tel Aviv University, Tel Aviv, Israel

⁴Department of Biology, University of Waterloo, Waterloo, Ontario, Canada

Correspondence to: sroux@lbl.gov, eaeloefadrosh@lbl.gov

25

Table of Contents

Supplementary Text.....	2
1. Overview of the predicted CRISPR repeat and spacer database content.....	2
2. Comparison of approaches for the identification of CRISPR spacers from metagenomic reads.....	3
3. Estimation of population-wide CRISPR locus diversity from metagenomic reads.....	3
4. Ecological and taxonomic distribution of spacer sets with unusually large/limited spacer diversity.....	4
5. Spacer hits to atypical phages and archaeal viruses.....	4
6. Spacer hits to phages and archaeal viruses with known hosts.....	5
7. Targeting patterns for UViGs connected to repeats assigned to multiple phyla.....	5
Supplementary Tables (legends).....	8
Source Data Files (list).....	9
Supplementary Data Files (list and links).....	10
Supplementary Figures.....	11
References.....	24

Supplementary Text

30 1. Overview of the predicted CRISPR repeat and spacer database content

The entire repeat database was built from ~3.8M predicted CRISPR repeats, which dereplicated into 581,889 non-redundant repeat sequences. Among these, 43,483 were detected in at least one bacteria or archaea genome. Thanks to the genomic context available for these repeats, a majority can be assigned to a given taxon (99.1%) and associated with a known CRISPR type (73.7%). Repeats found in a genome but not clearly associated with a known CRISPR type include the ones from “orphan” arrays with no co-localized Cas genes. Overall, these “genome” repeats reflect the known diversity of CRISPR-Cas systems, and include for instance the canonical sequences used for establishing the different CRISPR types and subtypes as well as several CRISPR-Cas systems studied *in vitro* for e.g. PAM identification (Supplementary Table S1). CRISPR types could be predicted for most of these “genome” repeats, either based on co-occurring Cas genes or based on the repeat sequence itself, and showed a predominance of CRISPR-Cas systems of types I (I-C, I-E, and I-B primarily), and types III (III-A, III-D, and III-B primarily). This distribution is consistent with known distribution of CRISPR types in bacteria¹, which represent the main source of repeats in our database. In contrast, the main types of repeats assigned to the domain Archaea in our data were I-B, I-A, I-D, and III-A, again consistent with previous survey of these genomes¹. Furthermore, among repeats that could be classified taxonomically, most (61.3%) were assigned to a single family or genus (Fig. S1A). This pattern was even stronger for spacers, which were typically (99.0%) associated with a single repeat sequence, or associated with a single genus when linked to multiple repeat sequences (78.5%). This indicates that a majority of CRISPR repeats and spacers are specific to a relatively narrow clade or taxon, and our sequence clustering approach can distinguish these cases from CRISPR repeats and spacers shared across a broader taxonomic range, which is critical for robust interpretation of CRISPR spacer matching for host prediction. Finally, a majority (n=5.3M, 67.1%) of spacers were derived from these “genome” repeats, even if these represent only a small fraction (7.47%) of the overall repeat sequence diversity. Taken together, we interpret the “genome” repeats and their associated spacers as representing the most common CRISPR systems across microbiomes, which for the most part correspond to “known” and previously described CRISPR-Cas systems, or minor variants thereof.

In contrast, a large fraction of dereplicated repeats were only detected from metagenomes, in particular short contigs (n=442,389 contigs) often containing only the predicted CRISPR array without any surrounding gene. As previously noted², the short length of these contigs likely reflects the challenge of assembling CRISPR array loci from metagenomes: because these loci are often variable between members of a same population, assemblers can not reconstruct a clear consensus sequence from the assembly graph, leading to CRISPR arrays being found on disconnected contigs. These repeat sequences were in majority (68.75%) singletons, i.e. detected in a single contig in a single metagenome, and associated with ~ 37x less spacers per repeat compared to “genome” repeats (see above). Overall, we interpret these “metagenome-only” predicted repeats as reflecting CRISPR-Cas systems encoded on genomes from relatively rare organisms and/or strains. This would explain both the fragmented nature of the assemblies that prevents taxonomic assignment and CRISPR type prediction, and their absence from genome databases. Considering that new repeat sequences detected in fully assembled genomes can often be associated with a known CRISPR type, we expect this to be the case for these “metagenome-only” repeats as well, and interpret the absence of CRISPR type prediction as “unknown” CRISPR type (i.e. with more data, these repeats could likely be connected to a known CRISPR type), rather than “novel” CRISPR type (i.e. these repeats are likely associated with CRISPR-Cas systems outside of the current classification). Finally, some of these “metagenome-only” repeats may also originate from repeat loci not associated with a CRISPR-Cas system, especially repeats for which no spacer set (i.e. one sample with at least 5 distinct spacers) could be identified (n=258,857). While these would represent erroneous predictions, since no spacer was collected for these repeats, they did not have any impact on downstream analysis of CRISPR array diversity and CRISPR targeting.

Instead, these downstream analyses rely mostly on “genome” repeats, as these yielded 67.1% of all spacers (Fig. 1).

2. Comparison of approaches for the identification of CRISPR spacers from metagenomic reads

Because of the repetitive nature of CRISPR array loci, including the presence of near-identical spacers in the same array and the presence of distinct arrays with near-identical repeats on the same genome or in the same microbiome, robustly identifying the set of spacers associated with a given repeat is challenging. Both assembly-based and read-based methods can typically only recover a portion of all spacers encoded, and read-based methods in particular can yield artefactual spacers due to sequencing errors. To illustrate the efficiency of these methods across different levels of sequencing error and genome coverage, we compared our SpacerExtractor approach to the state of the art programs in the field, CRASS² and MetaCRAS³, using simulated reads from isolate genomes from which CRISPR arrays were previously predicted using CRISPR-Cas Typer⁴, as done in ref. ⁵.

Overall, this benchmarking recapitulated expected trends and highlighted some limitations of CRISPR spacer recovery from metagenomic reads (Fig. S2). Across all tools, both recall (percentage of expected spacers recovered) and false-discovery rate (FDR, percentage of unexpected spacers recovered among all predicted spacers) increased with coverage, i.e., as more reads are available for a given genome, all tools recover more of the original spacers, along with additional erroneous spacers (Fig. S2). The increase in FDR was more important at 100x compared to 10x, and when simulating reads using higher sequencing error rates (HiSeq 2500, vs ~ half of the error of HiSeq 2500 as expected for more recent Illumina sequencing technologies such as NovaSeq). When considering both recall and FDR (F1-measure), MetaCRAS³ and SpacerExtractor showed similar performances for both error profiles at 10x and 100x, while SpacerExtractor performed better for low-coverage genomes (0.8x) with both a higher recall and lower FDR (Fig. S2). The FDR remained relatively high for genomes with high coverage (100x) and high sequencing error rate despite the denoising procedure included in SpacerExtractor. Ignoring singleton spacers substantially decreased this FDR at high coverage, but at the same time led to a strong reduction in recall for low-coverage genomes (Fig. S2).

This benchmarking thus suggests that (i) all tools only recover a fraction of the CRISPR spacers predicted from the original genome assembly, including at high coverage, and (ii) in some cases (high genome coverage, high sequencing error rate), a substantial number of erroneous spacers can also be predicted, often as singleton spacers. Given that most genomes will be present at relatively low coverage in a metagenome, we opted in this study for using SpacerExtractor with standard denoising and including singleton spacers, however studies focusing specifically on genomes with deep sequencing may want to use a more aggressive denoising procedure and/or remove singleton spacers from their analysis. Importantly, these benchmarks only focus on a set of known repeats, i.e. repeats previously predicted by CRISPR-Cas Typer, and do not take into account the potential of some of these tools to discover entirely new CRISPR arrays (i.e. new repeats and spacers).

3. Estimation of population-wide CRISPR locus diversity from metagenomic reads

Most published work on CRISPR-Cas systems relies on assembled genomes. This includes the analysis of assembled CRISPR arrays, which already revealed some of the key mechanisms governing spacer acquisition and CRISPR arrays evolution processes⁶⁻⁸. Because our analysis is based on short-read metagenomes, and CRISPR arrays are expected to be variable between members of a same population⁹, we did not assemble individual CRISPR array loci from our data. Instead, we opted to analyze spacer diversity as part of “spacer sets”, defined here as the non-redundant collection of spacers associated with a given repeat in a given sample. While these spacer sets do not allow for an analysis of spacer ordering within arrays, and are not directly comparable to previous observations of number and diversity of spacers identified in assembled arrays, we interpret these sets as reflecting the diversity of spacers encoded by members of a given population in a sample. As such, these

120 spacer sets uniquely provide an estimation of the population-level variability of spacer array content in natural microbiomes.

Most spacers sets seemed to be composed of a few high-coverage spacers followed by a long tail of rare spacers (Fig. 2B). This pattern may, however, be influenced by singleton spacers, i.e. spacers detected on a single read. These singleton spacers likely include a combination of spacers only acquired and/or encoded by rare
125 population members, spacers having undergone mutation event(s) after acquisition as part of the host genome replication process, and artefactual spacers deriving from errors introduced during the sequencing process. If most of the singleton spacers derive from sequencing errors, the spacer diversity observed in spacer sets may be artificially inflated.

To evaluate the likelihood of singleton spacers deriving mostly from a methodological artifact, we clustered
130 spacers within each spacer set at 95% and 80% identity. We reasoned that “real” rare spacers deriving from independent spacer acquisition events would most likely not be similar to another spacer in the set. Meanwhile, spacers having undergone post-acquisition mutation or deriving from sequencing errors would likely be similar to at least one other spacer in the set, and thus should not be singletons anymore when clustering at lower identity percentage. Across all spacer sets, when clustering spacers at 80% identity within a set, 37.1% of
135 singletons were clustered with another spacer, and may possibly originate from one of the latter two scenarios (spacer mutation after acquisition, sequencing error). The remaining 62.9% likely represent genuine rare spacers (Fig. 2C). This suggests that, overall, CRISPR spacer arrays are most often highly diverse and variable between members of a given microbial population, and that the observed diversity is not primarily explained by artefactual spacers, which is also consistent with previous PCR-based targeted exploration of CRISPR spacers¹⁰.

