

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Metagenomic Analysis of CRISPR-Mediated Host-Virus Interactions in Microbial Communities

Permalink

<https://escholarship.org/uc/item/5n12m3d3>

Author

Sun, Christine

Publication Date

2013

Supplemental Material

<https://escholarship.org/uc/item/5n12m3d3#supplemental>

Peer reviewed|Thesis/dissertation

Metagenomic Analysis of CRISPR-Mediated Host-Virus
Interactions in Microbial Communities

By

Christine Ling Sun

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Microbiology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Jillian Banfield, Chair

Professor Steven Brenner

Professor Jennifer Doudna

Professor Steven Lindow

Spring 2013

Abstract

Metagenomic Analysis of CRISPR-Mediated Host-Virus Interactions in Microbial Communities

by

Christine Ling Sun

Doctor of Philosophy in Microbiology

University of California, Berkeley

Professor Jillian F. Banfield, Chair

Viruses of Bacteria (bacteriophages) and Archaea have the ability to significantly alter the structure and function of microbial communities. Thus, it is critical to obtain a greater understanding of the dynamic interaction between microbial hosts and their associated viral populations. The CRISPR-Cas system in Bacteria and Archaea serves as a method to connect viruses to their hosts and provides insight into virus-host interactions. A clustered regularly interspaced short palindromic repeat (CRISPR) locus and CRISPR-associated (Cas) proteins function together in the CRISPR-Cas adaptive immune system. Transcripts of the spacers that separate the repeats in the CRISPR locus confer immunity through sequence identity with targeted viral, plasmid, or other foreign DNA. The CRISPR locus can be interpreted as a historical timeline of virus exposure as spacers are incorporated in a unidirectional manner at the leader end of the CRISPR locus.

Metagenomic approaches were employed to simultaneously analyze CRISPR loci in microbial hosts and sequences of their associated viruses to examine: 1) the viral response to CRISPR spacer diversification in a closed model system, 2) the retention of older spacers without co-existing targets through time, 3) the information about population histories revealed from the CRISPR locus, and 4) the factors impacting phage and viral diversity in a model natural system. The host-phage system of *Streptococcus thermophilus* DGCC7710 and phage 2972 was used as a closed model system whereas the low diversity microbial communities growing as biofilms atop acid mine drainage (AMD) within the Richmond Mine at Iron Mountain, CA, USA was used as a model natural system.

S. thermophilus DGCC7710 was challenged with phage 2972 and the resulting host and phage populations were examined after one week of co-culturing in order to explore how co-existing, co-evolving hosts and phage populations establish. Additions of new spacers converted the clonal CRISPR locus into a diversified locus, with multiple sub-dominant CRISPR strain lineages present in the final *S. thermophilus* population. All phage mutations that circumvented three early-acquired spacers were localized in the proto-spacer adjacent motif (PAM) or near the PAM end of the proto-spacer, suggesting a strong selective advantage for the phage that mutate in this region. The sequential fixation or near fixation of these single mutations indicates

selection events so severe that single phage genotypes ultimately gave rise to all surviving lineages.

The CRISPR loci of Bacteria and Archaea and their associated viral and plasmid populations from AMD biofilm microbial communities were examined from samples collected over eight years. Notably, CRISPR loci were present in most populations. It was shown that CRISPR loci in some AMD microorganisms retain older CRISPR spacers over long time periods, despite not targeting any abundant coexisting viruses or plasmids. In order to investigate this phenomenon, the two CRISPR loci from a dominant archaeal G-plasma population were reconstructed and viral targets were identified via spacer matches. A polyclonal bloom of viruses was detected, and G-plasma population with highly diverse CRISPR loci emerged. In collaboration, a mathematical model was developed to link documented patterns of genomic conservation in CRISPR loci to an evolutionary advantage against persistent viruses. The subset of hosts that retain old spacers seemingly lacking any matches to current viruses may indicate tuning of CRISPR-mediated immunity against low abundance viruses that may re-emerge.

