Title

Long-term CRISPR locus dynamics and stable host-virus co-existence in subsurface fractured shales

Permalink

https://escholarship.org/uc/item/9369m9pk

Journal

Current Biology, 33(15)

ISSN

0960-9822

Authors

Amundson, Kaela K
Roux, Simon
Shelton, Jenna L
et al.

Publication Date

2023-08-01

DOI

10.1016/j.cub.2023.06.033

Current Biology

Article

Long-term CRISPR locus dynamics and stable host-virus co-existence in subsurface fractured shales

Highlights

- Host and viral genomes were recovered from temporal sampling of six shale wells
- CRISPR loci were identified in the majority of host genomes
- Viral linkages were made via CRISPR spacers for 90 host genomes spanning 25 phyla
- Among linked genomes, occurrences of host-viral co-existence increased temporally

Authors

Kaela K. Amundson, Simon Roux,
Jenna L. Shelton, Michael J. Wilkins

Correspondence

kaela.amundson@colostate.edu

In brief

Amundson et al. analyze microbial and viral communities from temporal sampling of subsurface shale wells and use spacers from CRISPR-Cas systems to make linkages between host and viral populations. They identify abundant CRISPR-Cas loci at community and host-population levels and observe increases in host-viral co-existence through time.

Amundson et al., 2023, Current Biology *33*, 3125–3135
August 7, 2023 © 2023 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.cub.2023.06.033

CellPress

Current Biology

CellPress
OPEN ACCESS

Article

Long-term CRISPR locus dynamics and stable host-virus co-existence in subsurface fractured shales

Kaela K. Amundson,[1,4,*] Simon Roux,[2] Jenna L. Shelton,[3] and Michael J. Wilkins[1]
[1]Colorado State University, Department of Soil & Crop Sciences, 301 University Ave., Fort Collins, CO 80523, USA
[2]DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA
[3]United States Geological Survey, 12201 Sunrise Valley Dr., Reston, VA 20192, USA
[4]Lead contact
*Correspondence: kaela.amundson@colostate.edu
https://doi.org/10.1016/j.cub.2023.06.033

SUMMARY

Viruses are the most ubiquitous biological entities on Earth. Even so, elucidating the impact of viruses on microbial communities and associated ecosystem processes often requires identification of unambiguous host-virus linkages—an undeniable challenge in many ecosystems. Subsurface fractured shales present a unique opportunity to first make these strong linkages via spacers in CRISPR-Cas arrays and subsequently reveal complex long-term host-virus dynamics. Here, we sampled two replicated sets of fractured shale wells for nearly 800 days, resulting in 78 metagenomes from temporal sampling of six wells in the Denver-Julesburg Basin (Colorado, USA). At the community level, there was strong evidence for CRISPR-Cas defense systems being used through time and likely in response to viral interactions. Within our host genomes, represented by 202 unique MAGs, we also saw that CRISPR-Cas systems were widely encoded. Together, spacers from host CRISPR loci facilitated 2,110 CRISPR-based viral linkages across 90 host MAGs spanning 25 phyla. We observed less redundancy in host-viral linkages and fewer spacers associated with hosts from the older, more established wells, possibly reflecting enrichment of more beneficial spacers through time. Leveraging temporal patterns of host-virus linkages across differing well ages, we report how host-virus co-existence dynamics develop and converge through time, possibly reflecting selection for viruses that can evade host CRISPR-Cas systems. Together, our findings shed light on the complexities of host-virus interactions as well as long-term dynamics of CRISPR-Cas defense among diverse microbial populations.

INTRODUCTION

Viruses are abundant and important constituents of microbial communities in nearly all ecosystems. Consequently, bacteria and archaea, like all living things, are subject to near constant threat of viral predation. In response, many bacteria (~40%–60%) and archaea (~90%) deploy CRISPR-Cas viral defense systems.[1–4] CRISPR-Cas works by recording memories of viral interactions via integration of small pieces of viral DNA (“spacers”) within the hosts’ CRISPR array that are interspaced with identical repeat sequences and flanked by Cas (CRISPR-associated) genes.[5–13] These saved memories help to protect the host against recurrent invasion by the same viral population by more rapidly identifying and degrading the invading nucleic acids, analogous to antibodies in the human immune system.[5–7,10,12,14]

Spacers within CRISPR arrays therefore provide a record of past interactions between a host and viral population, and host-viral linkages can be made by matching the hosts’ CRISPR spacers to protospacers in viral genomes.[11,15–27] However, the presence of CRISPR-Cas systems within the microbial community is often a limiting step to making strong host-virus connections; CRISPR-Cas defense is most likely advantageous in ecosystems where host and viral populations repeatedly interact, such as environments dominated by biofilms or those hosting lower microbial and viral diversity.[28–30] Additionally, CRISPR-Cas has been shown to be more widespread in some ecosystems relative to others, such as anoxic environments or those with elevated temperatures.[14,28,31–33]

Despite the important role of CRISPR-Cas in viral defense, much remains to be understood about CRISPR-Cas frequency, size, and how the presence of these defense systems might influence the temporal dynamics of host and viral populations within diverse microbial communities. Successful incorporation of a spacer should provide the host with future defense upon interaction with the same viral population. However, there are many factors that may influence CRISPR-Cas defense function. For example, CRISPR arrays do not grow exponentially and spacers can indeed be lost,[34–37] and host-viral co-existence—despite CRISPR-Cas defense—has been observed.[38] Additionally, spacers nearest to the leading end of the CRISPR arrays are most likely to be effective, as they typically represent more recent viral interactions with less time for mutations to occur within the viral protospacer, although recombination can also influence CRISPR array architecture.[8] Thus, it has been hypothesized and shown in laboratory experiments that select spacers may be more favorably retained if they target evolutionarily conserved portions of the viral genome, providing more effective long-term viral defense.[13,39] Although ecosystem resources and genome size are not necessarily limiting factors to array size,[40,41] other studies have modeled the optimum CRISPR cassette size based on other factors, such as viral diversity and trade-offs between Cas machinery and array size.[42–44] As a result, it has been suggested that maintaining smaller arrays, in the order of a few dozen to a hundred spacers, may be the optimal size for CRISPR arrays that provide broad protection against a range of viruses but do not overwhelm CRISPR-Cas machinery.[42,44,45] However, many of these insights are derived from modeling or laboratory experiments, and there remains a need to understand patterns of CRISPR-Cas defense in environmental systems with diverse microbial communities.

To address this knowledge gap, we used a temporally resolved dataset from six fractured shale wells to interrogate host-virus dynamics and CRISPR-Cas loci in a subsurface ecosystem. Subsurface fractured shales, which are relatively closed ecosystems with limited immigration, elevated temperatures, lower microbial diversity, and likely dominated by biofilms, present an opportunity to address these questions through strong CRISPR-based host-viral linkages.[21,23,46–49] We hypothesized and found that CRISPR-Cas viral defense systems were widely encoded across hosts within shale microbial communities. Building on this, we applied multiple bioinformatic approaches to identify CRISPR spacers in both recovered host genomes and metagenomes and made strong host-virus linkages for many of the recovered host genomes. This approach also facilitated temporal investigations into host utilization of CRISPR-Cas at the community level and host-population levels to better understand CRISPR-Cas defense in this natural ecosystem. Finally, we leveraged over 2,000 viral linkages to investigate host-viral dynamics and saw evidence for increased host-virus co-existence through time. To our knowledge, our study represents one of the most extensive analyses of long-term, host-viral temporal dynamics with CRISPR-based linkages in an environmental system to date.

RESULTS AND DISCUSSION

Fractured shale ecosystems provide a unique opportunity to investigate virus-host temporal dynamics

We sampled fluids from two sets of hydraulically fractured oil and gas wells in the Denver-Julesburg (DJ) Basin (Colorado, USA) for nearly 800 days. The two sets of wells were defined by their age relative to the initial fracturing process: the "established" wells operated for nearly 3 years prior to the initiation of our sampling campaign (DJB-1, DJB-2, DJB-3), while we began sampling the "new" wells shortly after they had been hydraulically fractured (DJB-4, DJB-5, DJB-6) (Figure 1). All three wells within each group were located on the same frack pad and subject to the same drilling and hydraulic fracturing process, resulting in three replicate wells for each group.

From a total of 78 metagenomes across all six wells (Figure 1; Data S1), we recovered 202 unique metagenome assembled genomes (MAGs) representing 29 phyla and 2,176 unique viral MAGs (vMAGs) > 10 kb from the subsurface communities. The microorganisms that persist in this ecosystem likely originate from water, sand, and chemical inputs used during the hydraulic fracturing process. Many of the dominant and persisting MAGs—encompassing bacterial taxa affiliated with *Clostridia*, *Thermotogae*, *Fusobacteriia*, and *Synergistia*, and archaeal taxa affiliated with *Methanosarcinia*, *Methanomicrobia*, and *Thermococci* (Figure S1)—have been reported in other engineered subsurface environments,[21,50–54] and their relative abundances in this system reflected patterns observed in complementary 16S rRNA gene analyses (Figure S1). Although vMAGs were recovered from metagenomes and not viromes, only a small portion of viruses were predicted to be temperate by presence of integrase genes (n = 192) or hidden Markov model (HMM) searches of domains associated with temperate viruses (n = 293).[55]

Microbial communities, however, are not static through time. Taxa unable to tolerate high temperatures and elevated salinity are likely outcompeted, while biofilms and spatially distinct niches likely emerge and expand in this closed ecosystem.[56,57] Thus, we expect that microbial communities within the established wells are more spatially heterogeneous, while microbial communities in the new wells are initially well mixed and more spatially homogeneous.[49,58] In agreement with these assumptions, we observed higher host (bacterial and archaeal) and viral alpha diversity in the established wells relative to the new wells (Wilcoxon, $p = 5.563e{-}06$) (Figure S2). Alpha diversity also generally increased through time in all wells, likely reflecting the development of niches fostering more diverse taxa (Figure S2). This trend contrasts findings from previous fractured shale studies that reported a rapid decrease in microbial diversity.[46,47] More broadly, host alpha diversity in the DJ Basin was also higher than many other fractured shale ecosystems studied to date[50,52,59,60] and similar to those reported previously for produced fluids from the DJ Basin.[51] Notably, microbial communities from the new wells became more similar to those in the established wells over time, likely reflecting the maturation of the well ecosystem (Figure S2). Together, these results illustrate the temporal juxtaposition of the two sets of wells and the connectedness of host-viral dynamics in increasingly diverse microbial communities within a closed, subsurface ecosystem.