140

4. Ecological and taxonomic distribution of spacer sets with unusually large/limited spacer diversity

After normalizing for coverage depth variations, unusually large spacer sets were most often observed in aquatic biomes, especially thermal springs, and engineered systems. Unusually large spacer sets were also more frequently observed in uncultivated taxa known or predicted to engage in microbe-microbe interactions
145 such as *Candidatus* Kryptonium, *Candidatus* Thiodubiliella, or *Candidatus* Magnetomorum^{11–13}(Fig. S8). This suggests that specific taxa-environment combinations may select for large and diverse spacer sets, i.e. some populations may be uniquely able to accommodate, and likely benefit from, expanded CRISPR spacer repertoires.

On the other end, a small minority (4.9%) of sets showed a much lower intra-population variation and were
150 instead composed of $\geq 50\%$ of “common” spacers (Fig. S9). These sets were primarily observed for repeats of type V CRISPR arrays encoded by *Clostridia*, *Bacteroidia*, and *Spirochaetia*, which were frequently associated with a relatively low spacer diversity overall (Fig. S9). These repeats may thus represent arrays with limited rates of spacer acquisition, and for which most spacers may still be retained as “memory” spacers¹⁴. Further studies of individual CRISPR arrays will however be required to characterize in more detail the eco-evolutionary drivers
155 of spacer acquisition and retention in natural populations^{5,15}.

5. Spacer hits to atypical phages and archaeal viruses

Caudoviruses (tailed phages) are the most common phages and archeoviruses in IMG/VR and represent the majority of viruses with spacer hits. Beyond caudoviruses however, several other virus taxa also showed sign of
160 CRISPR spacer targeting (Extended Data Fig. 6). In particular, spacer hits were observed to other dsDNA and ssDNA phages including microviruses (*Malgrandviricetes*) and inoviruses (*Faserviricetes*); atypical archaeal viruses such as bicaudaviruses, fuselloviruses, and globuloviruses; and RNA phages in the levi, cysto, and picobirnavirus groups. Across all these taxa, the overall patterns of targeting in terms of spacer taxonomy and ecosystem were consistent with expectations. Specifically, dsDNA and ssDNA phages were targeted by a
165 relatively broad diversity of CRISPR spacers found across all types of ecosystems. Conversely, taxa known to infect extremophile archaea were almost exclusively targeted by spacers from archaea-assigned repeats

obtained from extreme aquatic environments (Extended Data Fig. 6). One possible exception was *Aeropyrum* coil-shaped virus (family *Spiraviridae*, IMGVR_UViG_2974665388_000001) for which spacer hits from “marine” samples were obtained (BioSample SAMN06761480), however further examination of the sample metadata revealed it was taken near Vulcano islands, known to harbor extreme archaea including e.g. *Pyrococcus*. Taken together, these results confirm that spacer hits detected through our global database matching approach seem to represent genuine CRISPR spacer targeting of atypical viruses.

Beyond expected taxa and ecosystems, these spacer hits also revealed potential new hosts for some of these atypical phage and archeovirus taxa, for which relatively few isolates are available. First, these data confirmed that both microviruses and inoviruses seem to interact with a broad range of microbial taxa, with 193 and 197 distinct families with ≥ 2 spacer targeting at least one of these ssDNA phage sequences (0 or 1 mismatch), respectively^{16,17}. Meanwhile, relatively few spacer hits were identified for known and/or predicted clades of RNA phages, namely leviviruses, cystoviruses, and picobirnaviruses. Specifically, even with the global spacer database, no cystovirus or picobirnavirus could be connected to a CRISPR repeat by more than 1 spacer, and only 48 leviviruses could, out of 82,176 total levivirus sequences in IMG/VR. This low level of CRISPR targeting contrasts with a recent description of a putative new clade of RNA phages, related to partitiviruses, that are highly targeted by a *Roseiflexus* CRISPR array¹⁸. This may be due to a combination of (i) CRISPR-Cas systems only rarely being used to defend against RNA phages, and/or (ii) the fact that the partiti-like virus targeting was detected based on paired metagenomes and metatranscriptomes sampled from the same site, indicating that RNA phages and CRISPR spacers may diverge too quickly to lead to numerous near-exact matches in a global database search, and targeted sampling may be required to identify these interactions.

6. Spacer hits to phages and archaeal viruses with known hosts

To explore the relationship between CRISPR targeting and virus-host interaction, we focused on bacteriophages and archaeal viruses in IMG/VR for which a “known” host is available, either because the virus was isolated and grown on a host culture (“isolate”, $n=3,438$), or because it was detected as part of the whole genome shotgun sequencing of a single bacteria or archaea (summarized as “prophage”, $n=230,544$). The taxonomy (GTDB r214) of these “known hosts” was then compared to the taxonomy assigned to CRISPR repeats (using the same GTDB r214) for which at least one spacer hit was identified to this virus, and 3 cases were identified: (i) both taxonomies were identical from domain to genus rank (ii) taxonomies were inconsistent, e.g. pointing to a different genus, family, or order, and (iii) taxonomies were consistent but one of the two assignments (the “known host” or the CRISPR repeat) was unclassified at the genus rank (and sometimes at higher ranks as well). Cases in this latter category (one of the two taxonomies unclassified at genus rank or higher) were ignored as taxonomy consistency could not be validated in these cases.

For virus-taxon connected by less than 10 spacer hits, 45.9% of cases showed consistent taxonomies. Meanwhile, for virus-taxon connected by 10 spacer hits or more, 76.8% showed consistent taxonomies. These percentages were consistent between “isolate” references (22.5% consistency for <10 spacer hits, and 77.5% consistency for ≥ 10 spacer hits) and “prophage” references (46.2% consistency for <10 spacer hits, and 76.8% consistency for ≥ 10 spacer hits). This high number of spacer hits (≥ 10 hits) was not specific to isolate or provirus genomes: across all HQ viruses, nearly half (48.1%) showed at least one repeat with ≥ 10 near-exact spacer hits, and may thus be connected to a putative host even when using this relatively conservative cutoff.

7. Targeting patterns for UViGs connected to repeats assigned to multiple phyla

In previous analyses, several UViGs were highlighted as potentially targeted by repeats assigned to different phyla. Some of these UViGs were also targeted by multiple repeats assigned to different phyla in our analysis, and are displayed in panels A and B of Figure 5. Below is a more detailed description of the targeting patterns observed for these two figure panels.

215 I24, I20: Spacer hits were detected for multiple repeats assigned to Patescibacteria and *Actinomycetota*, including repeats with high confidence (see Methods) for each phylum. In both cases, the number of spacer hits was low for *Actinomycetota* (≤ 5), and most hits displayed 2 or 3 mismatches between spacer and (predicted) protospacer.

220 I33: Spacer hits were detected for multiple repeats assigned to Patescibacteria with high confidence, but only one repeat assigned to *Actinomycetota* with high-confidence. No spacer hits were detected from other repeats assigned to *Actinomycetota* with medium or low confidence. A total of 17 hits were detected between the *Actinomycetota* spacers and this UViG, but only a single hit was detected with 1 mismatch or less.

225 U105: Spacer hits were detected for multiple repeats assigned to Patescibacteria and Bacillota_A, including multiple repeats with high confidence for each phylum. The total number of spacer hits was low for Bacillota_A (≤ 2 per taxon), and higher for Patescibacteria (up to $\sim 3,000$ for a single repeat).

230 U060: Spacer hits were detected for multiple repeats assigned to Patescibacteria, Bacillota_A, and *Bacillota*, including multiple repeats with high confidence for each phylum. The total number of spacer hits with 0 or 1 mismatch was low for *Bacillota* and Bacillota_A (≤ 2 per taxon), and higher for Patescibacteria (up to 363 for a single repeat).

235 I14: Spacer hits were detected to multiple repeats assigned to Bacillota_A, *Bacillota*, and *Pseudomonadota* with high confidence. Spacer hits were also detected to two repeats assigned to Patescibacteria, but this assignment was made with low confidence (repeats assigned based on detection in a single MAG). A relatively low number of hits were obtained for repeats assigned to *Pseudomonadota* or *Bacillota* (≤ 9 , and ≤ 5 with 0 or 1 mismatches), while a larger number of hits were obtained for repeats assigned to Bacillota_A (up to 47, up to 14 hits with 0 or 1 mismatches).

240 I49: Spacer hits were obtained from 3 repeats, all with a low-confidence taxonomic assignment (taxonomic assignment based on detection of the repeat in a single MAG).

245 I60: Spacer hits were detected from multiple repeats assigned to Desulfobacterota, including one with high confidence. The number of hits for these repeats was also relatively high (multiple repeats with ≥ 80 hits with 0 or 1 mismatch). Spacer hits were also detected for a single repeat assigned to *Streptococcus* with medium confidence (repeat detected in a single isolate genome), and one assigned to *Elusimicrobiota* with low confidence. The number of hits associated with these latter two repeats was lower (< 25).

250 I15: Spacer hits were detected from repeats assigned to *Bacillota* and Bacillota_C, one repeat with high-confidence assignment for each phylum (*Bacilli* ; *Streptococcus* and *Negativicutes* ; *Megasphaera*), as well as one repeat assigned with medium confidence to *Halobacteriota* (*Methanomicrobia* ; *Methanofollis*). Most individual repeats showed a relatively low number of spacer hits, with ≤ 2 hits with 0 or 1 mismatch for the two repeats with high-confidence assignment.

255 I86: Spacer hits were detected to one repeat assigned to *Halobacteriota* with medium confidence, as well as single repeats assigned to *Pseudomonadota* or *Verrucomicrobiota* with low confidence. All repeats showed a relatively high number of hits (> 50 hits), the vast majority without any mismatch between spacer and target.