Through retention of older spacers as well as the acquisition of new spacer sequences, CRISPR-Cas systems can provide insight into recent population history. The link between spacers and their viral targets and the relative position and order of spacers in the loci are key characteristics that can be uncovered via population metagenomic analysis. New bioinformatics methods were developed and used to analyze the CRISPR loci of *Leptospirillum* group II bacterial population and its associated bacteriophage AMDV1 in biofilms sampled over five years in the AMD site. Spacers throughout the locus target the same phage population (AMDV1), but there are blocks of consecutive spacers without AMDV1 targets and only newer spacers target plasmid populations. This suggests the consistent co-existence of the bacteria with the AMDV1 phage population, with periods when this phage was prominent, and a fluctuating plasmid pool. The approach of examining the pattern of CRISPR spacers with targets may have direct application to tracking the potential sources of medically- and defense-relevant microbial strains. Spacer matches can identify phage or plasmids previously existing in the host's environment and indicate phage and plasmid diversity levels over time, thus constraining the past history of the population.

CRISPR spacers can be used to identify sequences derived from viruses or plasmids within a metagenomic dataset. In fact, with metagenomic datasets sequenced from AMD biofilms, CRISPR spacer targeting enabled detection of a very wide variety of previously unknown viruses and plasmids. Fine scale examination of the AMDV1 phage population diversity over time revealed meter-scale spatial variation, but somewhat reproducible seasonal genotypic abundance patterns. There is evidence in CRISPR loci of *Leptospirillum* group II that excision events removed large blocks of spacers that only match phage or plasmid sequences present at earlier times. These CRISPR locus and phage diversity patterns suggest that there is viral-imposed selection for host strains as well as host-related selection for virus types.

The function of the CRISPR-Cas system was only recognized a little over five years ago. In the intervening period, the majority of research has focused on unraveling the biochemical and mechanistic underpinnings of Cas and spacer function. In this thesis, the emphasis has been on understanding the importance of population diversification for long-term CRISPR-Cas system-based immunity, the significance of spacer retention, and the utility of locus analysis for

ecological and evolutionary studies. Hosts and viruses were examined in two different systems. The challenge experiment between *S. thermophilus* and phage 2972 isolate cultures demonstrated the rapid evolution of host and phage populations as the direct result of the addition of CRISPR spacers and the fixation of phage mutations. Examination of CRISPR loci and associated viruses and plasmids populations in the AMD system suggests CRISPR loci reconstruction enables investigation of past viral and plasmid exposure. It was shown that older spacers may be retained in order to target reappearing or low abundance viruses. Spacers with co-existing matches provide host immunity and offer a method for detection and genomic analysis of new viruses. Evaluating results from both systems has provided a more complete understanding of the CRISPR-mediated host-virus dynamics in microbial communities.

Table of Contents

Introduction	ii
Acknowledgements	vii
Chapter 1 Phage mutations in response to CRISPR diversification in a bacterial population	1
Chapter 2 Persisting low-abundance viral sequences shape microbial CRISPR-based immunity	13
Chapter 3 Metagenomic reconstructions of bacterial CRISPR loci constrain population histories	35
Chapter 4 Stability and factors impacting phage and viral diversity over a multi-year period in a model microbial ecosystem	53
References	79

Introduction

Despite their relatively small size, microorganisms are crucial to life. In ecosystems, microbes can impact the flow of various global geochemical cycles and processes (Zehr and Ward, 2002; Arrigo, 2005; Giovannoni and Vergin, 2012; Stocker, 2012). For example, in some environments, communities of microorganisms have the ability both to contribute substantially to contamination levels (Edwards et al., 2000) or, alternatively, to remediate sites (Furukawa, 2003). In biotechnology, specifically chosen combinations of microbial cultures are used in bioleaching for metal recovery (Bosecker, 1997) or the production of dairy products, such as yogurt and cheese (Leroy and De Vuyst, 2004). In all of these instances, the microorganisms are not working as single entities but rather as populations or as communities.