Evidence for active viral predation in deep subsurface shales' microbial communities

Although community composition did vary with time, host and viral community dynamics generally mirrored one another (Figure 2). We quantified these temporal changes in community structure using Bray-Curtis dissimilarity values (Figure 2). In this analysis, higher dissimilarity values indicate greater change in community composition relative to the previous time point. Shifts in host and viral communities were strongly and positively correlated, often mirroring one another in their dynamics (Figure 2)—a trend which was also reflected in host and viral alpha diversity (Spearman's Rho: established wells = 0.71, 0.96, 0.076; new wells = 0.6, 0.66, 0.91). Viruses depend on their hosts for replication, yet host populations are often impacted as a direct result of this proliferation. Thus, the strong relationship observed here suggests that host and viral communities were continually changing and that viral predation was likely

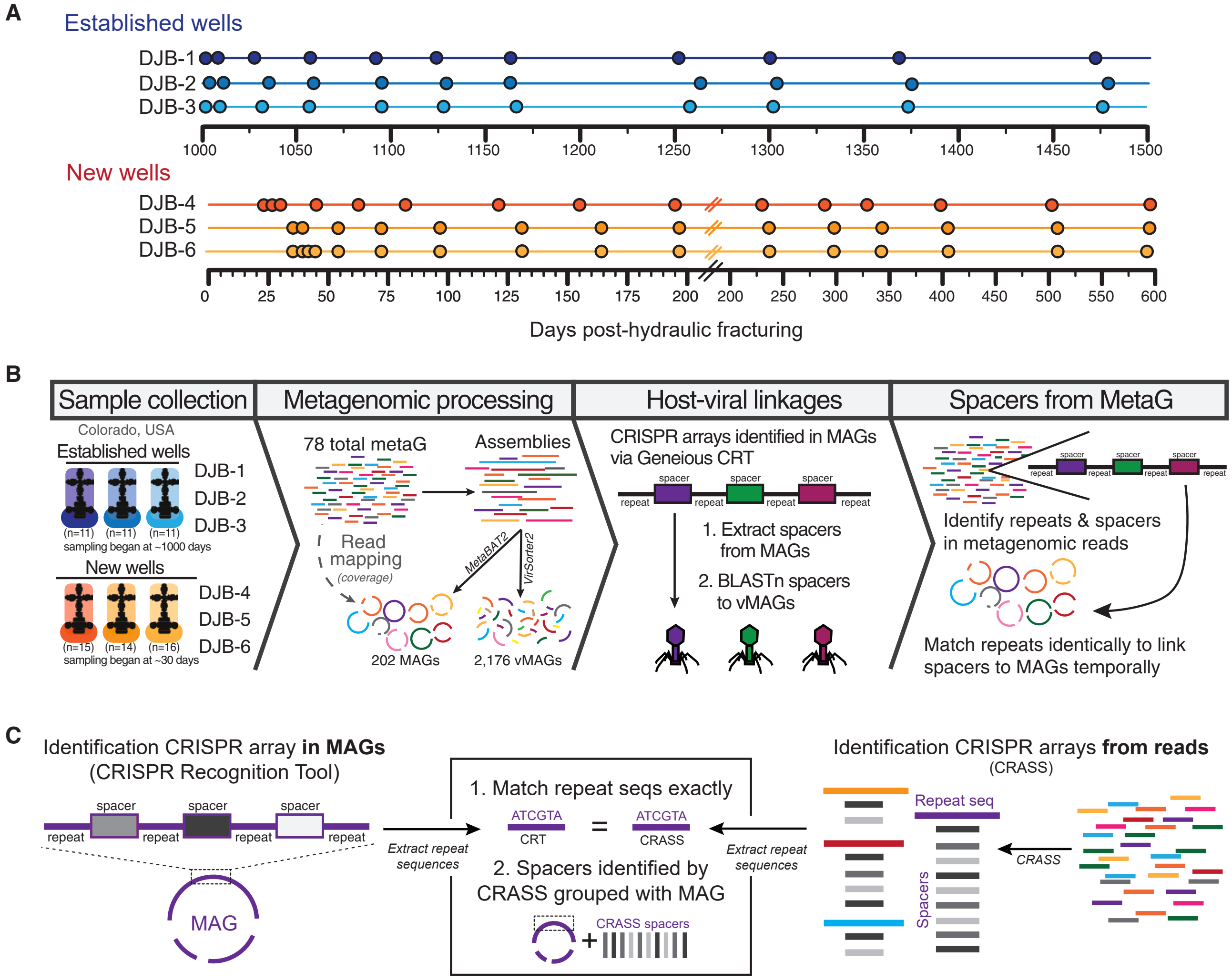

Figure 1. Sampling scheme and methods

(A) Sampling time points for all six wells split by their age group. Dashes at 200 days on the axis of the new wells sampling scheme represents a change in scale.

(B) Overview of methods used to recover representative host and viral genomes, make linkages between MAGs and vMAGs, and identify spacers from metagenomic reads for a community-level insight.

(C) Overview of methods used to link additional spacers, identified with CRASS, to representative host genomes.

See also Figure S1.

occurring. Interestingly, we did not observe trends toward community stability in either grouping of wells, which would be indicated by consistently decreasing Bray-Curtis dissimilarity values. Finally, there was a stronger relationship between host and viral communities in the established wells compared with the new wells, potentially reflecting the temporal loss of viral populations that lack hosts and subsequent enrichment of interdependent host and viral populations.

Only a small portion of recovered viruses encoded genes indicative of a temperate lifestyle (Data S3), suggesting that these strong correlations are not driven by integrated prophage. A temperate lifestyle may not be necessary for a virus to survive in this ecosystem, as has been reported for other, more extreme environments due to the availability of growth substrates (i.e., organic carbon) from additives used in the fracturing process and reduced environmental stress on microbiomes due to the lower salinity of the DJ Basin (avg. 47 mS/cm). Additionally, only a very small proportion of vMAGs that encoded an integrase gene (<5 in each well) closely matched their hosts' coverage, indicating that temperate viruses are unlikely to be solely responsible for the trends observed. Instead, lytic viruses are likely recovered during the filtration step in sample processing as produced fluids are often viscous, containing small particulates that clog the filter pores and elevated levels of ferric iron that can bind viral capsids.[23,61] Thus, the strong relationships observed here are likely due to dynamic virus-host interactions and driven by both lytic and lysogenic infections.

Community-level responses to viral interactions recorded by CRISPR-Cas arrays

Viral diversity has been shown to impact the success of, and selection for, CRISPR-Cas systems, as the "memory" recording of

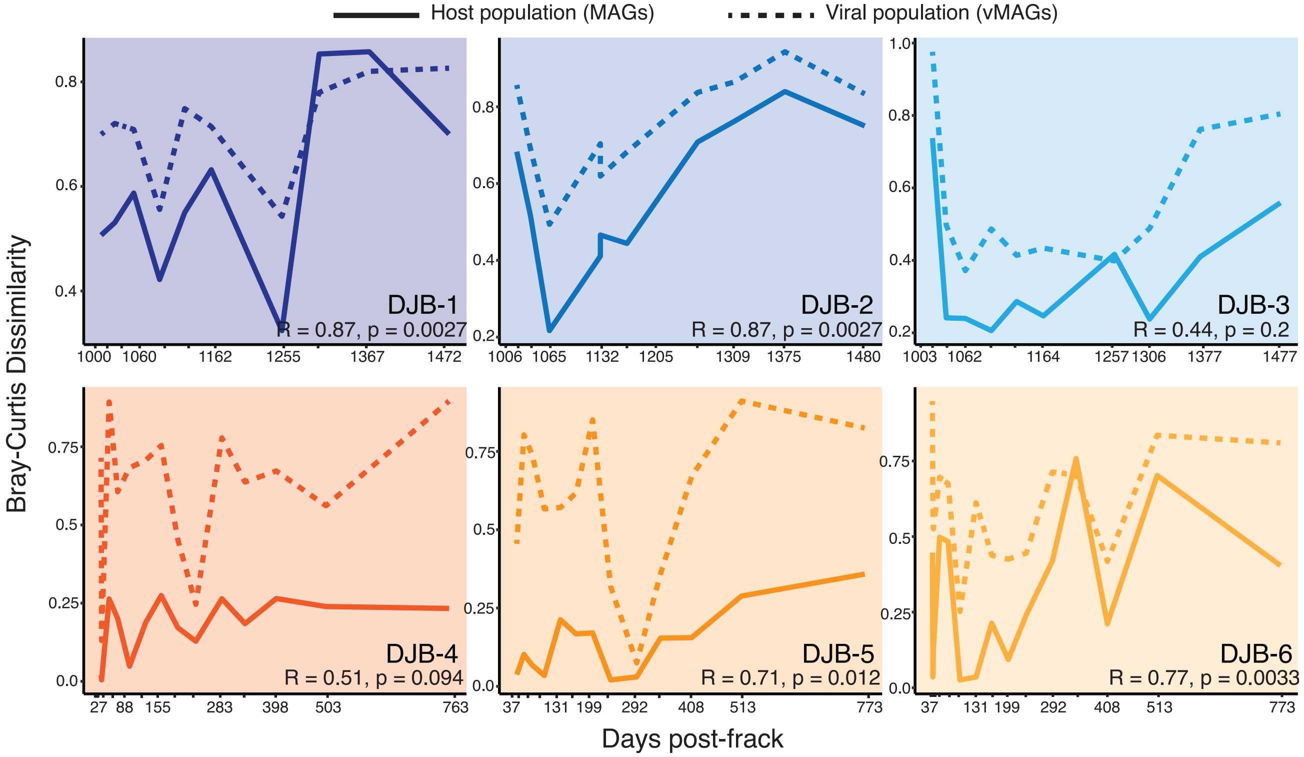

Figure 2. Temporal dynamics of host (bacterial and archaeal MAGs) and viral (vMAGs) communities

Bray-Curtis community dissimilarity through time for both host (solid line) and viral (dashed line) communities illustrating the change in each community composition from the previous time point, with larger dissimilarity values indicating greater change in community structure. Spearman's rho and p values highlight the strong positive relationship between the temporal changes in host and viral populations in all wells.
See also Figure S2.

a viral interaction as an integrated spacer is more effective in ecosystems where repeated interactions between host and viral populations occur.[29] The closed nature of these subsurface ecosystems should promote such repeated interactions, and thus we sought to identify evidence of hosts using CRISPR-Cas defense at the community level.