260 I87: Spacer hits were detected from multiple repeats assigned to *Methanoculleus* with medium confidence. The number of this for these repeats was relatively high (multiple repeats with ≥ 50 hits with 0 or 1 mismatch).

Spacer hits were also detected for a single repeat assigned to *Thermoplasmatota* and a single repeat assigned to *Thermotogota*, both with low confidence. The number of hits associated with the *Thermotogota* repeat was also high (>100 hits) while this number was lower for the *Thermoplasmatota* repeat (8 hits).

265 I10: Spacer hits were detected from multiple repeats assigned to *Halobacteriota* with medium confidence, and for one repeat assigned to *Bacillota_A* with low confidence (repeat detected in a single MAG). A large (≥ 500) number of hits with 0 or 1 mismatch were detected for repeats assigned to *Halobacteriota*, and a lower but still substantial number (45) were detected for the repeat assigned to *Bacillota_A*.

270 I48: All spacer hits were detected for repeats with low-confidence taxonomic assignment (repeats detected in a single MAG). A large (≥ 30) number of hits with 0 or 1 mismatches were detected for several repeats.

I53: Spacer hits were detected to multiple repeats assigned to *Bacteroidota* with medium confidence, as well as a single repeat assigned to *Armatimonadota* with low confidence. Both repeats assigned to *Bacteroidota* and
275 *Armatimonadota* showed a relatively high number of hits (≥ 30).

Supplementary Tables (legends)

Table S1. **Characteristic of known CRISPR repeats detected in the SpacerExtractor database, as well as the 50 most common repeats (i.e. repeats with the most spacers).** The first part of the table lists known references CRISPR-Cas systems (described in refs.^{1,19-26}), and the corresponding repeat cluster in the SpacerExtractor ("SE") database. For each repeat, the information inferred from the repeats and spacer hits in our database is indicated in columns prefixed with "SE", while information from other columns are derived from the published literature. The bottom table presents the same information for the 50 repeat clusters with the most associated spacers.

Table S2. **PAM motifs predictions for individual repeats based on conserved neighborhood of hits to IMG/VR and IMG/PR sequences.** Only hits with 0 or 1 mismatch between spacer and predicted protospacer were considered. Only repeats with ≥ 50 distinct neighborhoods (10 bases upstream and downstream of the hit) and repeats with < 50 distinct neighborhoods but a predicted PAM corresponding to a known motif for their predicted type are included in this supplementary table. All motifs are shown as if upstream from the protospacer, as the spacer orientation can be ambiguous in our dataset. This means that some of these PAMs are not the exact sequence shown here, but instead the reverse-complement downstream from the spacer.

Table S3. **Functional annotation of spacer hits for different types of spacers (single vOTU vs multi-vOTU).** For each type of spacer (either spacers hitting a single vOTU, or spacers hitting multiple vOTUs), 8 random subsamples of 100,000 spacers were collected and functionally annotated when overlapping a CDS. The table shows the average and individual percentage of each functional category and match type (intergenic, intragenic, or overlapping gene edge) for each subsample for single vOTU spacers first, then for multi-vOTU spacers.

Table S4. **Targeting patterns for selected UViGs connected to different phyla.** The connection between UViGs and arrays depicted as virus-taxon networks in Fig. 6 c and d are detailed here for each individual repeat. The confidence of the taxonomic assignment is indicated along with the percentage of target covered by spacers, the number of spacer hits at different mismatch levels (from 0 to 3) and the corresponding correlation level between spacer hit number and mismatch number.

Source Data Files (list)

source_data_fig1_array-eco.tsv: Source data for Fig. 1, panel D
310 source_data_fig2b_10k_slice.tsv: Source data for Fig. 2, panel B
source_data_fig2d_spacer_dist.tsv: Source data for Fig. 2, panel D
source_data_fig3_multitargeting-100k-slice.tsv: Source data for Fig. 3, panel E (dot plot)
source_data_fig4_known_hosts_hits.tsv: Source data for Fig. 4, panel B
source_data_fig5_edges.tsv: Source data for Fig. 5, panels A & B (network's edges)
315 source_data_fig5_nodes.tsv: Source data for Fig. 5, panels A & B (network's nodes)
source_data_ef2_array-eco.tsv: Source data for Extended Data Fig. 2
source_data_ef3_genus_eco.tsv: Source data for Extended Data Fig. 3
source_data_ef4_common_repeats.tsv: Source data for Extended Data Fig. 4
source_data_ef5_lowdiv.tsv: Source data for Extended Data Fig. 5 panel A
320 source_data_ef5_lowdiv_nosingleton.tsv: Source data for Extended Data Fig. 5 panel B
source_data_ef5_frozen.tsv: Source data for Extended Data Fig. 5 panel C
source_data_ef6_atypical_targets.tsv: Source data for Extended Data Fig. 6 (bottom panels)
source_data_ef7_pam.tsv: Source data for Extended Data Fig. 7
source_data_ef8_IMGVR_UViG_638275558_000001-RM378-spacer_hits.tsv: Source data for Extended Data Fig. 8
325 (top panel)
source_data_ef8_IMGVR_UViG_2911423748_000003-Agathobacter-spacer_hits.tsv: Source data for Extended Data
Fig. 8 (bottom panel)

Supplementary Data Files (list and links)

- 330 Data file S1. **List of repeat sequences used to mine CRISPR spacers in metagenomes.**
https://portal.nersc.gov/dna/microbial/prokpubs/spacer_database_resources/Supplementary_Data_File_1_repeats.tsv
<https://figshare.com/ndownloader/files/66995219>
- 335 Data file S2. **List of metagenomes mined for CRISPR spacers.**
https://portal.nersc.gov/dna/microbial/prokpubs/spacer_database_resources/Supplementary_Data_File_2_samples.tsv
<https://figshare.com/ndownloader/files/66995225>
- 340 Data file S3. **List of predicted Cas operons for CRISPR repeats in SpacerExtractor database.**
https://portal.nersc.gov/dna/microbial/prokpubs/spacer_database_resources/Supplementary_Data_File_3_cas.tsv
<https://figshare.com/ndownloader/files/66995222>

Supplementary Figures

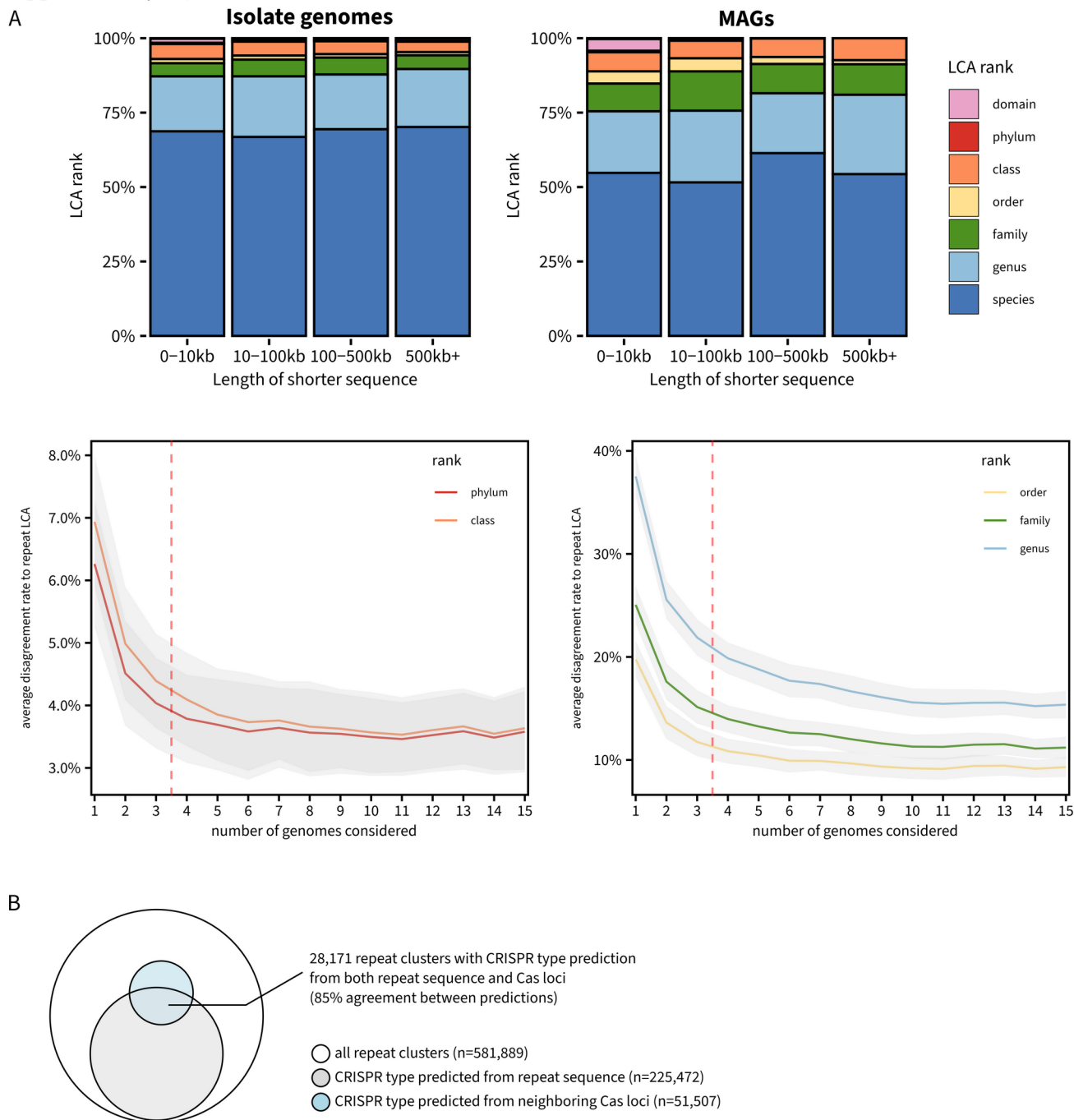


Figure S1. Taxonomic and type distribution of predicted CRISPR repeats. A. The top panel shows the lowest common ancestor (LCA) rank for individual CRISPR repeats detected in ≥ 2 genomes for isolate genomes (left) and metagenome-assembled genomes (MAGs, right). The bottom panels show the consistency in predicted LCA when considering only a subset of the genomes (x-axis) at different taxonomic ranks (colors). The orange dashed line indicates the cutoff for minimum number of genomes (≥ 4) required to consider a taxonomic assignment as “high-confidence” (if isolates) or “medium-confidence” (if MAGs, see Methods). B. Comparison of

the number of CRISPR type predictions obtained from the repeat sequence (gray) or neighboring Cas loci (blue).