Predation, either by single cellular eukaryotes or viruses, can directly impact the abundance of Bacteria and Archaea within microbial communities (Weinbauer and Höfle, 1998; Fuhrman, 1999; Suttle, 2005, 2007). Single cellular eukaryotes kill cells relatively indiscriminately, and often target different types of cells non-specifically (Sherr and Sherr, 2002). However, viruses of Bacteria (bacteriophages) and viruses of Archaea have specifically defined ranges of microbial hosts they are able to infect (Hyman and Abedon, 2010). It is well established that bacteriophages and archaeal viruses (collectively referred to here as viruses) can directly impact communities of microorganisms not only through predation but also through the transfer of specific genes (Fuhrman, 1999; Canchaya et al., 2003). Understanding the intricate relationship between viruses and their microbial hosts is important for many environmental and biotechnological processes and has implications for ecosystem functioning.

Historically, it has been difficult to study the dynamic interactions between viruses and their hosts. The vast majority of tests used to examine the relationship between a host and virus requires that both are isolated and grown in pure culture (Hyman and Abedon, 2010). For environmental microorganisms, it is estimated that only a minute number have been cultured due to the difficulty of isolation (Staley and Konopka, 1985; Hugenholtz, 2002). In the recent literature, many researchers have conducted studies utilizing metagenomic techniques [e.g., (Tyson et al., 2004; Venter et al., 2004)], which consist of obtaining total DNA sequence information from a mixture of organisms from a community. However, without culturing, it has been difficult or impossible to determine host range solely from host and virus genome sequences. The relatively recent discovery of CRISPR-Cas systems has allowed for direct inference of virus host range.

The CRISPR-Cas system comprises a “clustered regularly interspaced short palindromic repeat” (CRISPR) locus and a set of CRISPR-associated (Cas) proteins that function as an adaptive immune system in many Bacteria and Archaea [reviewed extensively in (Horvath and Barrangou, 2010; Karginov and Hannon, 2010; Marraffini and Sontheimer, 2010)]. The CRISPR-Cas system was first discovered in bacterial genome sequences due to the fact that the locus consists of non-unique sequences (23-47 nucleotides) interrupted by unique sequences of similar length (21-72 nucleotides) (Horvath and Barrangou, 2010). The non-unique sequences were termed repeats due to the conservation of these sequences within a single locus and the unique intervening sequences were termed spacers. A suite of CRISPR-associated (*cas*) genes and a short sequence (~150 nucleotides), termed the leader sequence, immediately flank CRISPR loci (Makarova et al., 2011).

Little was known regarding the function of these loci until three different studies reported that many spacer sequences have sequence similarity to virus and plasmid sequences in public

databases (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). A fourth study examined the genes that flank CRISPR loci and realized that they may be functionally similar to those involved in RNAi, a defense system in eukaryotes (Makarova et al., 2006). The CRISPR-Cas system was not definitively shown to be an adaptive immune system until 2007 when Barrangou et al. experimentally proved that a CRISPR-Cas system in a bacterium, *Streptococcus thermophilus*, was able to provide immunity to a phage strain (Barrangou et al., 2007). This study was the first of many to describe the requirements for CRISPR-mediated immunity. Briefly, it was determined phage-immune bacteria need at least a single spacer in a functioning CRISPR-Cas system that perfectly matches a sequence in the phage (Barrangou et al., 2007).

Spacers are derived from targeted sequences, such as viruses (bacteriophage and archaeal viruses), plasmids, transposons, and other mobile elements (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005; Horvath et al., 2008). Proto-spacers refer to sequences within targets that become incorporated as spacers in CRISPR loci (Mojica et al., 2009). Each CRISPR-Cas system has a specific proto-spacer adjacent motif (PAM), usually a few nucleotides in length, which directly flanks and thus identifies sequences in targets that can become spacers (Mojica et al., 2009; Horvath and Barrangou, 2010). It is not uncommon for a spacer to appear in multiple locations within a single CRISPR locus due to independent sampling events (Tyson and Banfield, 2008). While availability and sequences of potential spacers may change due to mutations in the PAM or proto-spacer region, the potential pool of spacers can be predicted based on PAM locations in the targeted genome.