First, repeats and spacers from CRISPR-Cas arrays (which tend to break during metagenomic assembly) were identified in all samples using CRASS.[62] Overall, we recovered a total of 918,724 spacers from all 78 metagenomes. All six wells had a significant positive relationship between the number of spacers recovered and time, especially in the new wells, where the total number of spacers rapidly increased through the first 100 days (Figure 3). Although the established wells initially contained a greater total number of spacers compared with the new wells, at later time points, the new wells began to approach the established wells' totals, highlighting the speed at which spacers may be incorporated by microorganisms.

Finally, linking this temporal relationship with observations that the viral community is continually changing, we observed that the number of spacers also had a significant and strong positive relationship with the total number of unique viral populations (vMAGs) across all samples (Figure 3). Together, these results demonstrate the presence of widely encoded CRISPR-Cas defense systems in the microbial community and suggest that microorganisms using CRISPR-Cas are likely responding to ongoing viral interactions and integrating matching spacers into CRISPR arrays through time.

CRISPR-Cas and other viral defense systems within host genomes

Moving beyond community-level analyses, we sought to link individual CRISPR-Cas viral defense systems to representative host MAGs. Our host MAGs, by nature, are composite genomes that likely represent a host population as opposed to a single host cell. Linking spacers assembled from metagenome reads to CRISPR-Cas loci associated with a given MAG thus provides an overview of the total complement of spacers encoded by members of this population—and a window into the population-level diversity and dynamics at these CRISPR-Cas loci. In total, 123 of our 202 MAGs (~60%) spanning 25 phyla contained a detectable CRISPR array (Figure 4; Data S2). We identified CRISPR-Cas loci in a higher proportion of MAGs from the established wells (67%) relative to the new wells (54%), highlighting the persistent widespread use of this defense system in an environment where viral predation is likely recurrent. Type I-B was the most common CRISPR-Cas system type (22%) out of all those identified and classified, followed by type III-A (17%) (Data S2). The high proportion of hosts containing a CRISPR-Cas array is not unexpected, as CRISPR-Cas systems are reported to be more widely encoded in closed ecosystems,

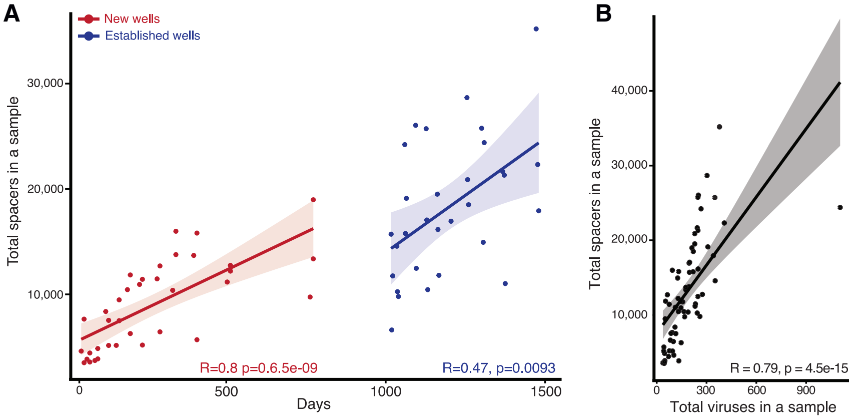

Figure 3. Community-level responses to viral interactions through time as recorded by CRISPR-Cas loci within the microbial community

(A) Spearman's correlation between the number of spacers recovered in a sample and the number of viruses in a sample.

(B) Spearman's correlations between the number of spacers in a sample and days for the new and established wells.

See also Figure S3.

biofilms, and ecosystems with elevated temperatures—three characteristics of fractured subsurface shales.

We next leveraged CRASS to identify additional spacers associated with our host MAGs.[62] Briefly, repeat sequences identified within host arrays were matched to those identified with CRASS, and spacers grouped to the CRASS-identified repeat were then associated with the MAG, representing a host population (Figure 1). This approach identified thousands of additional spacers from metagenomic reads and associated many of these spacers with host genomes from as many time points as possible. Importantly, this facilitated the identification of temporal trends in CRISPR-Cas loci sizes, as insights into host arrays were not limited to a single time point where the host MAG was recovered.

For many host populations, the number of CRISPR spacers generally exhibited a strong positive relationship with MAG coverage, a proxy for relative abundance in the community (Figure S3). This trend was obscured when MAGs were grouped together at higher taxonomic levels (Figure S3), highlighting the importance of genome-resolved analyses and the need for even more resolved analyses, possibly at the single-cell level, to better understand loss and gain of spacers through time in natural communities. In contrast, there was greater variability in the relationship between spacers and time at the genome level (Figure S3), with few host populations exhibiting consistent increases or decreases in the number of spacers over time. The overall increase in number of CRISPR spacers through time observed at the community levels (Figure 2) is thus probably not only due to increases in spacer number within individual populations but also to an overall increase in host populations encoding active CRISPR-Cas systems.

CRISPR-Cas is one of many viral defense systems, many of which have been only recently described.[63] Not all MAGs encoded a detectable CRISPR array, and thus we hypothesized that other viral defense systems are likely deployed by hosts. We found that 87% of all MAGs contained another viral defense system, with no significant difference in the proportion of MAGs between the new and established wells (Data S2). Most of the MAGs that contained a CRISPR array also tended to encode another defense system in both the new and established wells (81% and 78% of MAGs, respectively). However, there was a greater diversity of different viral defense systems encoded in the established wells compared with the new wells, with 41 and 34 different systems detected across MAGs from the established and new wells, respectively.[63] In all wells, the most common viral defense system was a restriction modification system, which works more promiscuously than CRISPR-Cas to degrade nucleic acids,[64] while known abortive systems that induce cell death (i.e., Abi2, AbiEii) were a small proportion of the overall number of systems in both established (6%) and new (4%) wells.[65,66] Together, this provides further evidence that there is a benefit to the host for encoding viral defense mechanism(s) and offers insights into the types of defense systems that may be paired with CRISPR-Cas in an environmental system.

Fewer CRISPR spacers associated with hosts in the established wells may reflect selection toward more effective CRISPR-Cas arrays

With continued interactions between hosts and viruses, we hypothesized that MAGs (representing host populations) in the established wells would generally be associated with more CRISPR spacers reflecting these events. However, loss of spacers through time as well as a theoretical optimum for CRISPR array size[34–36,42,44] have both been previously reported, which could lead to populations from established wells encoding fewer spacers. Here, despite higher host and viral diversity in the established wells, we observed on average a greater number of spacers per host population in the new wells (avg. 288 ± 410), compared with the established wells (avg. 180 ± 306) (Figure S4). The high number of spacers associated with hosts may be due to MAGs representing host populations rather than a single host cell and likely mirrors trends within the subsurface host populations and reflects differences in terms of population diversity among the group of wells for CRISPR loci.

A greater number of viral linkages were also made per representative host MAG from the new wells relative to the established wells, averaging 22 and 7 unique viral linkages, respectively (Figure S4). We hypothesize that fewer linkages per host in the established wells may be driven by interactions with fewer different viruses, potentially due to more heterogeneous and confined spatial structures where host and viral populations interact. Although MAGs from the established wells encoded fewer spacers and linked to fewer viruses, we observed less redundancy within those viral linkages. For all linkages made (i.e., every host linked to any virus) for MAGs from the established wells, an average of 71% of those linkages were to unique viruses (Figure 5). In the new wells, we

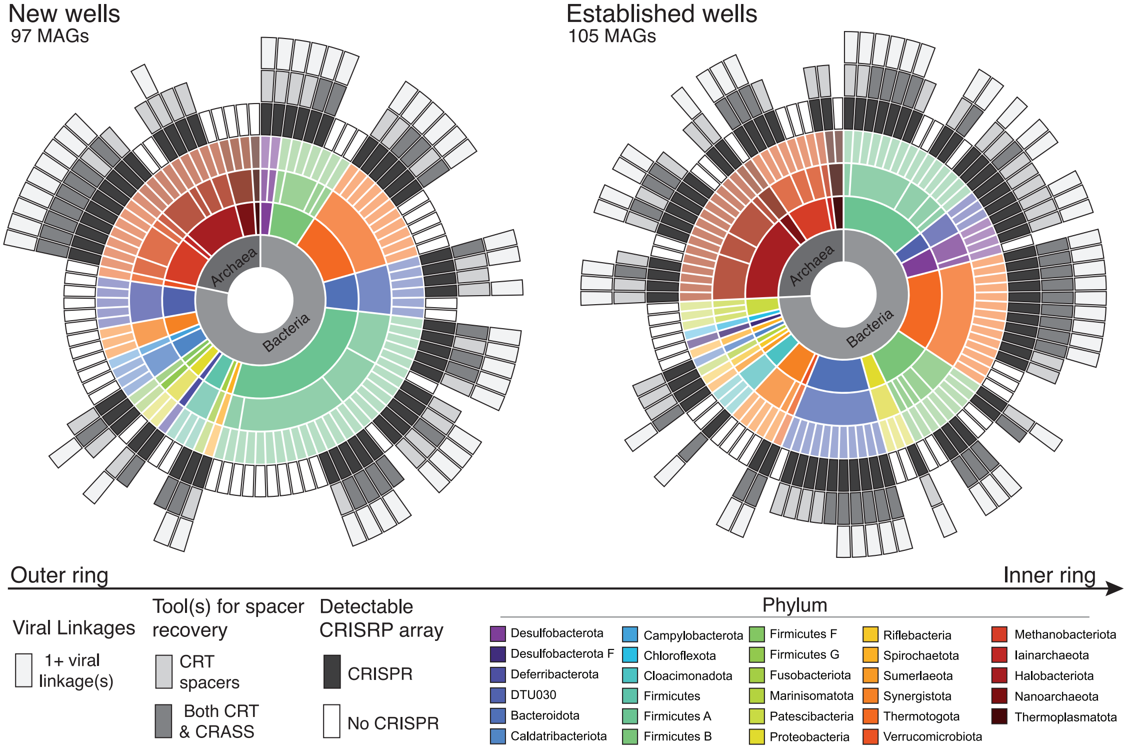

Figure 4. CRISPR arrays and viral linkages in representative host genomes (MAGs) recovered in the new and established wells, organized by phylum

The first four rings (from inside out) provide information on the taxonomy of host MAGs and how many MAGs were recovered for a given phyla. Inner-most ring splits MAGs by bacteria or archaea. The second ring illustrates the different phyla represented. The third ring illustrates how many different taxonomic classes are represented by MAGs from a given phyla. And the fourth ring shows how many individual host MAGs were recovered for a given phyla and class. The last three rings present information on CRISPR-Cas arrays (presence/absence), how spacers were recovered (just CRISPR recognition tool or CRT and CRASS), and whether at least one viral linkage was made for the MAG.