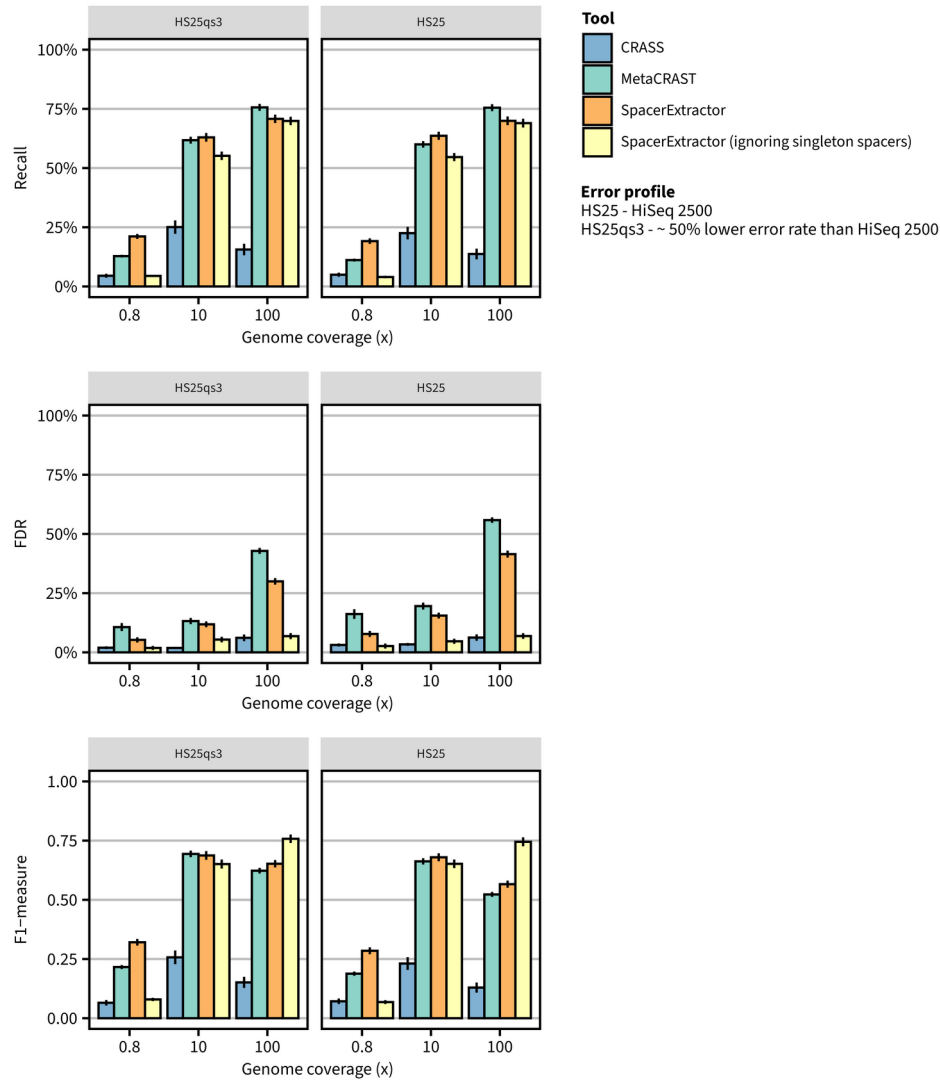


Figure S2. **Evaluation of spacer recovery from short reads across coverages, error rates, and tools.** Three metrics were computed based on the prediction of spacers by CRASS, MetaCRAST, and SpacerExtractor: recall (top row), false-discovery rate (middle row), and F1 score (bottom row). For SpacerExtractor, these metrics were calculated both with all high-quality spacers, or with all high-quality spacers excluding singletons, i.e. requiring spacer coverage >1. Two error profiles were used when simulating reads, corresponding to ~50% of the error rate of the Illumina HiSeq 2500 sequencer (left column) or ~100% of the error rate of the Illumina HiSeq 2500 sequencer (right column), and three levels of genome coverage were used (0.8x, 10x, 100x, x-axis). For each metric, the average and standard error across 200 genomes is indicated.

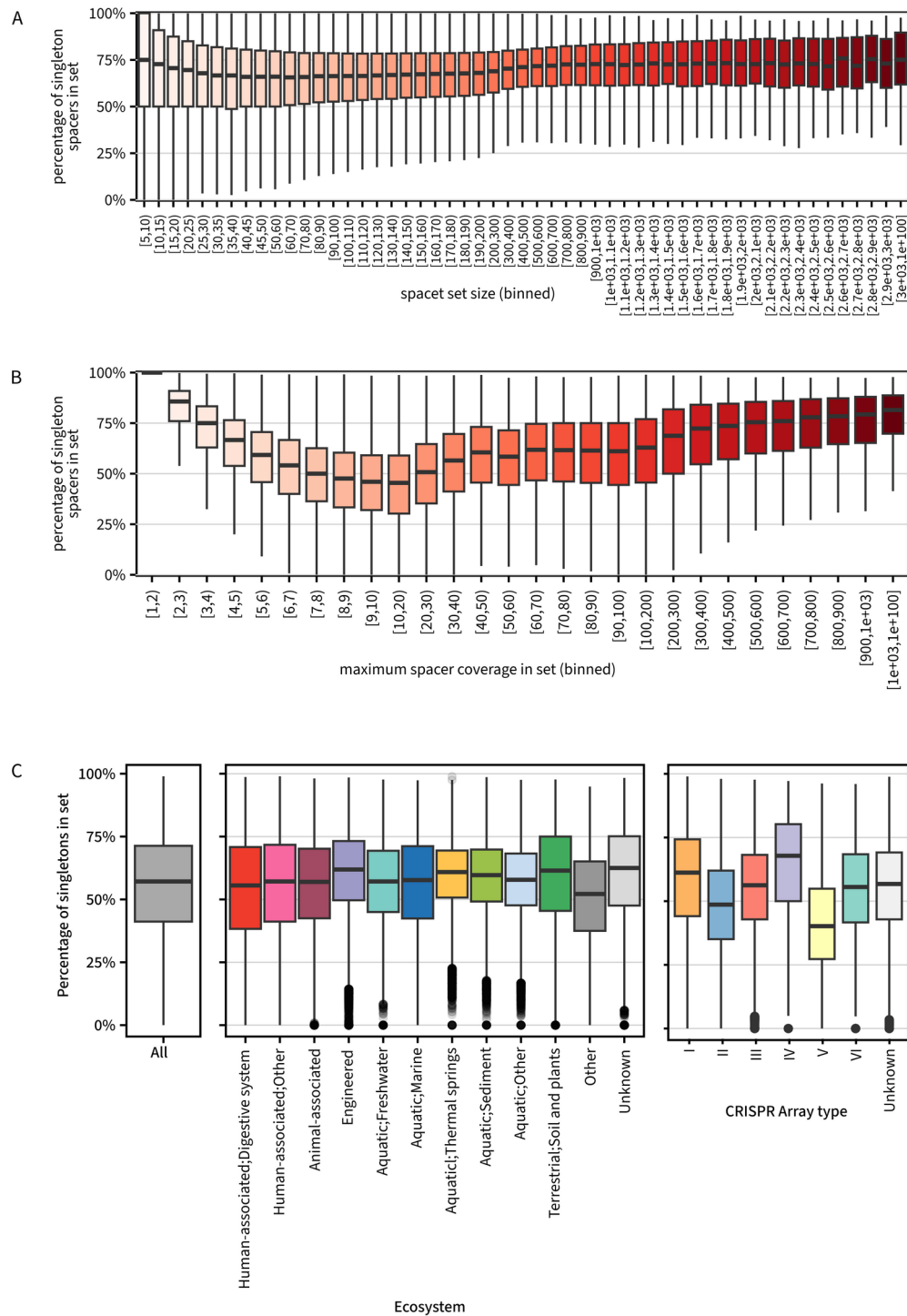


Figure S3. Percentage of singletons across spacer set size, maximum spacer coverage in set, ecosystem, and CRISPR array type. A. Relationship between the percentage of singletons in a spacer set (y-axis) and the size of spacer sets (x-axis). B. Relationship between the percentage of singletons in a spacer set (y-axis) and the maximum spacer coverage in the spacer set (x-axis). C. Distribution of the percentage of singletons in a spacer set (y-axis) for all sets (left), by ecosystem (middle), and CRISPR array type (right). Only spacer sets for which the maximum spacer coverage was 20x or higher were included in plots for the bottom row.