Repeat sequences are usually identical within a single active CRISPR locus (Horvath et al., 2008). The term “CRISPR” was coined because many of the first-studied repeat sequences were palindromic. In addition, some repeats were hypothesized to form secondary structures, such as loops (Kunin et al., 2007). It is not uncommon for Bacteria and Archaea to contain more than one locus, but repeat sequences are often different in different loci within a single bacterial or archaeal host (Andersson and Banfield, 2008). CRISPR-Cas systems have also been found on plasmid sequences, suggesting that the system is capable of horizontal transfer (Goltsman et al., 2009). However, different groups of repeat types correspond to certain combinations of Cas protein types (Makarova et al., 2011). In general, Cas proteins have high sequence diversity, thus can vary significantly between CRISPR loci (Makarova et al., 2011).

The Cas proteins are responsible for spacer acquisition, conversion of CRISPR locus into crRNA transcripts, and targeting and degradation of viral or plasmid sequences [reviewed in (Horvath and Barrangou, 2010)]. Initially, it was difficult to identify the *cas* gene cassettes due to the high level of gene diversity as well as the number of different combinations of genes (Makarova et al., 2006; Makarova et al., 2011). Recently, the different combinations of Cas proteins have been reorganized into different types (I, II, III, U) with different sub-types to separate the different CRISPR-Cas systems (Makarova et al., 2011). Only Cas1 and Cas2 proteins are found in all functional CRISPR system types. The main basis of separation is the identity and organization of the core Cas proteins (Makarova et al., 2011). For a host cell to obtain CRISPR-mediated resistance to a virus or plasmid, the host must have: a functional and complete CRISPR-Cas system, at least one spacer matching a proto-spacer in the target sequence and a PAM flanking the proto-spacer target region (Horvath and Barrangou, 2010). Recently, it was discovered that 100% identity was not required for resistance for Type II systems (Sapranaukas et al., 2011; Semenova et al., 2011; Wiedenheft et al., 2011). Only mutations localized near the PAM end of the proto-spacer were effective in evading CRISPR immunity (Sapranaukas et al., 2011; Semenova et al., 2011; Wiedenheft et al., 2011).

Spacers are incorporated into the CRISPR locus in a unidirectional manner (Barrangou et al., 2007). All new spacers are added to the leader end of the locus; older spacers accumulate at the other end (trailer) (Barrangou et al., 2007; Tyson and Banfield, 2008). Thus, the CRISPR locus can be seen as a historical timeline of virus and plasmid exposure. While there may be spacer deletions internally within the locus (Tyson and Banfield, 2008), the order of spacers in the locus provides historical information. Since spacers are derived from viruses and plasmids, spacers serve as a direct link that can be used to infer host range. Spacers with matches to co-existing viral or plasmid sequences have been shown to generally cluster near the leader end (Andersson and Banfield, 2008). While these relatively newer spacers have targets, the older spacers at the trailer end tend not to have any matches due to the appearance of escape mutations in the targeted sequences (Andersson and Banfield, 2008; Tyson and Banfield, 2008). Over time, the spacers are expected to lose effectiveness due to mutation in the virus or plasmid. Thus, there is a delicate balance as the hosts with active CRISPRs strive to acquire spacers while the viruses evade via sequence mutations.