See also Figure S4 and Data S2.

observed greater redundancy in linkages to the same virus, as only 39% of all linkages were to unique viruses. Therefore, while MAGs from the established wells contained fewer spacers and linked to fewer viruses, they could be matched to a proportionally greater number of different viruses, suggesting that the retained spacers may help to protect the host against a wider suite of viruses with less redundancy.

One way host populations with fewer spacers may still provide efficient defense against viral predation is if the retained spacers target regions of the viral genome with fewer mutations. We used single nucleotide polymorphism (SNP) frequency as a proxy for sequence variation in viral genes to investigate possible spacer effectiveness. Overall, very few spacers persisted in both sets of wells, highlighting the continual fluctuations in host and viral communities and the loss and gain of spacers within hosts populations. More spacers from the established wells persisted (4.5%) for at least half the sampling time points compared with the new wells (1.3%). Further, we observed that spacers recovered at only one time point generally matched viral protospacers with the highest SNP frequency (Figure S5). Additionally, the percentage of targeted viral genes with zero SNPs increased with spacer persistence in the new wells but not the established wells (Figure S5). That is, spacers that were present across the most time points tended to target viral genes that had less sequence variation within the community. These trends may be associated with the increased selection and retention of spacers that may confer viral resistance for longer periods of time.

Temporal increase in patterns of host-virus co-existence

Spacers within CRISPR-Cas arrays can be uniquely leveraged to make strong inferences about host-virus dynamics, as spacers from the host array often identically match the viral protospacer "target."[15] Leveraging additional spacers identified via CRASS, we were able to identify 2,110 viral linkages across 90 MAGs representing 25 different phyla (Figure 4). Indeed, matching all spacers associated with a host MAG to our vMAG database yielded at least one viral linkage for a majority of MAGs encoding a CRISPR-Cas array. We observed $\geq$1 viral linkage for 68% of MAGs with CRISPR-Cas arrays in the established wells (48 of 70 MAGs) and 79% of MAGs with CRISPR-Cas arrays in the new wells (42 of 53 MAGs) (Figure 4; Data S2). There was no significant relationship between MAG coverage and the number of viral linkages in the established wells, and a weak positive relationship in the new wells (Spearman's rho: R = 0.39; p = 1.4e−07), suggesting that there is not a sustained relationship

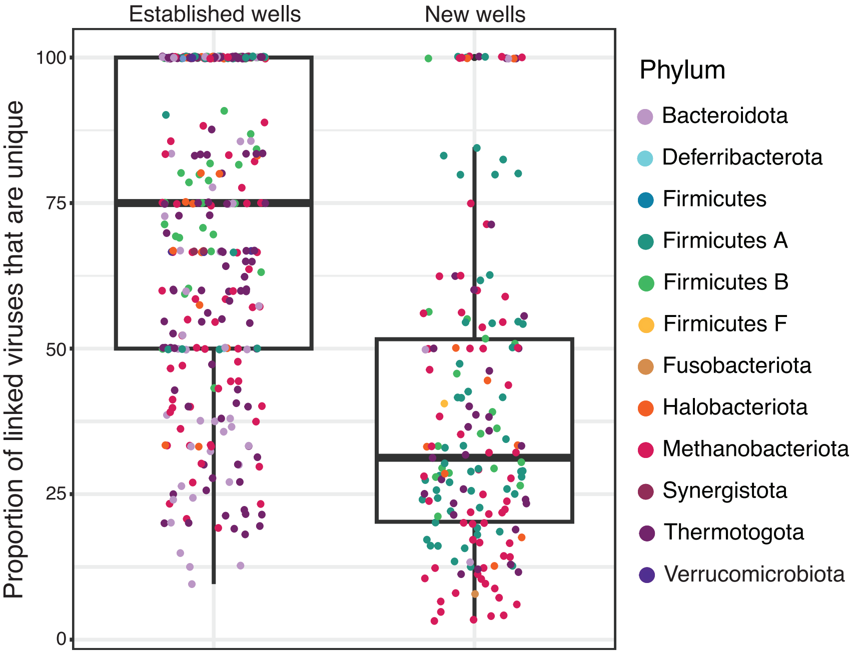

Figure 5. Redundancy of viral linkages

Boxplots illustrating the proportion of linked viruses per MAG that are unique. Each point represents a single host population at a single time point where spacers were recovered and linkages were made, colored by phylum. The proportion of unique viruses linked was calculated as the number of different viruses linked out of the total number of linkages at a given time point. The boxes represent the 25th to 75th percentiles, and the middle line represents the median.

See also Figures S4 and S5.

between the hosts' overall success and the number of different interacting viruses.

Even among widespread CRISPR-Cas defense and the presence of matching (linking) spacers, many linked viruses persisted. Therefore, we next evaluated how often viruses can persist and interact with the host population, despite theoretical CRISPR-Cas defense. We leveraged our 2,110 host-virus linkages to quantify differences in host-virus co-occurrence patterns in both sets of wells and studied how these dynamics may develop through time. We quantified occurrences of three scenarios for every host-virus linkage: (1) when only the host was present, (2) when both the virus and host were present, and (3) when only the virus was present (Figure 6). Scenarios where both host and virus were absent were excluded. In this analysis, "absence" of host or virus is likely not complete absence of the virus in the ecosystem but rather indicates that their true abundance was very low and there was insignificant evidence for their presence.

We observed differing patterns of host-virus dynamics between the new and established wells, potentially reflecting the establishment of microbial communities through time in this closed ecosystem. Notably, we tracked a decreasing trend in occurrences of only the host present in the new wells, approaching values observed in the established wells. Concurrently, we observed a slight increasing trend in host-virus co-existence in both sets of wells. This provides evidence that CRISPR-Cas may be most effective when microbial communities are first introduced into the newly formed ecosystem. CRISPR-Cas may then become less effective through time as the selection of viruses able to evade host defenses results in greater frequency of host-virus co-existence (Figure 6).

Anti-CRISPR (*acr*) genes are one mechanism employed by viruses to persist despite host defense, as they can suppress CRISPR-Cas systems. Putative anti-CRISPR genes were identified in 16 different vMAGs that were persistent in the established wells. Although this is a small proportion of total vMAGs, *acr* genes are poorly characterized and infrequently observed in natural ecosystems and are thus likely under-detected in this dataset. Together, these findings shed light on the complexities of host-virus dynamics temporally and how subsurface closed ecosystems may develop toward an equilibrium of host-virus co-existence, as opposed to dominance by host or viral populations,[67] or "red queen" dynamics of constant evolution and population turnover.[68,69]

Here, we leveraged time-resolved samples from six hydraulically fractured shale wells to establish CRISPR-based host-viral linkages and study long-term host-virus co-existence and CRISPR-Cas dynamics in a natural, closed ecosystem. Timeseries data (>800 days) from all six wells allowed us to recover CRISPR spacers from metagenomes and MAGs, which facilitated community-level and host-population level analyses of CRISPR-Cas defense through time. At the community level, we observed evidence that viral predation is active through time and that hosts are likely incorporating new spacers into their arrays in response to viral interactions. Next, at the genome level, we observed that CRISPR-Cas viral defense systems were widely encoded across a majority of MAGs. In total, we identified CRISPR arrays in ~60% of MAGs across 25 of 29 different phyla representing host populations from the deep shale microbial communities. We observed that host populations (represented by MAGs) in the established wells were associated with fewer spacers and that there was less redundancy in viral linkages, potentially reflecting selection for retention of more effective spacers through time in a closed ecosystem.

Leveraging CRISPR spacers to link viruses to hosts, we next identified potential viruses for a majority of hosts containing a CRISPR-Cas array, with over 2,000 total linkages identified across 90 different host MAGs. The proliferation of microenvironments (e.g., biofilms) over time in these subsurface ecosystems may constrain the number of interactions between diverse host and viral populations, resulting in fewer linkages in the established wells. Alternatively, such patterns may be attributed to lack of viral recovery due to successful host defense. Finally, given the prevalence of CRISPR-Cas systems and the important role this defense might have in host-viral co-existence, we used host-viral linkages to interrogate host-virus temporal dynamics. We found that co-existence of host and viral populations

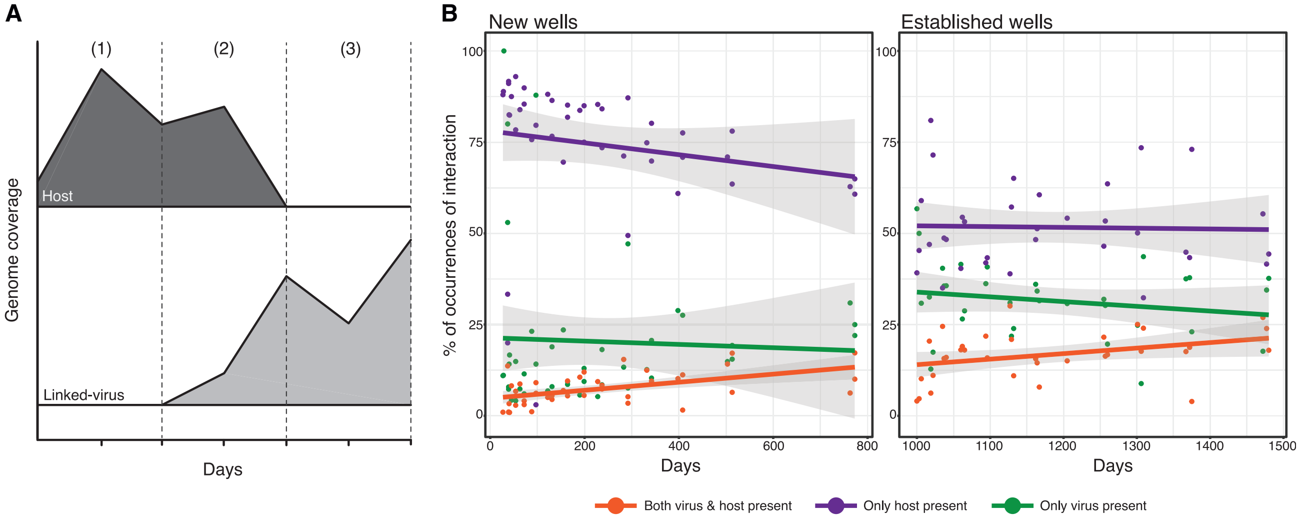

Figure 6. Patterns of host-virus co-existence

(A) Conceptual diagram of different "interaction" types: (1) where only host is present, but the virus is not present (below detection), (2) where host and virus are both present, and (3) where only the virus is present. Axes are purposely left blank, given this conceptual illustration.