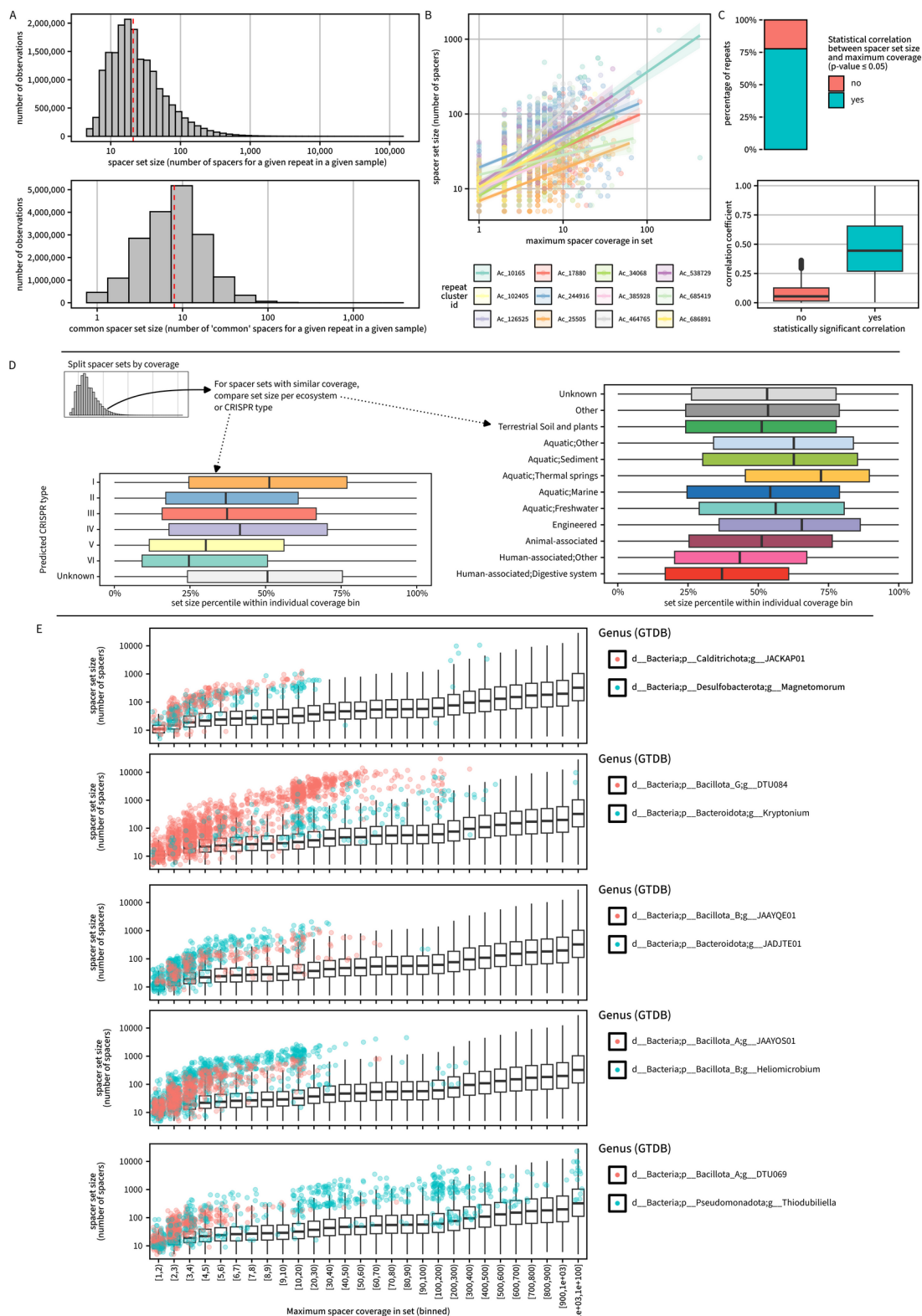


Figure S4. **Spacer set diversity as a function of spacer set size, ecosystem, and CRISPR array type.** **A.** Distribution of the number of distinct spacers across spacer sets, considering all spacers (top) or only common spacers (bottom). **B.** Relationship between spacer set size (y-axis) and maximum spacer coverage in a set (x-axis).

axis), on a log10-log10 scale, for 10 randomly selected repeat clusters with ≥ 100 spacer sets. Colored lines show the results of a linear regression over all spacer sets for a given repeat cluster. **C.** Percentage of repeat clusters with ≥ 10 spacer sets showing a significant positive correlation between spacer set size and maximum spacer coverage (top), and distribution of the corresponding correlation coefficients (bottom). **D.** Distribution of the set size percentiles within their coverage bin across ecosystems. For each set, the set size was transformed in a percentile when compared to all other sets within the same coverage bin. The distribution of percentiles is then plotted by ecosystem (right) or CRISPR type (bottom left), to highlight ecosystems/types for which spacer sets tend to be always on the larger or smaller end of the distribution within their coverage bin. **E.** Distribution of set size for the 10 genera with the largest percentage of sets in the 80th size percentile or above. For each plot, the distribution of set sizes for all taxa is indicated by a boxplot, and the set size corresponding to the genera selected are indicated with colored dots (2 per plot).

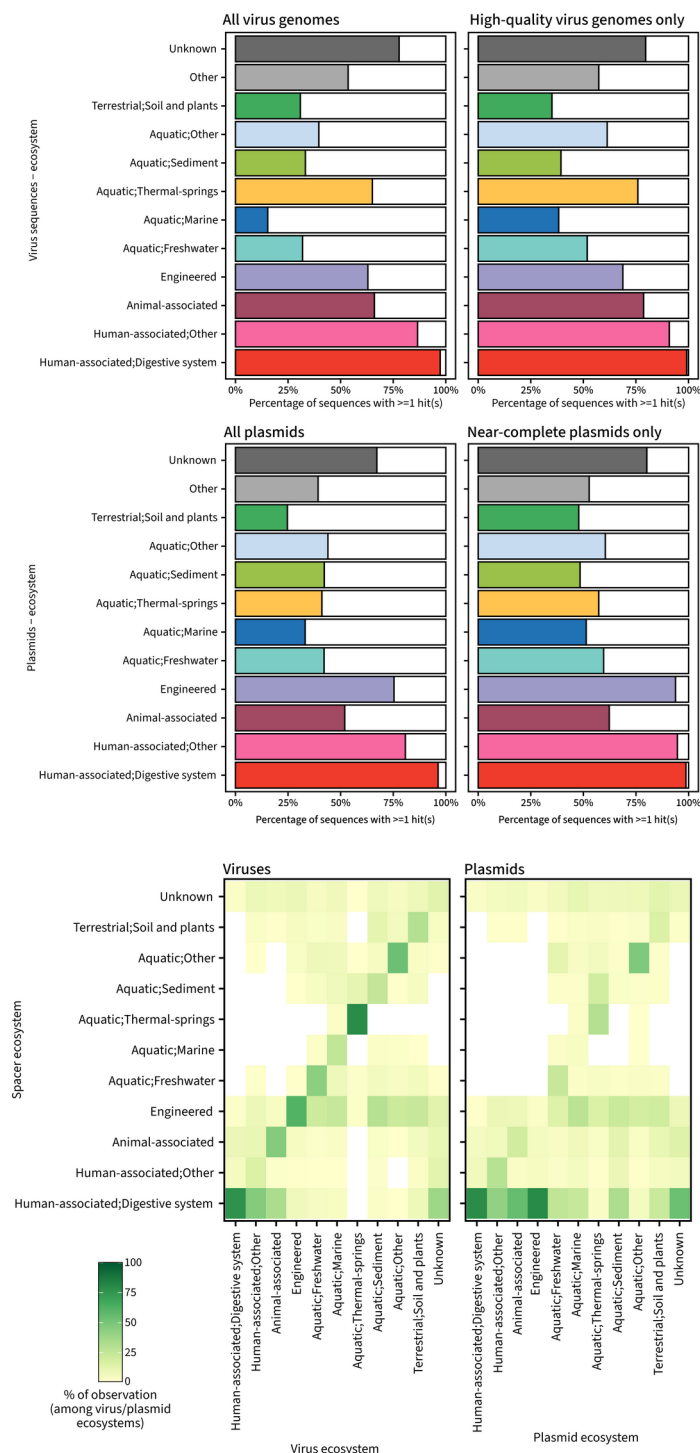


Figure S5. Spacer hits to IMG/VR and IMG/PR across ecosystems. The percentage of targets (top panel: viruses, middle panel: plasmids) with at least 1 hit in the global spacer database, based on the ecosystem from which the corresponding target was assembled from (y-axis). For each target, the left panel has all sequences (viruses or plasmids), and the right panel has high-quality sequences only (high-quality UViGs or near-complete plasmids). The bottom panel shows the frequency of matches organized by target ecosystem (x-axis) and spacer ecosystem (y-axis). The heatmap is colored based on the percentage of hits within the target ecosystem.