Almost all prior studies of CRISPR-Cas systems have focused on pure cultures that were investigated over very short time frames. In part, this reflects the significant bias within the CRISPR research field toward biochemical characterization of Cas proteins and complexes. Despite this emphasis, it can be argued that the CRISPR-Cas system evolved to function in the context of populations, and to maintain long-term balance between viral survival and host survival. From the perspective of a single individual, CRISPR function may represent a poor gamble in a game that it can never win in the long term. However, at the population level, the maintenance of high levels of diversity in host immunity is an effective strategy for species survival. For this reason, understanding of CRISPR-Cas as an immune system requires that it be investigated over long time scales and in non-clonal systems. The first step toward this goal is to initiate a multi-generation experiment with a clonal host and single phage, allowing both to evolve – the host by spacer acquisition and the phage by acquisition of escape mutations (mutations that defeat spacer-based immunity). The logical choice for an experimental system for such research is *Streptococcus thermophilus* and its isolated phage, as this is currently the only case in which spacer addition can be promoted simply by phage challenge under laboratory conditions.

The CRISPR-Cas systems in *Streptococcus thermophilus* DGCC7710 are arguably the most thoroughly studied. *S. thermophilus* is a low G+C gram-positive bacteria that is responsible for the manufacture of dairy products, such as yogurt (Bolotin et al., 2004). The genome of strain DGCC7710 has four CRISPR-Cas systems (CRISPR1-4) (Horvath et al., 2008). Via challenges with phage, it was determined that only CRISPR1 and CRISPR3 are functional and capable of acquiring new spacers and providing resistance (Horvath et al., 2008). While both CRISPR1 and CRISPR3 are type II-A CRISPR-Cas systems, they have different CRISPR repeats. The most active, CRISPR1, has 32 spacers in the wildtype strain while CRISPR3 contains 12 spacers (Horvath et al., 2008). The wildtype spacers have some similar and exact matches to different strains of phage. However, these spacers have no sequence matches to phage 2972, a lytic phage capable of successful infection of *S. thermophilus* (Horvath et al., 2008). Phage 2972 is a pac-type phage that belongs to *Siphoviridae* (Levesque et al., 2005). Both host CRISPR and phage genome have been sequenced (Levesque et al., 2005; Horvath et al., 2008).

Characterization of co-evolutionary change in diversifying phage and host populations requires the ability to deeply sample sequence and spacer content variation in coexisting individuals. Metagenomics methods are ideally suited to such a task. In metagenomic studies,

whole community DNA (including DNA from both hosts and viruses) is extracted, fragmented, and subsequently sequenced [theory and methodology reviewed in (Handelsman, 2004; Riesenfeld et al., 2004)]. The sequences are then reassembled *in silico*, producing complete or near complete reconstructed genomes. The first metagenomic studies were completed with Sanger sequencing technology (Tyson et al., 2004; Venter et al., 2004). However, second generation sequencing methods, such as Roche 454 and Illumina technologies have rapidly replaced Sanger sequencing due to significant decreases in price as well as increases in throughput.

Co-culturing *S. thermophilus* with infective phage 2972 and examining the host and phage populations after interaction could reveal genomic sites and frequency of phage mutations and document overall levels of sequence levels, as well as rates of change in these parameters. Such experiments can also be used to document the types, combinations and frequency of spacer addition events in the CRISPR locus. Specifically, with bioinformatics tools, both the local spacer sequences in CRISPR loci and viral genomes can be reconstructed *in silico*. In this closed system, any spacer additions to the CRISPR1 and CRISPR3 locus or nucleotide mutations occurring in the phage genome should result from direct interaction. This fine scale examination of the host-phage relationship is challenging in more complex, open environments because other variables can come into play, such as immigration of other bacteria and phage strains.