(B) Temporal trends in percent of each interaction type in the new and established wells. Lines represent linear trends, while shaded gray areas indicate the 95% confidence interval.

See also Data S3.

generally increased through time, potentially due to the selection for viruses able to evade host defenses, specifically CRISPR-Cas defense, in this closed ecosystem. Together, this study offers new insights into the long-term dynamics between host and viral populations and CRISPR-based host-viral linkages within a subsurface ecosystem.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - ○ DNA extraction and metagenomic sequencing
  - ○ 16S rRNA gene sequencing and analysis
  - ○ Metagenomic assembly, binning, and viral recovery
  - ○ Calculating MAG and vMAG coverage and relative abundance
  - ○ Detection of viral defense systems and recovery of spacers
  - ○ Making CRISPR-based host-virus linkages
  - ○ Host-viral co-occurrence patterns
  - ○ Analysis of single nucleotide polymorphisms in vMAGs
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2023.06.033.

ACKNOWLEDGMENTS

This work was supported through an award (EAR-1847684) from the National Science Foundation Geobiology and Low Temperature Geochemistry program to M.J.W. The work conducted by the US Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science user facility, is supported by the Office of Science of the US Department of Energy, operated under contract no. DE-AC02-05CH11231 (S.R.). This work was also supported by the US Geological Survey's Energy Resources Program (Alicia Lindauer, program coordinator) (J.L.S.). Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

AUTHOR CONTRIBUTIONS

M.J.W. and J.L.S. designed field sampling; K.K.A., M.J.W., and J.L.S. collected samples; K.K.A. performed data analysis, with guidance from M.J.W. and S.R.; K.K.A. and M.J.W. wrote the manuscript, with feedback and support from S.R. and J.L.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 8, 2023
Revised: May 18, 2023
Accepted: June 12, 2023
Published: July 3, 2023

REFERENCES

1. Makarova, K.S., Wolf, Y.I., Iranzo, J., Shmakov, S.A., Alkhnbashi, O.S., Brouns, S.J.J., Charpentier, E., Cheng, D., Haft, D.H., Horvath, P., et al. (2020). Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. Nat. Rev. Microbiol. *18*, 67–83. https://doi.org/10.1038/s41579-019-0299-x.
2. Makarova, K.S., Wolf, Y.I., and Koonin, E.V. (2013). Comparative genomics of defense systems in archaea and bacteria. Nucleic Acids Res. *41*, 4360–4377. https://doi.org/10.1093/nar/gkt157.

3. Staals, R.H.J., and Brouns, S.J.J. (2013). Distribution and mechanism of the Type I CRISPR-Cas systems. In CRISPR-Cas Systems: RNA-Mediated Adaptive Immunity in Bacteria and Archaea, R. Barrangou, and J. van der Oost, eds. (Springer), pp. 145–169. https://doi.org/10.1007/978-3-642-34657-6_6.

4. Burstein, D., Sun, C.L., Brown, C.T., Sharon, I., Anantharaman, K., Probst, A.J., Thomas, B.C., and Banfield, J.F. (2016). Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nat. Commun. *7*, 10613. https://doi.org/10.1038/ncomms10613.

5. Hampton, H.G., Watson, B.N.J., and Fineran, P.C. (2020). The arms race between bacteria and their phage foes. Nature *577*, 327–336. https://doi.org/10.1038/s41586-019-1894-8.

6. Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science *315*, 1709–1712. https://doi.org/10.1126/science.1138140.

7. Jackson, S.A., McKenzie, R.E., Fagerlund, R.D., Kieper, S.N., Fineran, P.C., and Brouns, S.J.J. (2017). CRISPR-Cas: adapting to change. Science *356*, eaal5056. https://doi.org/10.1126/science.aal5056.

8. Hille, F., Richter, H., Wong, S.P., Bratovič, M., Ressel, S., and Charpentier, E. (2018). The biology of CRISPR-Cas: backward and forward. Cell *172*, 1239–1259. https://doi.org/10.1016/j.cell.2017.11.032.

9. Koonin, E.V., and Makarova, K.S. (2019). Origins and evolution of CRISPR-Cas systems. Philos. Trans. R. Soc. Lond. B Biol. Sci. *374*, 20180087. https://doi.org/10.1098/rstb.2018.0087.

10. Barrangou, R., and Marraffini, L.A. (2014). CRISPR-Cas systems: prokaryotes upgrade to adaptive immunity. Mol. Cell *54*, 234–244. https://doi.org/10.1016/j.molcel.2014.03.011.

11. Andersson, A.F., and Banfield, J.F. (2008). Virus population dynamics and acquired virus resistance in natural microbial communities. Science *320*, 1047–1050. https://doi.org/10.1126/science.1157358.

12. Horvath, P., and Barrangou, R. (2010). CRISPR/Cas, the immune system of bacteria and Archaea. Science *327*, 167–170. https://doi.org/10.1126/science.1179555.

13. Horvath, P., Romero, D.A., Coûté-Monvoisin, A.C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C., and Barrangou, R. (2008). Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. J. Bacteriol. *190*, 1401–1412. https://doi.org/10.1128/JB.01415-07.

14. Watson, B.N.J., Steens, J.A., Staals, R.H.J., Westra, E.R., and Houte, S. van (2021). Coevolution between bacterial CRISPR-Cas systems and their bacteriophages. Cell Host Microbe *29*, 715–725. https://doi.org/10.1016/j.chom.2021.03.018.

15. Edwards, R.A., McNair, K., Faust, K., Raes, J., and Dutilh, B.E. (2016). Computational approaches to predict bacteriophage–host relationships. FEMS Microbiol. Rev. *40*, 258–272. https://doi.org/10.1093/femsre/fuv048.

16. Anderson, R.E., Brazelton, W.J., and Baross, J.A. (2011). Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. FEMS Microbiol. Ecol. *77*, 120–133. https://doi.org/10.1111/j.1574-6941.2011.01090.x.

17. Sanguino, L., Franqueville, L., Vogel, T.M., and Larose, C. (2015). Linking environmental prokaryotic viruses and their host through CRISPRs. FEMS Microbiol. Ecol. *91*, fiv046. https://doi.org/10.1093/femsec/fiv046.

18. McKay, L.J., Nigro, O.D., Dlakić, M., Luttrell, K.M., Rusch, D.B., Fields, M.W., and Inskeep, W.P. (2022). Sulfur cycling and host-virus interactions in Aquificales-dominated biofilms from Yellowstone's hottest ecosystems. ISME J. *16*, 842–855. https://doi.org/10.1038/s41396-021-01132-4.

19. Emerson, J.B., Andrade, K., Thomas, B.C., Norman, A., Allen, E.E., Heidelberg, K.B., and Banfield, J.F. (2013). Virus-host and CRISPR dynamics in archaea-dominated hypersaline lake Tyrrell, Victoria, Australia. Archaea *2013*, 370871. https://doi.org/10.1155/2013/370871.

20. Emerson, J.B., Roux, S., Brum, J.R., Bolduc, B., Woodcroft, B.J., Jang, H.B., Singleton, C.M., Solden, L.M., Naas, A.E., Boyd, J.A., et al. (2018). Host-linked soil viral ecology along a permafrost thaw gradient. Nat. Microbiol. *3*, 870–880. https://doi.org/10.1038/s41564-018-0190-y.

21. Amundson, K.K., Borton, M.A., Daly, R.A., Hoyt, D.W., Wong, A., Eder, E., Moore, J., Wunch, K., Wrighton, K.C., and Wilkins, M.J. (2022). Microbial colonization and persistence in deep fractured shales is guided by metabolic exchanges and viral predation. Microbiome *10*, 5. https://doi.org/10.1186/s40168-021-01194-8.

22. Berg, M., Goudeau, D., Olmsted, C., McMahon, K.D., Yitbarek, S., Thweatt, J.L., Bryant, D.A., Eloe-Fadrosh, E.A., Malmstrom, R.R., and Roux, S. (2021). Host population diversity as a driver of viral infection cycle in wild populations of green sulfur bacteria with long standing virus-host interactions. ISME J. *15*, 1569–1584. https://doi.org/10.1038/s41396-020-00870-1.

23. Daly, R.A., Roux, S., Borton, M.A., Morgan, D.M., Johnston, M.D., Booker, A.E., Hoyt, D.W., Meulia, T., Wolfe, R.A., Hanson, A.J., et al. (2019). Viruses control dominant bacteria colonizing the terrestrial deep biosphere after hydraulic fracturing. Nat. Microbiol. *4*, 352–361. https://doi.org/10.1038/s41564-018-0312-6.

24. Stern, A., Mick, E., Tirosh, I., Sagy, O., and Sorek, R. (2012). CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. Genome Res. *22*, 1985–1994. https://doi.org/10.1101/gr.138297.112.

25. Roux, S., Brum, J.R., Dutilh, B.E., Sunagawa, S., Duhaime, M.B., Loy, A., Poulos, B.T., Solonenko, N., Lara, E., Poulain, J., et al. (2016). Ecogenomics and potential biogeochemical impacts of globally abundant ocean viruses. Nature *537*, 689–693. https://doi.org/10.1038/nature19366.

26. Minot, S., Bryson, A., Chehoud, C., Wu, G.D., Lewis, J.D., and Bushman, F.D. (2013). Rapid evolution of the human gut virome. Proc. Natl. Acad. Sci. USA *110*, 12450–12455. https://doi.org/10.1073/pnas.1300833110.

27. Paez-Espino, D., Eloe-Fadrosh, E.A., Pavlopoulos, G.A., Thomas, A.D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N.N., and Kyrpides, N.C. (2016). Uncovering Earth's virome. Nature *536*, 425–430. https://doi.org/10.1038/nature19094.