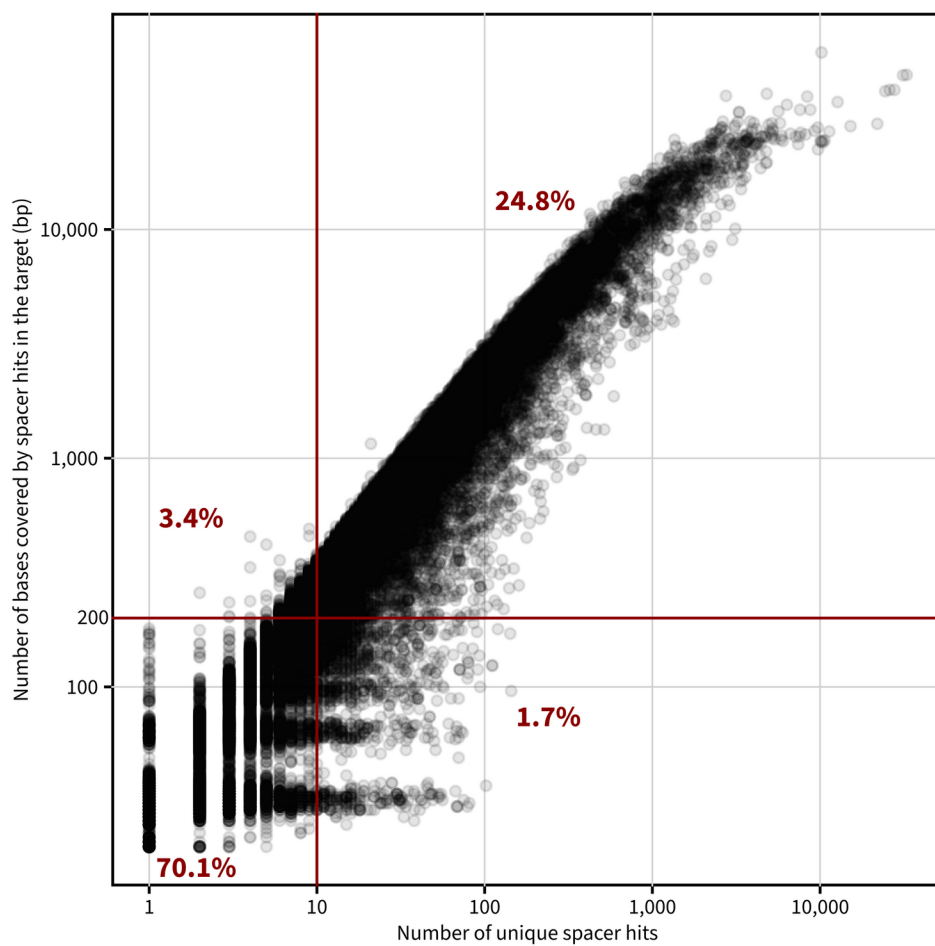
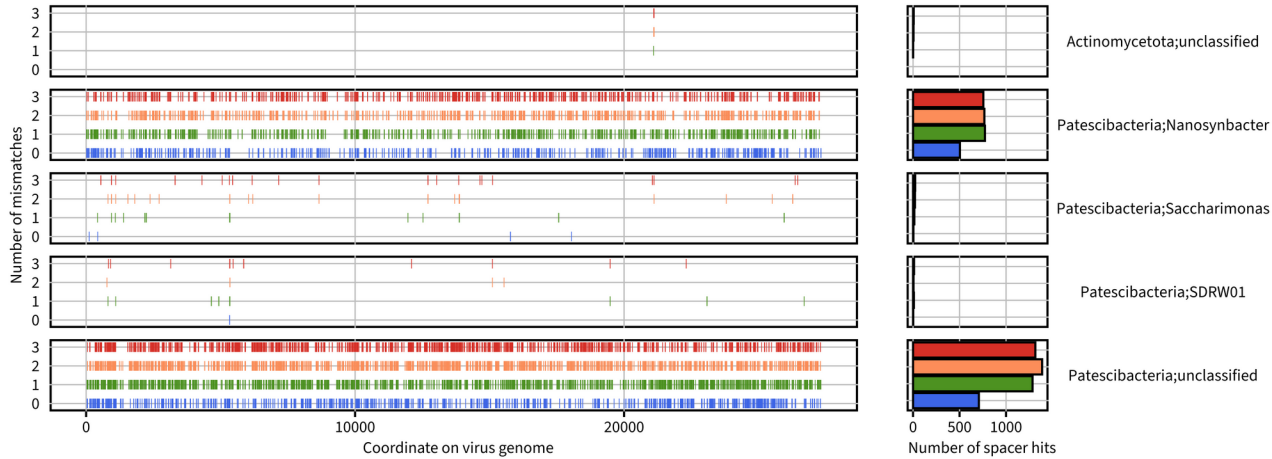
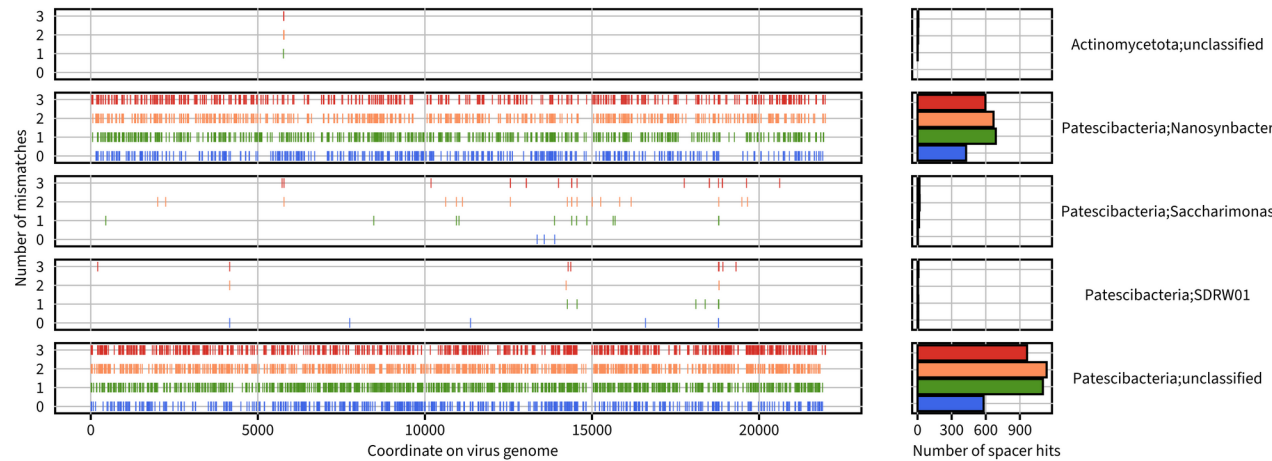


Figure S6. **Correlation between number of unique spacer hits and number of bases in the target covered by spacer hits.** The plot shows the relationship between the number of unique spacer hits (x-axis) and the number of bases covered by spacer hits (y-axis) for 100,000 randomly sampled pairs of high-quality viruses and repeat clusters. Red lines outline cutoffs of 10 distinct spacer hits and 200bp covered, and the percentage of observations within each part of the quadrant is indicated on the plot.