The majority of previous studies have investigated CRISPR regions in microbes in pure culture. While this approach is necessary to determine the mechanisms of CRISPR function and diversification of CRISPRs in microbes that can be genetically manipulated, it is equally important to comprehend how these CRISPRs behave in the context of naturally occurring communities of Bacteria, Archaea, and Viruses. However, as this system is artificial, this interaction may not be a reflection of how natural populations behave. Examining a model natural system would enable the investigation of long-term population histories across years, which is difficult to complete using isolates. An ideal natural model system is the acid mine drainage (AMD) site within the Richmond Mine at Iron Mountain, CA, USA (Denef et al., 2010a). AMD is created due to exposure of pyrite (FeS_2) to oxygen and water due to extensive mining (Johnson and Hallberg, 2005). Members of the microbial communities that grow as biofilms on top of the highly acidic, metal rich AMD solution significantly accelerate iron oxidation (Baker and Banfield, 2003). Due to environmental conditions, the AMD system has a relatively low diversity of microorganisms, as compared to environments such as the ocean and soil (Tringe et al., 2005).

The dominant bacterium in the Richmond Mine AMD system is *Leptospirillum* group II, a member of the *Nitrospira* class (Schrenk et al., 1998; Baker and Banfield, 2003). The dominant archaea in the system are lineages from *Thermoplasmata* class, such as E-plasma and G-plasma (Edwards et al., 2000; Baker and Banfield, 2003). The majority of Bacteria and Archaea in the AMD site have at least one CRISPR-Cas system (Andersson and Banfield, 2008). *Leptospirillum* group II has one CRISPR-Cas system while dominant archaea G-plasma and E-plasma both have multiple loci within their genomes (Andersson and Banfield, 2008). Using spacers to determine host range, it was shown the phage AMDV1 targets *Leptospirillum* group II. A number of archaeal viruses such as AMDV2 and AMDV3 target E-plasma and G-plasma, respectively (Andersson and Banfield, 2008). These phage and archaeal viral genomes have been partially reconstructed *in silico*. Another advantage of the Richmond Mine system is that similar samples can be collected from the same set of locations over long periods. For example, genomic sequence information is now available for biofilm communities collected over ten years. This

provides the opportunity to track CRISPR and virus distribution patterns over time and space in a natural system.

The power of the CRISPR-Cas immune system lies in its ability to adapt quickly, likely resulting in host strains containing different CRISPR-based immunities. Examining the CRISPR loci of host populations in these two systems provided new perspective on the response of viral populations to CRISPR-based defense and new insight into the patterns of spacer addition and loss, and the evolutionary consequences of these steps.

Acknowledgements

It would have not been possible to write this dissertation without the mentorship, support, and encouragement of Professor Jillian F. Banfield. Jill has introduced me to the field of community genomics (and to the biofilms of acid mine drainage). I can only hope that I can one day inspire students with the same passion and dedication she has shown towards research and teaching. I would like to thank the other members of my dissertation committee—Professor Steven Brenner, Professor Jennifer Doudna, and Professor Steven Lindow—for their invaluable guidance and for their evaluation of this dissertation.

Both previous and current members of the Banfield lab have significantly assisted in my development as a scientist. I would like to especially thank Brian Thomas for mentoring on various projects, invaluable discussions, and introducing me to Ruby programming. In addition, I would like to thank Vincent Deneff, Brett Baker, Paul Wilmes, Greg Dick, Sheri Simmons, Anders Andersson, Kelly Wrighton, Ryan Mueller, Chris Miller, Anders Norman, David Paez, Joanne Emerson, Daniela Goltsman, Nicholas Justice, and Susan Spaulding for helpful discussions and assistance with laboratory work and computer analyses.

Research completed in this dissertation was the result of collaborative work. I would like to thank everyone involved in the studies contained in this dissertation, with special thanks to Rodolphe Barrangou and Philippe Horvath from DuPont Nutrition and Health, Kimberly Pause Tucker and Mya Breitbart from University of South Florida, and Wayne Getz from University of California, Berkeley.

Funding for work completed in this thesis was provided: the DOE Genomics: GTL Program (DE-FG02-07ER64505), the Army Research Office (W911NF-10-0046), and DuPont Nutrition and Health. I would like to thank Mr. T.W. Arman (owner, Iron Mountain Mines) and Dr. R. Sugarek (EPA) for providing site access to Richmond Mine, Mr. R. Carver for on-site assistance, and all the current as well as previous members of the Banfield lab for sampling at Iron Mountain. In addition, I would like to thank Rocio Sanchez and Dana Jantz for administrative assistance in the Department of Plant and Microbial Biology.