28. Weinberger, A.D., Wolf, Y.I., Lobkovsky, A.E., Gilmore, M.S., and Koonin, E.V. (2012). Viral diversity threshold for adaptive immunity in prokaryotes. mBio *3*. e00456-e00412. https://doi.org/10.1128/mBio.00456-12.

29. Meaden, S., Biswas, A., Arkhipova, K., Morales, S.E., Dutilh, B.E., Westra, E.R., and Fineran, P.C. (2022). High viral abundance and low diversity are associated with increased CRISPR-Cas prevalence across microbial ecosystems. Curr. Biol. *32*, 220–227.e5. https://doi.org/10.1016/j.cub.2021.10.038.

30. Broniewski, J.M., Meaden, S., Paterson, S., Buckling, A., and Westra, E.R. (2020). The effect of phage genetic diversity on bacterial resistance evolution. ISME J. *14*, 828–836. https://doi.org/10.1038/s41396-019-0577-7.

31. Bernheim, A., Calvo-Villamañán, A., Basier, C., Cui, L., Rocha, E.P.C., Touchon, M., and Bikard, D. (2017). Inhibition of NHEJ repair by type II-A CRISPR-Cas systems in bacteria. Nat. Commun. *8*, 2094. https://doi.org/10.1038/s41467-017-02350-1.

32. Westra, E.R., Dowling, A.J., Broniewski, J.M., and van Houte, S. (2016). Evolution and ecology of CRISPR. Annu. Rev. Ecol. Evol. Syst. *47*, 307–331. https://doi.org/10.1146/annurev-ecolsys-121415-032428.

33. Weissman, J.L., Laljani, R.M.R., Fagan, W.F., and Johnson, P.L.F. (2019). Visualization and prediction of CRISPR incidence in microbial trait-space to identify drivers of antiviral immune strategy. ISME J. *13*, 2589–2602. https://doi.org/10.1038/s41396-019-0411-2.

34. Deveau, H., Barrangou, R., Garneau, J.E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P., and Moineau, S. (2008). Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J. Bacteriol. *190*, 1390–1400. https://doi.org/10.1128/JB.01412-07.

35. Bradde, S., Vucelja, M., Teşileanu, T., and Balasubramanian, V. (2017). Dynamics of adaptive immunity against phage in bacterial populations. PLoS Comput. Biol. *13*, e1005486. https://doi.org/10.1371/journal.pcbi.1005486.

36. Garrett, S.C. (2021). Pruning and tending immune memories: spacer dynamics in the CRISPR array. Front. Microbiol. *12*, 664299.

37. Lopez-Sanchez, M.J., Sauvage, E., Da Cunha, V., Clermont, D., Ratsima Hariniaina, E., Gonzalez-Zorn, B., Poyart, C., Rosinski-Chupin, I., and Glaser, P. (2012). The highly dynamic CRISPR1 system of Streptococcus agalactiae controls the diversity of its mobilome. Mol. Microbiol. *85*, 1057–1071. https://doi.org/10.1111/j.1365-2958.2012.08172.x.

38. Guerrero, L.D., Pérez, M.V., Orellana, E., Piuri, M., Quiroga, C., and Erijman, L. (2021). Long-run bacteria-phage coexistence dynamics under natural habitat conditions in an environmental biotechnology system. ISME J. *15*, 636–648. https://doi.org/10.1038/s41396-020-00802-z.

39. Sun, C.L., Thomas, B.C., Barrangou, R., and Banfield, J.F. (2016). Metagenomic reconstructions of bacterial CRISPR loci constrain population histories. ISME J. *10*, 858–870. https://doi.org/10.1038/ismej.2015.162.

40. Levin, B.R., Moineau, S., Bushman, M., and Barrangou, R. (2013). The population and evolutionary dynamics of phage and bacteria with CRISPR-mediated immunity. PLoS Genet. *9*, e1003312. https://doi.org/10.1371/journal.pgen.1003312.

41. Vale, P.F., Lafforgue, G., Gatchitch, F., Gardan, R., Moineau, S., and Gandon, S. (2015). Costs of CRISPR-Cas-mediated resistance in Streptococcus thermophilus. Proc. Biol. Sci. *282*, 20151270. https://doi.org/10.1098/rspb.2015.1270.

42. Martynov, A., Severinov, K., and Ispolatov, I. (2017). Optimal number of spacers in CRISPR arrays. PLoS Comput. Biol. *13*, e1005891. https://doi.org/10.1371/journal.pcbi.1005891.

43. McGinn, J., and Marraffini, L.A. (2016). CRISPR-Cas systems optimize their immune response by specifying the site of spacer integration. Mol. Cell *64*, 616–623. https://doi.org/10.1016/j.molcel.2016.08.038.

44. Bradde, S., Nourmohammad, A., Goyal, S., and Balasubramanian, V. (2020). The size of the immune repertoire of bacteria. Proc. Natl. Acad. Sci. USA *117*, 5144–5151. https://doi.org/10.1073/pnas.1903666117.

45. Childs, L.M., Held, N.L., Young, M.J., Whitaker, R.J., and Weitz, J.S. (2012). Multiscale model of Crispr-induced coevolutionary dynamics: diversification at the interface of Lamarck and Darwin. Evolution *66*, 2015–2029. https://doi.org/10.1111/j.1558-5646.2012.01595.x.

46. Daly, R.A., Borton, M.A., Wilkins, M.J., Hoyt, D.W., Kountz, D.J., Wolfe, R.A., Welch, S.A., Marcus, D.N., Trexler, R.V., MacRae, J.D., et al. (2016). Microbial metabolisms in a 2.5-km-deep ecosystem created by hydraulic fracturing in shales. Nat. Microbiol. *1*, 16146. https://doi.org/10.1038/nmicrobiol.2016.146.

47. Cluff, M.A., Hartsock, A., MacRae, J.D., Carter, K., and Mouser, P.J. (2014). Temporal changes in microbial ecology and geochemistry in produced water from hydraulically fractured Marcellus Shale Gas wells. Environ. Sci. Technol. *48*, 6508–6517. https://doi.org/10.1021/es501173p.

48. Mouser, P.J., Borton, M., Darrah, T.H., Hartsock, A., and Wrighton, K.C. (2016). Hydraulic fracturing offers view of microbial life in the deep terrestrial subsurface. FEMS Microbiol. Ecol. *92*, fiw166. https://doi.org/10.1093/femsec/fiw166.

49. Booker, A.E., Hoyt, D.W., Meulia, T., Eder, E., Nicora, C.D., Purvine, S.O., Daly, R.A., Moore, J.D., Wunch, K., Pfiffner, S.M., et al. (2019). Deep-subsurface pressure stimulates metabolic plasticity in shale-colonizing Halanaerobium spp. Appl. Environ. Microbiol. *85*, e00018-19. https://doi.org/10.1128/AEM.00018-19.

50. Wang, H., Lu, L., Chen, X., Bian, Y., and Ren, Z.J. (2019). Geochemical and microbial characterizations of flowback and produced water in three shale oil and gas plays in the central and western United States. Water Res. *164*, 114942. https://doi.org/10.1016/j.watres.2019.114942.

51. Hull, N.M., Rosenblum, J.S., Robertson, C.E., Harris, J.K., and Linden, K.G. (2018). Succession of toxicity and microbiota in hydraulic fracturing flowback and produced water in the Denver–Julesburg Basin. Sci. Total Environ. *644*, 183–192. https://doi.org/10.1016/j.scitotenv.2018.06.067.

52. Murali Mohan, A., Hartsock, A., Bibby, K.J., Hammack, R.W., Vidic, R.D., and Gregory, K.B. (2013). Microbial community changes in hydraulic fracturing fluids and produced water from shale gas extraction. Environ. Sci. Technol. *47*, 13141–13150. https://doi.org/10.1021/es402928b.

53. Murali Mohan, A., Hartsock, A., Hammack, R.W., Vidic, R.D., and Gregory, K.B. (2013). Microbial communities in flowback water impoundments from hydraulic fracturing for recovery of shale gas. FEMS Microbiol. Ecol. *86*, 567–580. https://doi.org/10.1111/1574-6941.12183.

54. Struchtemeyer, C.G., and Elshahed, M.S. (2012). Bacterial communities associated with hydraulic fracturing fluids in thermogenic natural gas wells in North Central Texas, USA. FEMS Microbiol. Ecol. *81*, 13–25. https://doi.org/10.1111/j.1574-6941.2011.01196.x.

55. Hockenberry, A.J., and Wilke, C.O. (2021). BACPHLIP: predicting bacteriophage lifestyle from conserved protein domains. PeerJ *9*, e11396. https://doi.org/10.7717/peerj.11396.

56. Whitman, W.B., Coleman, D.C., and Wiebe, W.J. (1998). Prokaryotes: the unseen majority. Proc. Natl. Acad. Sci. USA *95*, 6578–6583.

57. McMahon, S., and Parnell, J. (2014). Weighing the deep continental biosphere. FEMS Microbiol. Ecol. *87*, 113–120. https://doi.org/10.1111/1574-6941.12196.

58. Flemming, H.C., and Wuertz, S. (2019). Bacteria and archaea on Earth and their abundance in biofilms. Nat. Rev. Microbiol. *17*, 247–260. https://doi.org/10.1038/s41579-019-0158-9.

59. Tinker, K., Gardiner, J., Lipus, D., Sarkar, P., Stuckman, M., and Gulliver, D. (2020). Geochemistry and microbiology predict environmental niches with conditions favoring potential microbial activity in the Bakken shale. Front. Microbiol. *11*, 1781.

60. Stemple, B., Tinker, K., Sarkar, P., Miller, J., Gulliver, D., and Bibby, K. (2021). Biogeochemistry of the Antrim shale natural gas reservoir. ACS Earth Space Chem. *5*, 1752–1761. https://doi.org/10.1021/acsearthspacechem.1c00087.

61. John, S.G., Mendez, C.B., Deng, L., Poulos, B., Kauffman, A.K.M., Kern, S., Brum, J., Polz, M.F., Boyle, E.A., and Sullivan, M.B. (2011). A simple and efficient method for concentration of ocean viruses by chemical flocculation. Environ. Microbiol. Rep. *3*, 195–202. https://doi.org/10.1111/j.1758-2229.2010.00208.x.

62. Skennerton, C.T., Imelfort, M., and Tyson, G.W. (2013). Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. Nucleic Acids Res. *41*, e105. https://doi.org/10.1093/nar/gkt183.

63. Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G., and Sorek, R. (2018). Systematic discovery of antiphage defense systems in the microbial pangenome. Science *359*, eaar4120. https://doi.org/10.1126/science.aar4120.