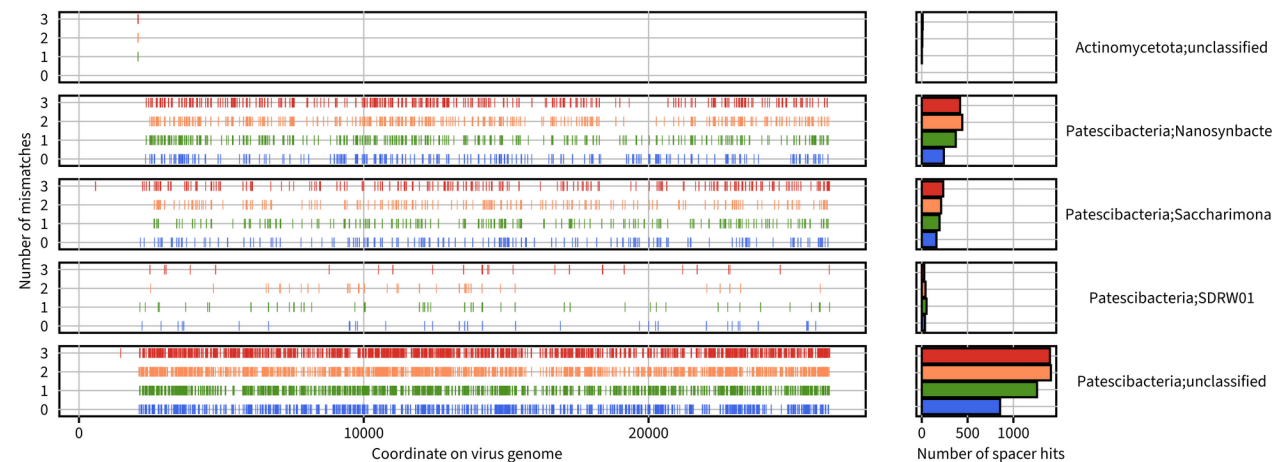
I24: IMGVR_UViG_3300008155_000024



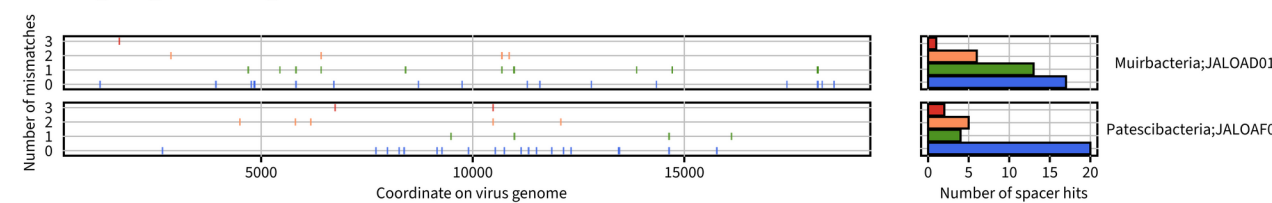
I20: IMGVR_UViG_3300008436_000020



I33: IMGVR_UViG_3300008506_000033

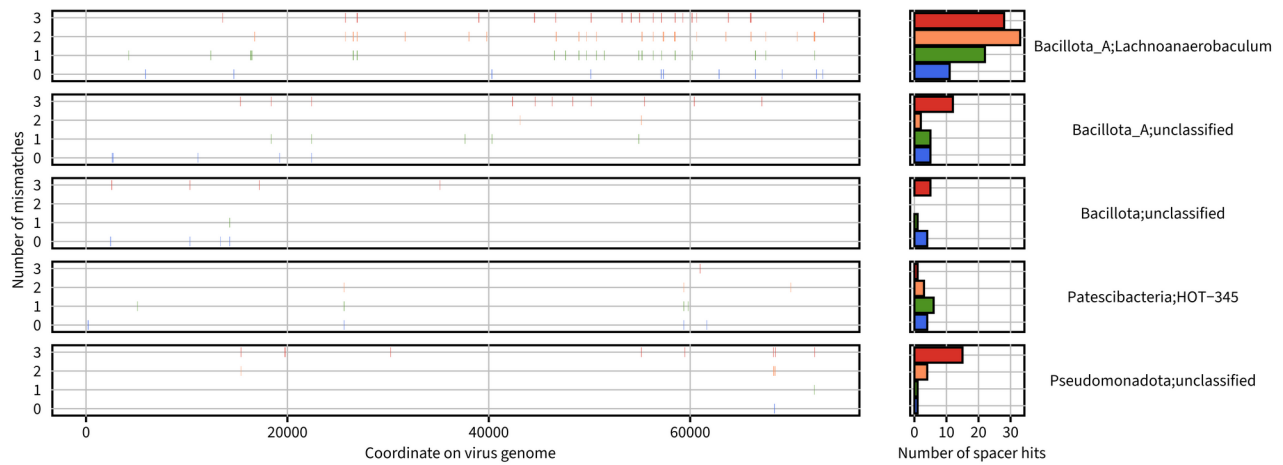


I49: IMGVR_UViG_3300028603_000349

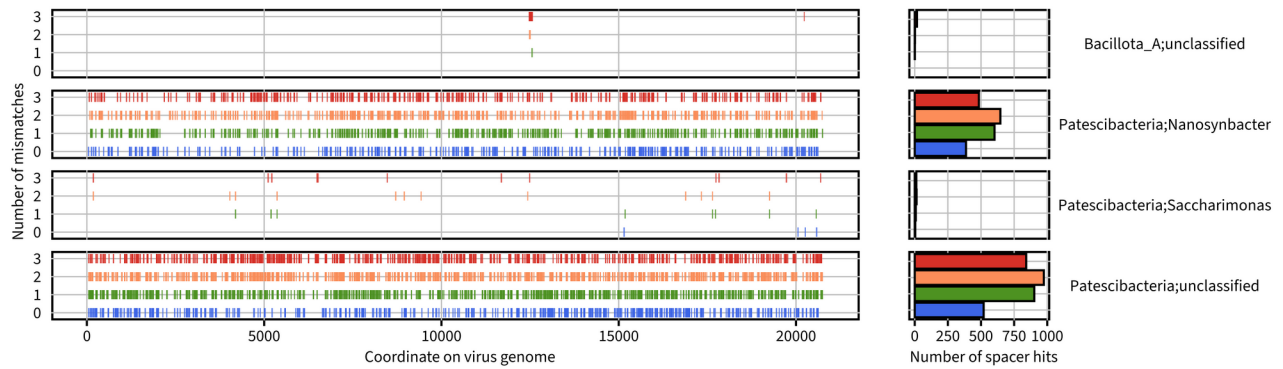


405 **Figure S7. Distribution of spacer hits on selected virus sequences.** For each virus sequence, the distribution of
spacer hits along the virus genome (x-axis) is indicated on the left side, with the number of mismatches
between spacer and virus sequences indicated on the y-axis. The right side shows the total number of spacer
hits for each number of mismatches across the entire sequence. Spacer hits are gathered by predicted taxon of
the corresponding repeat, at the genus rank (indicated on the right). For virus sequences, the simplified
410 identifier is indicated next to the full identifier in IMG/VR or UGV.

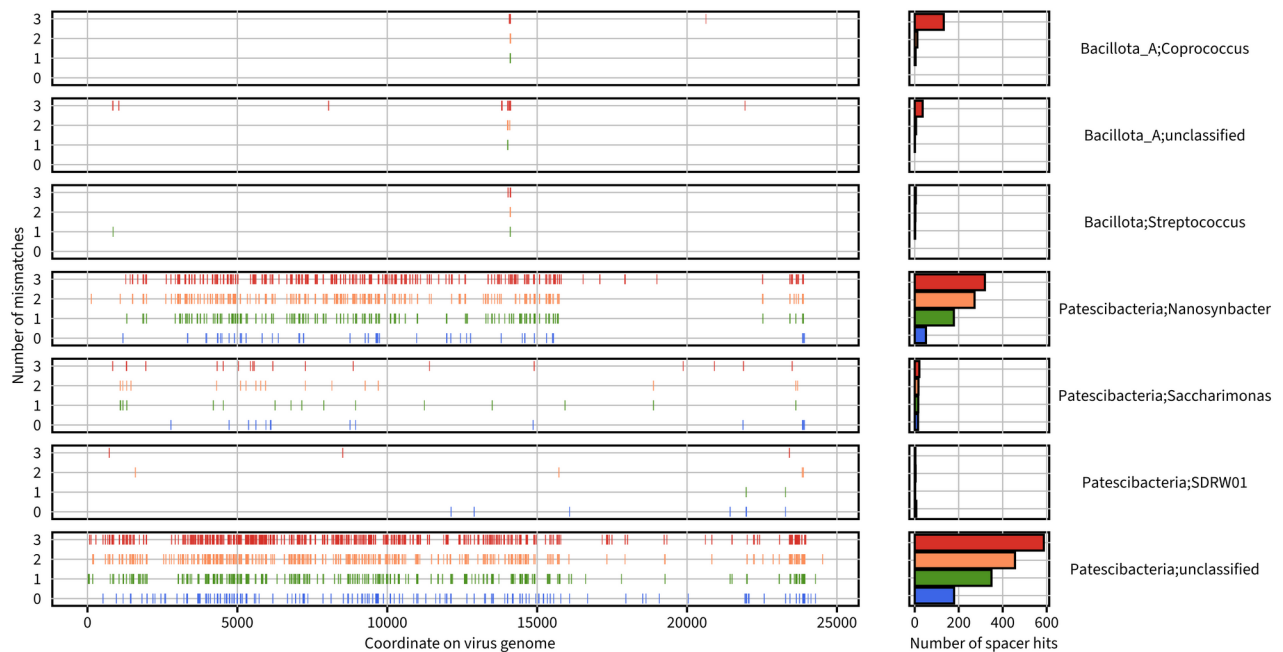
I14: IMGVR_UVIG_7000000432_000014



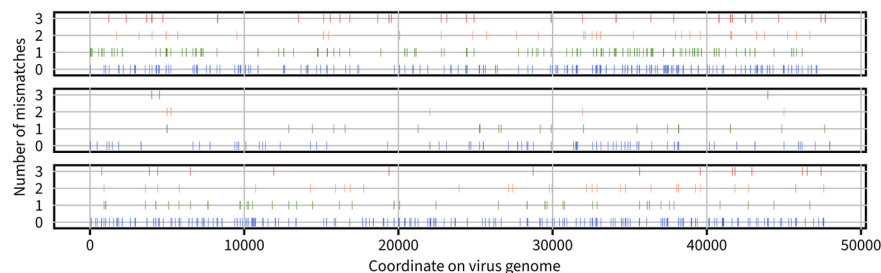
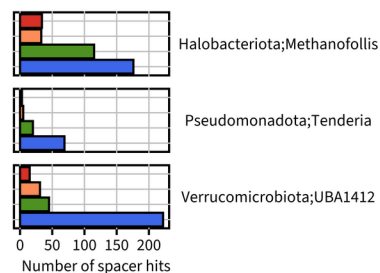
U060: UGV-GENOME-0152060



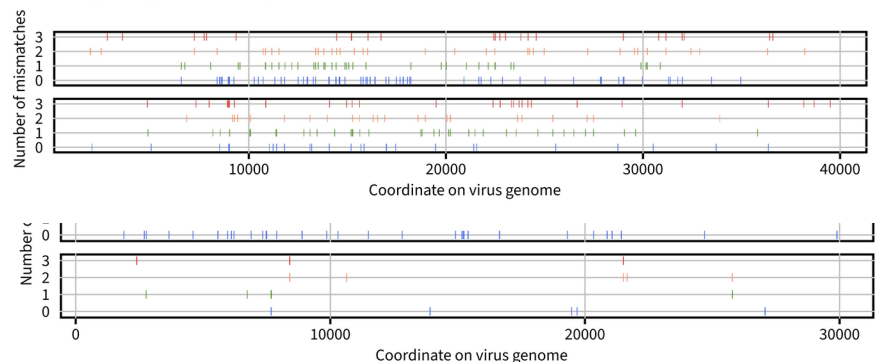
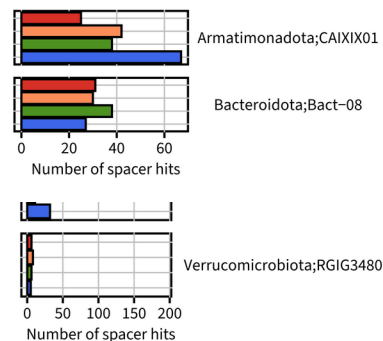
U105: UGV-GENOME-0183105



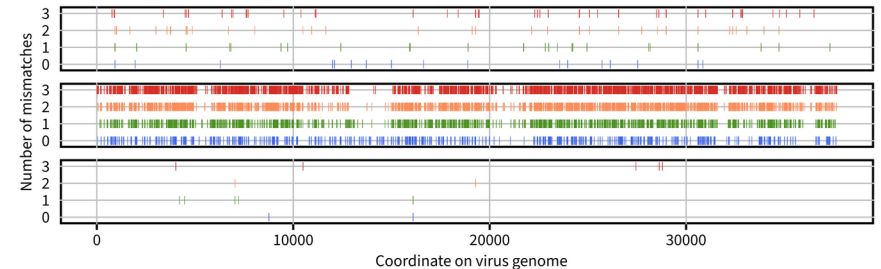
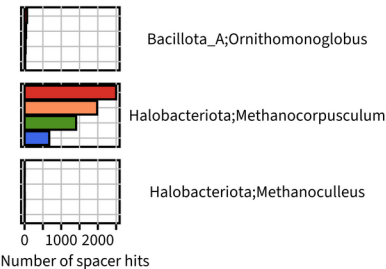
186: IMGVR_UViG_3300028603_000086