Finally, I would like to thank my family and friends who have supported me through my entire graduate career as well as through the writing of this dissertation.

Chapter 1

Phage mutations in response to CRISPR diversification in a bacterial population

Abstract

Interactions between Bacteria and their co-existing phage populations impact evolution and can strongly influence biogeochemical processes in natural ecosystems. Periodically, mutation or migration results in exposure of a host to a phage to which it has no immunity; alternatively, a phage may be exposed to a host it cannot infect. To explore the processes by which co-existing, co-evolving hosts and phage populations establish, we cultured *Streptococcus thermophilus* DGCC7710 with phage 2972 and tracked CRISPR diversification and host-phage co-evolution in a population derived from a colony that acquired initial CRISPR-encoded immunity. After one week of co-culturing, the co-existing host-phage populations were metagenomically characterized using 454 FLX Titanium sequencing. The evolved genomes were compared to reference genomes to identify newly incorporated spacers in *S. thermophilus* DGCC7710 and recently acquired single nucleotide polymorphisms (SNPs) in phage 2972. Following phage exposure, acquisition of immune elements (spacers) lead to a genetically diverse population with multiple sub-dominant strain lineages. Phage mutations that circumvented three early immunization events were localized in the proto-spacer adjacent motif (PAM) or near the PAM end of the proto-spacer, suggesting a strong selective advantage for the phage that mutate in this region. The sequential fixation or near fixation of these single mutations indicates selection events so severe that single phage genotypes ultimately gave rise to all surviving lineages and potentially carry to fixation traits unrelated to immunity.

Introduction

Viruses of bacteria and archaea have the ability to alter microbial community composition and structure via predation and to impact microbial evolution through lateral gene transfer (Fuhrman, 1999; Brussow et al., 2004; Weinbauer and Rassoulzadegan, 2004). The fastest evolving regions of microbial genomes are arguably those involved in resistance to invasive elements (Zheng et al., 2004). Phage mutations to circumvent host immunity lead to an arms race and results in rapid diversification of both host and phage populations (Banfield and Young, 2009).

The CRISPR/Cas (clustered regularly interspaced short palindromic repeats/CRISPR associated sequences) system in bacteria and archaea is based on short nucleotide sequences (spacers) derived from viruses, plasmids, and host chromosomes that provide sequence-based immunity (Horvath and Barrangou, 2010). These sequences are derived from sites (proto-spacer regions) adjacent to short recognition sequences in the mobile element genome (proto-spacer adjacent motif, PAM) (Deveau et al., 2008; Mojica et al., 2009). The spacers are added to the CRISPR locus unidirectionally at the leader end, where transcription of the locus initiates (Barrangou et al., 2007; Deveau et al., 2008; Horvath et al., 2008; Horvath and Barrangou, 2010). CRISPR-encoded immunity is mediated through small interfering CRISPR RNAs (crRNAs) that match the targeted mobile element sequence (Brouns et al., 2008). Phages and plasmids can potentially circumvent the CRISPR/Cas system through mutations in the proto-spacer sequence or PAM (Deveau et al., 2008; Garneau et al., 2010; Horvath and Barrangou, 2010; Sapranauskas et al., 2011). In some CRISPR systems, studies show that mutations within the ‘seed’ of the proto-spacer (nucleotides closest to the PAM) allow for phage escape of the CRISPR system while mutations elsewhere in the proto-spacer may be ineffective (Sapranauskas

et al., 2011; Semenova et al., 2011; Wiedenheft et al., 2011). Spacers that retain effectiveness against co-existing phage are concentrated on the leader end whereas older spacers provide a historical record of viral and plasmid immunity (Andersson and Banfield, 2008; Tyson and Banfield, 2008; Heidelberg et al., 2009).