64. Oliveira, P.H., Touchon, M., and Rocha, E.P.C. (2014). The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. Nucleic Acids Res. *42*, 10618–10631. https://doi.org/10.1093/nar/gku734.

65. Chopin, M.C., Chopin, A., and Bidnenko, E. (2005). Phage abortive infection in lactococci: variations on a theme. Curr. Opin. Microbiol. *8*, 473–479. https://doi.org/10.1016/j.mib.2005.06.006.

66. Dy, R.L., Przybilski, R., Semeijn, K., Salmond, G.P.C., and Fineran, P.C. (2014). A widespread bacteriophage abortive infection system functions through a type IV toxin–antitoxin mechanism. Nucleic Acids Res. *42*, 4590–4605. https://doi.org/10.1093/nar/gkt1419.

67. van Houte, S., Ekroth, A.K.E., Broniewski, J.M., Chabas, H., Ashby, B., Bondy-Denomy, J., Gandon, S., Boots, M., Paterson, S., Buckling, A., et al. (2016). The diversity-generating benefits of a prokaryotic adaptive immune system. Nature *532*, 385–388. https://doi.org/10.1038/nature17436.

68. Stern, A., and Sorek, R. (2011). The phage-host arms race: shaping the evolution of microbes. BioEssays *33*, 43–51. https://doi.org/10.1002/bies.201000071.

69. Weitz, J.S., Poisot, T., Meyer, J.R., Flores, C.O., Valverde, S., Sullivan, M.B., and Hochberg, M.E. (2013). Phage–bacteria infection networks. Trends Microbiol. *21*, 82–91. https://doi.org/10.1016/j.tim.2012.11.003.

70. Peng, Y., Leung, H.C.M., Yiu, S.M., and Chin, F.Y.L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics *28*, 1420–1428. https://doi.org/10.1093/bioinformatics/bts174.

71. Kang, D.D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., and Wang, Z. (2019). MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ *7*, e7359. https://doi.org/10.7717/peerj.7359.

72. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055. https://doi.org/10.1101/gr.186072.114.

73. Olm, M.R., Brown, C.T., Brooks, B., and Banfield, J.F. (2017). dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. *11*, 2864–2868. https://doi.org/10.1038/ismej.2017.126.

74. Chaumeil, P.A., Mussig, A.J., Hugenholtz, P., and Parks, D.H. (2019). GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics *36*, 1925–1927. https://doi.org/10.1093/bioinformatics/btz848.

75. Shaffer, M., Borton, M.A., McGivern, B.B., Zayed, A.A., La Rosa, S.L., Solden, L.M., Liu, P., Narrowe, A.B., Rodríguez-Ramos, J., Bolduc, B., et al. (2020). DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic Acids Res. *48*, 8883–8900. https://doi.org/10.1093/nar/gkaa621.

76. Guo, J., Bolduc, B., Zayed, A.A., Varsani, A., Dominguez-Huerta, G., Delmont, T.O., Pratama, A.A., Gazitúa, M.C., Vik, D., Sullivan, M.B., et al. (2021). VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome *9*, 37. https://doi.org/10.1186/s40168-020-00990-y.

77. Yi, H., Huang, L., Yang, B., Gomez, J., Zhang, H., and Yin, Y. (2020). AcrFinder: genome mining anti-CRISPR operons in prokaryotes and their viruses. Nucleic Acids Res. *48*, W358–W365. https://doi.org/10.1093/nar/gkaa351.

78. Brushnell, B. (2014). BBTools software package. SourceForge.net. BBMap download. https://sourceforge.net/projects/bbmap/.

79. Bland, C., Ramsey, T.L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N.C., and Hugenholtz, P. (2007). CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics *8*, 209. https://doi.org/10.1186/1471-2105-8-209.

80. Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S.A., and Sørensen, S.J. (2020). CRISPRCasTyper: automated identification, annotation, and classification of CRISPR-Cas loci. CRISPR J. *3*, 462–469. https://doi.org/10.1089/crispr.2020.0059.

81. Gregory, A.C., Gerhardt, K., Zhong, Z.P., Bolduc, B., Temperton, B., Konstantinidis, K.T., and Sullivan, M.B. (2022). MetaPop: a pipeline for macro- and microdiversity analyses and visualization of microbial and viral metagenome-derived populations. Microbiome *10*, 49. https://doi.org/10.1186/s40168-022-01231-0.

82. Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., and Holmes, S.P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. Nat. Methods *13*, 581–583. https://doi.org/10.1038/nmeth.3869.

83. Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A., et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nat. Biotechnol. *35*, 725–731. https://doi.org/10.1038/nbt.3893.

84. protocols.io (2021). Viral sequence identification SOP with VirSorter2. https://www.protocols.io/view/viral-sequence-identification-sop-with-virsorter2-5qpvoyqebg4o/v3.

85. Nayfach, S., Camargo, A.P., Schulz, F., Eloe-Fadrosh, E., Roux, S., and Kyrpides, N.C. (2021). CheckV assesses the quality and completeness of metagenome-assembled viral genomes. Nat. Biotechnol. *39*, 578–585. https://doi.org/10.1038/s41587-020-00774-7.

86. Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J.H., Roux, S., Adriaenssens, E.M., Brister, J.R., Kropinski, A.M., Krupovic, M., Lavigne, R., et al. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. Nat. Biotechnol. *37*, 632–639. https://doi.org/10.1038/s41587-019-0100-8.

87. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical commercial assays		
0.2µm PES membrane Nalgene vacuum filtration unit (Thermo Scientific)	ThermoFisher	Cat. No. 564-0020
DNeasy PowerSoil kit	Qiagen	Cat. No. 12888-100
Deposited data		
Newly Sequenced Metagenomic Reads and 16S RRNA Amplicon Sequencing	This study	BioProject PRJNA308326
Software and algorithms		
IDBA-UD (v.1.1.3)	Peng et al.[70]	https://github.com/loneknightpy/idba
MetaBAT2 (v2.12.1)	Kang et al.[71]	https://bitbucket.org/berkeleylab/metabat/src/master/
CheckM (v.1.1.2)	Parks et al.[72]	https://github.com/Ecogenomics/CheckM
dRep (v2.2.3)	Olm et al.[73]	https://github.com/MrOlm/drep
CRisprASSembler: CRASS (v1.0.1)	Skennerton et al.[62]	https://github.com/ctSkennerton/crass
GTDB-Tk v2.2.0	Chaumeil et al.[74]	https://github.com/Ecogenomics/GTDBTk
DRAM (v1.2.4)	Shaffer et al.[75]	https://github.com/WrightonLabCSU/DRAM
VirSorter2 (v2.2.2)	Guo et al.[76]	https://github.com/jiarong/VirSorter2
ArcFinder	Yi et al.[77]	https://github.com/HaidYi/acrfinder
BACPHLIP	Hockenberry and Wilke[55]	https://github.com/adamhockenberry/bacphlip
bbmap (v38.89)	SourceForge.net[78]	https://sourceforge.net/projects/bbmap/
coverM (v0.6.0)	NA	https://github.com/wwood/CoverM
CRISPR Recognition Tool (CRT) (v.1.2)	Bland et al.[79]	https://www.geneious.com/plugins/crt/
CRISPRCasTyper (v.1.8.0)	Russel et al.[80]	https://github.com/Russel88/CRISPRCasTyper
DefenseFinder (v.1.0)	Doron et al.[63]	https://github.com/mdmparis/defense-finder
MetaPop	Gregory et al.[81]	https://github.com/metaGmetapop/metapop

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Kaela Amundson (kaela.amundson@colostate.edu)

Materials availability

This study did not generate any unique reagents.

Data and code availability

- Metagenomic reads, 16S rRNA reads, MAGs, and vMAGs have been deposited and are available at NCBI: BioProject PRJNA308326. Specific accession numbers for metagenomic reads, 16S reads, MAGs, and vMAGs are listed in the respective supplementary datasets.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Produced fluid samples were collected from six hydraulically fractured wells from the Niobrara formation, within the Denver-Julesburg (DJ) Basin, in eastern Colorado between October 2018 and October 2020 (*n*=78) (Figure 1). The Niobrara shale formation consists of three benches that are located approximately 1890-1950 meters deep in the subsurface with a downhole temperature measuring approximately 112°C (recorded while drilling). The six wells within this formation were sampled are split equally into two groups defined by their age when sample collection began: the three 'established' wells (*n=33*) had been producing for approximately 1000 days prior to sample collection (DJB-1, DJB-2, and DJB-3), while the three 'new' wells (*n=45*) were sampled from day ~30 in production (DJB-4, DJB-5, and DJB-6). A small number of early produced fluid samples (those beginning with 'JMDJ#', <60 days) were collected directly from well heads and filtered through a 0.22μm pore size polyethersulfone membrane Sterivex filter due to field sampling constraints (MilliporeSigma) with a minimum of 500mL of fluid filtered. Most produced fluids (those beginning with 'DJKA') were collected directly from separator tanks into 1L Nalgene bottles with no head space and stored at 4°C until processing, which occurred within 24 hours from when the sample was collected. To the degree possible, samples were collected from separator tanks shortly after the last contents had been released to the central processing facility. 500-800mL of fluid was filtered through a 0.2μm PES membrane Nalgene vacuum filtration unit (Thermo Scientific). Filters were removed from the units and stored at -20°C until DNA extraction. Therefore, MAGs were recovered from produced fluids collected from the separator tank or well head for each well, though for brevity we refer to these MAGs as simply recovered from the well. Conductivity was measured on raw, unfiltered fluids at room temperature using a Myron L 6PIIFCE meter.

METHOD DETAILS

DNA extraction and metagenomic sequencing

Total nucleic acids were extracted from half of each sample's 0.2μm filter using DNAeasy PowerSoil Kit (Qiagen). Extraction blanks were run with each round of DNA extractions and all returned no detectable nucleic acids using the maximum amount of blank sample (20μL) via the Qubit dsDNA High Sensitivity assay kit (ThermoFisher Scientific). For all 78 samples, genomic DNA was prepared for metagenomic sequencing at the Genomics and Microarray Core at the University of Colorado, Denver's Genomics Shared Resource. Samples were prepared using the Illumina Nextera XT Library System according to manufacturer's instructions for 2x150bp libraries and were sequenced using the Illumina NovaSeq platform and paired-end reads were collected.