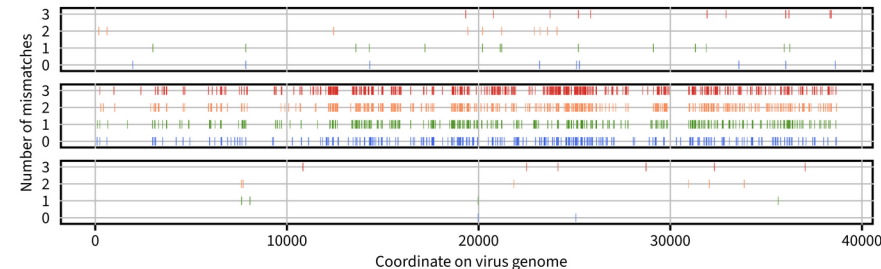
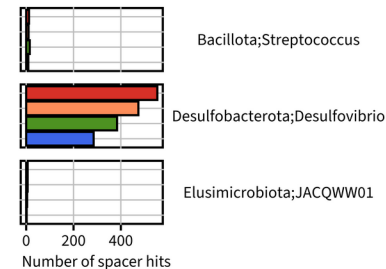
153: IMGVR_UViG_3300029288_007053



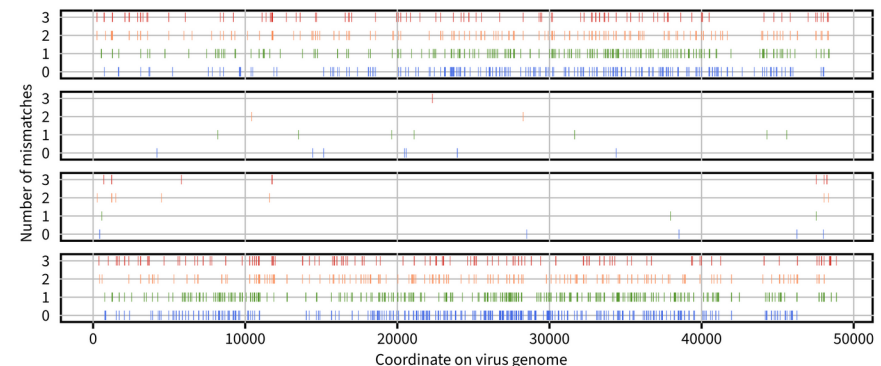
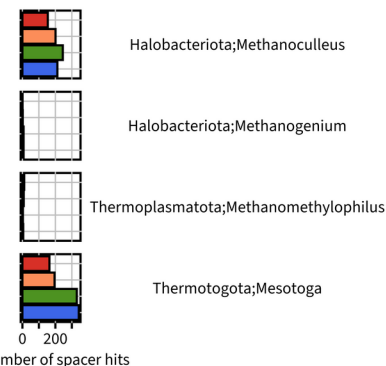
110: IMGVR_UViG_3300028601_000010



160: IMGVR_UViG_3300028601_000060



187: IMGVR_UViG_3300028602_002887



415 **Figure S8. Distribution of spacer hits on selected virus sequences.** For each virus sequence, the distribution of
spacer hits along the virus genome (x-axis) is indicated on the left side, with the number of mismatches
between spacer and virus sequences indicated on the y-axis. The right side shows the total number of spacer
hits for each number of mismatches across the entire sequence. Spacer hits are gathered by predicted taxon of
the corresponding repeat, at the genus rank (indicated on the right). For virus sequences, the simplified
420 identifier is indicated next to the full identifier in IMG/VR or UGV, and an asterisk is added next to the identifier
for viruses predicted to encode a DGR.

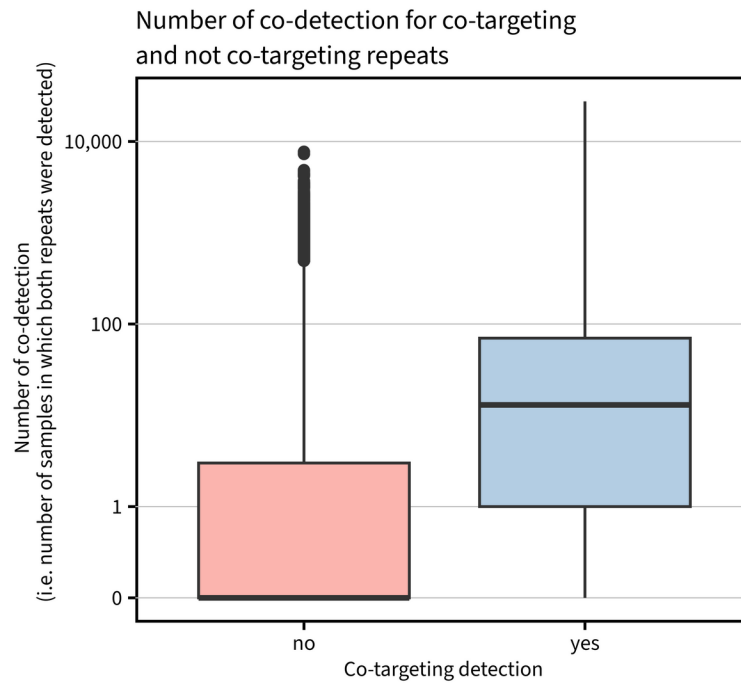


Figure S9. **Co-detection of CRISPR repeats assigned to distinct classes and targeting the same virus.** For all viruses targeted by repeats from distinct classes with ≥ 10 spacers, the corresponding individual repeats and their distribution across samples were collected. Next, repeat pairs were identified as co-targeting “yes” or “no” depending on whether they had at least one targeted virus in common (x-axis), and the number of samples in which both repeats were identified was tallied (y-axis). For the plot, only repeats identified in at least 100 samples were considered, and a random subsample of 350,000 repeat pairs was selected to balance the dataset between the co-targeting categories “yes” and “no”.

References

1. Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR–Cas systems. *Nat Rev Microbiol* **13**, 722–736 (2015).
2. Skennerton, C. T., Imelfort, M. & Tyson, G. W. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic acids research* **41**, e105 (2013).
3. Moller, A. G. & Liang, C. MetaCRAS: reference-guided extraction of CRISPR spacers from unassembled metagenomes. *PeerJ* **5**, e3788 (2017).
4. Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S. A. & Sørensen, S. J. CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. *The CRISPR Journal* **3**, 462–469 (2020).
5. Zhang, A.-N. *et al.* CRISPR-Cas spacer acquisition is a rare event in human gut microbiome. *Cell Genomics* **0**, (2024).
6. Pourcel, C., Salvignol, G. & Vergnaud, G. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology* **151**, 653–663 (2005).
7. Lopez-Sanchez, M.-J. *et al.* The highly dynamic CRISPR1 system of *Streptococcus agalactiae* controls the diversity of its mobilome. *Molecular Microbiology* **85**, 1057–1071 (2012).
8. Gudbergdottir, S. *et al.* Dynamic properties of the *Sulfolobus* CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. *Molecular Microbiology* **79**, 35–49 (2011).
9. Lam, T. J. & Ye, Y. Long reads reveal the diversification and dynamics of CRISPR reservoir in microbiomes. *BMC Genomics* **20**, 567 (2019).
10. Silas, S. *et al.* On the origin of reverse transcriptase- using CRISPR-Cas systems and their hyperdiverse, enigmatic spacer repertoires. *mBio* **8**, (2017).
11. Breusing, C. *et al.* Ecological differences among hydrothermal vent symbioses may drive contrasting patterns of symbiont population differentiation. *mSystems* **8**, e00284-23 (2023).
12. Eloë-Fadrosch, E. A. *et al.* Global metagenomic survey reveals a new bacterial candidate phylum in geothermal springs. *Nat Commun* **7**, 10476 (2016).
13. Schaible, G. A. *et al.* Multicellular magnetotactic bacteria are genetically heterogeneous consortia with metabolically differentiated cells. *PLOS Biology* **22**, e3002638 (2024).
14. Mick, E., Stern, A. & Sorek, R. Holding a grudge: Persisting anti-phage CRISPR immunity in multiple human gut microbiomes. *RNA Biology* **10**, 900–906 (2013).
15. López-Beltrán, A., Botelho, J. & Iranzo, J. Dynamics of CRISPR-mediated virus–host interactions in the human gut microbiome. *ISME J* **18**, wræ134 (2024).
16. Kirchberger, P. C. & Ochman, H. Microviruses: A World Beyond phiX174. *Annual Review of Virology* **10**, 99–118 (2023).
17. Roux, S. *et al.* Cryptic inoviruses revealed as pervasive in bacteria and archaea across Earth's biomes. *Nat Microbiol* **4**, 1895–1906 (2019).
18. Neri, U. *et al.* Expansion of the global RNA virome reveals diverse clades of bacteriophages. *Cell* **185**, 4023–4037.e18 (2022).

19. Makarova, K. S. *et al.* Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol* **18**, 67–83 (2020).
20. Rybnicky, G. A., Fackler, N. A., Karim, A. S., Köpke, M. & Jewett, M. C. Spacer2PAM: A computational framework to guide experimental determination of functional CRISPR-Cas system PAM sequences. *Nucleic Acids Research* **50**, 3523–3534 (2022).
21. Leenay, R. T. *et al.* Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. *Molecular Cell* **62**, 137–147 (2016).
22. Aviram, N., Thornal, A. N., Zeevi, D. & Marraffini, L. A. Different modes of spacer acquisition by the *Staphylococcus epidermidis* type III-A CRISPR-Cas system. *Nucleic Acids Res* **50**, 1661–1672 (2022).
23. Chylinski, K., Le Rhun, A. & Charpentier, E. The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. *RNA Biology* **10**, 726–737 (2013).
24. Briner, A. E. *et al.* Occurrence and Diversity of CRISPR-Cas Systems in the Genus *Bifidobacterium*. *PLOS ONE* **10**, e0133661 (2015).
25. Karah, N. *et al.* CRISPR-cas Subtype I-Fb in *Acinetobacter baumannii*: Evolution and Utilization for Strain Subtyping. *PLOS ONE* **10**, e0118205 (2015).
26. Shariat, N., Timme, R. E., Pettengill, J. B., Barrangou, R. & Dudley, E. G. Characterization and evolution of *Salmonella* CRISPR-Cas systems. *Microbiology* **161**, 374–386 (2015).