16S rRNA gene sequencing and analysis

Nucleic acids for all samples were also sent to Argonne National Laboratory for 16S rRNA gene sequencing (Data S1). Sequencing was performed with the Illumina MiSeq platform, using the Earth Microbiome Project primer set for amplification of the 251bp hypervariable V4 region. 16S rRNA gene sequences were obtained via Argonne's standard procedure, with the exception of performing 30 PCR amplification cycles. Paired-end reads were processed with QIIME2 (v 2021.2) EMP protocol, by first demultiplexing via exact-matching of barcodes, trimmed to 250bp and denoised with DADA2.[82] Representative sequences were taxonomically classified with SILVA (release 138). 16S community composition results are shown in Figure S1. All 16S rRNA gene sequencing reads were submitted to NCBI under BioProject PRJNA308326 and individual BioSample accession numbers are listed in Data S1.

Metagenomic assembly, binning, and viral recovery

For bacterial, archaeal, and viral recovery, total sequenced DNA from each sample was first trimmed from 5' to 3' ends with Sickle (https://github.com/najoshi/sickle) and individually assembled using IDBA-UD with default parameters.[70] Assembly information for each sample is provided in Data S1. Only scaffolds ≥5kb from metagenomic assemblies were used for binning bacterial and archaeal genomes with MetaBAT2 (v2.12.1) to recover metagenome assembled genomes (MAGs).[71] CheckM (v.1.1.2) lineage workflow ('lineage_wf') followed by the 'qa' command was used to assess completion and contamination for each metagenomic bin,[72] and medium (>50% completion, <10% contamination) and high (>90% completion, <5% contamination) quality bins were recovered from all samples from all six wells following the standard metrics for MAGs proposed by Bowers et al.[83] The two sets of unique MAGs (from the new and established wells) were individually determined by dRep v2.2.3 using default parameters.[73] MAGs were dereplicated based on their well groupings so that representative host populations were most reflective of true host populations in the subsurface communities, and to identify host repeats and associate spacers from CRASS. We anticipate differences such as the age differences (including possible differences in additive used), as well as physical separation of the well groupings from one another could impact host genomic content, specifically repeats in CRISPR arrays, and thus we created a MAG database unique to each grouping of wells. We refer to the final set of 202 MAGs as the 'host' community (Figure S1). All MAGs were taxonomically classified using GTDB-Tk v2.2.0.[74] Metagenomic assemblies and MAGs were annotated via DRAM (v1.2.4) using default parameters.[75] Additional details about MAGs can be found in Data S2. Metagenomic reads and MAGs were submitted to NCBI BioProject PRJNA308326 and individual accession numbers can be found in Data S1.

Viral MAGs (vMAGs) were also identified in metagenomic assemblies from scaffolds ≥10kb in length using VirSorter2 (v2.2.2)[76] and following the "Viral sequence identification SOP with VirSorter2" developed by the Sullivan Lab.[84] Following this protocol, quality of vMAGs were assessed using checkV (v0.8.1) and annotated using DRAM-v (v1.2.4).[75,85] Low confidence vMAGs were removed following the manual curation steps in the SOP. Viral genomic contigs (≥10kb) were clustered into viral populations (genus level) using the 'ClusterGenomes' (v 1.1.3) app in CyVerse using the parameters 95% average nucleotide identity and 90% alignment fraction of the smallest contig (https://github.com/simroux/ClusterGenomes). The resulting database of 2,176 vMAGs are considered our viral database. Viral taxonomy was determined by clustering vMAGs with viruses belonging to the viral reference taxonomy databases in NCBI Bacterial and Archaeal Viral RefSeq v211, and viruses from the

International Committee on Taxonomy of Viruses (ICTV) via vConTACT2 v0.11.3 with default settings.[86] Anti-CRISPR (arc) genes were identified in vMAGs using ArcFinder (using both homology-based and guilt by association based approaches) with default parameters.[77] Probable viral lifestyle (either lytic or lysogenic) was inferred via one of two methods: (1) presence of integrase genes via KEGG annotation and (2) >75% confidence of a temperate lifestyle assigned from BACPHLIP (HMM searching for temperate domains).[55] All vMAGs have been deposited under NCBI BioProject PRJNA308326 and additional details about vMAGs can be found in Data S3.

Calculating MAG and vMAG coverage and relative abundance

To calculate coverage and relative abundance of MAGs and vMAGs, all 78 pairs of trimmed metagenomics reads were rarified to the lowest metagenome sequencing depth of 9Gbp using the 'reformat' guide within bbmap.[78] Coverage for MAGs was calculated by competitively mapping rarified metagenomic reads to MAGs using bbmap (v38.89) with minid=90. Resulting sam files were converted to sorted bam files using samtools (v1.9).[87] Coverage for each MAG was calculated using coverM (v0.6.0) (https://github.com/wwood/CoverM) using two commands. First, coverM was run using –min-covered-fraction=90 to determine MAGs read recruitment to at least 90% of the genome. Second, coverage values were calculated using the -m reads_per_base command, which represents reads mapped/genome length, and thus multiplied this by read length (151bp) in order to calculate MAG coverage (simply, coverage = reads_per_base * 151 bp). Only MAGs with >1x coverage and with reads mapped to >90% of the genome were considered present in a sample. Relative abundance was thus calculated as the proportion of a given MAG's coverage out of the sum of all present MAGs' coverage, per sample.

Metagenomic reads were also mapped to vMAGs to determine coverage using bbmap with minid=95 (v38.89)[78] and sam files converted to bam files using samtools (v1.9).[87] Given vMAGs are viral contigs, coverM (https://github.com/wwood/CoverM) contig mode was applied with two commands. First, –min-covered-fraction 75 and next followed by -m reads_per_base to calculate coverage. Similar to requirements set for MAGs, here vMAGs must have a minimum covered fraction >75% to be considered present. Coverage values were calculated from the reads per base output*151 bp. Number of viruses present in a metagenome were determined by presence of vMAGs given this recruitment of metagenomic reads.

Detection of viral defense systems and recovery of spacers

CRISPR arrays in MAGs were identified using the Geneious (v.2020.0.5) plugin CRISPR Recognition Tool (CRT)[79] v.1.2 using the 'Find CRISPR loci' annotation tool with the following parameters: min number of repeats a CRISPR must contain: 4, minimum length of a CRISPR's repeated region: 19, maximum length of a CRISPR's repeated region: 55, minimum length of a CRISPR's non-repeated region (or spacer region): 19, maximum length of a CRISPR's non-repeated region (or spacer region), length of a search window used to discover CRISPR's: 8. CRISPR arrays were then classified into types/subtypes using CRISPRCasTyper (v.1.8.0)[80] via matching repeat sequences. Spacers were also detected in non-rarified and rarified trimmed metagenomics reads using CRisprASSembler: CRASS (v1.0.1).[62] Briefly, CRASS reassembles CRISPR-Cas arrays of repeats and spacers that tend to break during assemblies and groups spacers by the repeat sequences in CRISPR arrays. Only spacers recovered from rarified metagenomics reads were used to represent the 'total number of spacers in a metagenome' for all community-level correlations to not introduce bias from varying read depth. All recovered host genomes, regardless of detection of a CRISPR array, were also queried for 60 other known anti-phage systems using DefenseFinder (v.1.0).[63]

Making CRISPR-based host-virus linkages

Linkages between MAGs and vMAGs (hosts and viruses) were made exclusively via CRISPR spacers using two approaches (Figure 1). As a result of this, linkages could only be made with MAGs that had a detectable CRISPR array. First, CRISPR arrays were identified in MAGs using Geneious, and spacers and repeats were extracted from the CRISPR arrays. We then leveraged CRASS to make as many linkages as possible and evaluate the number of spacers associated with a MAG through time. Repeat sequences from MAGs were identically matched to direct repeat sequences from CRASS (same length, no mismatches). Spacers that were associated with a direct repeat sequence from CRASS were thus grouped with the MAG of the same repeat sequence. To make as many host-viral linkages as possible, spacers were extracted from CRASS applied to non-rarified reads. Next, spacers from all MAGs (linked via Geneious and CRASS) were queried against all vMAGs using BLASTn with the parameters to optimize short sequences BLAST: -dust no and -word_size 7. Finally, only identical or nearly identical (0 or 1 mismatch across spacer length) were used to match spacers to vMAGs and make host-viral linkages. Number of linkages per MAG is shown in Figure S4.

Host-viral co-occurrence patterns

All MAGs with at least one viral linkage were included in analyzing host-viral co-occurrence patterns. For each MAG and individual linked virus at every timepoint, all possibilities were evaluated for being one of three interaction types: (1) only host present but virus absent (below detection), (2) both host and virus are present and (3) when only the virus was present, but their linked host was absent (below detection). Instances where both host and virus were not present were excluded from any calculations and not counted in the total number of interaction occurrences, which was used to normalize occurrences. Thus, the percent of any interaction was calculated as the proportion of all interactions previously stated.

Analysis of single nucleotide polymorphisms in vMAGs

We combined SNP values for viral genes with the persistence of spacers that link host and virus to interrogate any possible relationship between gene variation and spacer retention for all MAGs with linkages (Figure S5). We utilized MetaPop[81] with default parameters to calculate the number of SNPs within all viral genes identified in our vMAGs. For genes that met MetaPop's default parameters, SNP frequency was calculated relative to the gene length. Genes containing linked protospacers that did not meet MetaPop's default parameters were not included in this analysis. Finally, SNP frequency for the gene containing the protospacer was combined with the persistence of the spacer (i.e., number of samples the spacer was recovered).

QUANTIFICATION AND STATISTICAL ANALYSIS

Alpha diversity (Shannon's index) and beta diversity (Bray-Curtis) values were calculated using vegan v2.6-2 in R. Alpha diversity was calculated using 16S rRNA amplicon data, while beta diversity and Bray-Curtis dissimilarity values were calculated based on the host and viral communities recovered via metagenomic sequencing and rarified reads. Metagenomics was used here since the viral community was recovered using metagenomics and thus the paired host communities were assessed similarly via MAGs (recovered from metagenomes). Bray-Curtis dissimilarity values were calculated as the difference in beta diversity from the previous timepoint. Spearman correlations and p-values were calculated using ggpubr to determine the strength and directionality of relationships between variables such as number of spacers, MAG/vMAG coverage, time, etc. Specifically, correlations between number of spacers and host coverage were only calculated for MAGs that were both present in at least 3 timepoints and also had spacer recovery from at least three timepoints.