Recent laboratory studies indicate that CRISPR spacer addition can occur on the timescale of hours (Barrangou et al., 2007) and field studies point to very high-level heterogeneity of CRISPR spacer content in natural populations (Andersson and Banfield, 2008; Tyson and Banfield, 2008; Heidelberg et al., 2009; Held and Whitaker, 2009). Further, ecosystem studies have revealed high levels of sequence variation within natural phage populations (Andersson and Banfield, 2008). Thus, CRISPR-based immunity could drive the emergence of polymorphic “clouds” consisting of co-existing, co-evolving bacterial and phage populations with different levels of immunity and infectivity (Banfield and Young, 2009). However, dramatic loss of diversity can result if a host population is exposed to a phage for which it has limited CRISPR-based immunity or a host population acquires immunity that is difficult for the phage to escape. In natural systems, a balance of these processes must occur periodically.

Notwithstanding recent advances in CRISPR research, the interplay between CRISPR diversification and phage-based selection that shapes co-existing population structures is not well understood. Similarly, the impact of spacer acquisition and host population diversification on phage genome evolution remains enigmatic. One approach to studying these interactions is to consider the case when a bacterial population is exposed to a phage for which it has no immunity or when a phage is introduced to an immune bacterial host population. Here, we conducted a laboratory experiment to investigate the initial stages of the establishment of co-existing, co-evolving bacterial and phage populations and to evaluate the nature and magnitude of selection events and the genetic nature of phage escape mutants.

Results

Streptococcus thermophilus DGCC7710 was challenged by infective phage 2972. Initially, *S. thermophilus* contained no spacer sequences that conferred immunity to phage 2972 in either of the two functional CRISPR loci (CRISPR1 and CRISPR3). After the primary challenge, we selected a single colony that contained both *S. thermophilus* that had acquired CRISPR-encoded immunity and its co-existing phage and inoculated it into milk. After three successive serial transfers, the coexisting host and phage populations (sample TS-H467) were metagenomically sequenced via 454 FLX Titanium. Additional sequencing was obtained for PCR amplicons at two later time points (samples TS-H467-a and TS-H467-b). An overview of the experimental is shown in Figure 1.1 (see Materials and Methods for further detail).

The CRISPR1 and CRISPR3 loci of each individual bacterium were randomly sampled with multiple short reads. On average, each base of the host genome was sampled 275 times (275 x genome coverage). At the conclusion of the experiment, no wild-type CRISPR1 and CRISPR3 sequences were recovered. All surviving host cells sampled contained at least one newly acquired spacer matching the phage 2972 genome. Typically, short reads did not span more than two repeat-spacer units at the leader end. In sample TS-H467, we retrieved 101 and 13 read sequences for CRISPR1 and CRISPR3, respectively, from the leader end that contain newly acquired spacers (CRISPR1: Figure 1.2A and Table 1.1; CRISPR3: Table 1.1). CRISPR1, the more active locus (Horvath et al., 2008), incorporated 37 different spacers while CRISPR3 only

added 3 spacers (Table 1.1). Only single spacers were incorporated into CRISPR3 while multiple spacers were iteratively added to CRISPR1 (Figure 1.2A).

Of those read sequences that sampled the region flanking the wild-type spacers in CRISPR1, all contained spacer1 (bright red box in Figure 1.2A). Reconstruction of the CRISPR1 loci shows greater spacer diversity in the position immediately flanking the leader sequence (newest spacer) than in the positions closest to the wild-type spacers, as expected. 12 out of the 13 newly acquired spacers that were only sampled once are immediately flanking the leader sequence. In addition, all the newly acquired spacers obtained from TS-H467 have proto-spacers with corresponding PAMs (Table 1.1, Figure 1.2B), indicating that PAMs are required for functional spacer acquisition. While 22% of the PAMs occur on the anti-sense strand, only 5% of the spacers derived from this strand, consistent with evidence of CRISPR loci preferentially incorporating spacers that match the sense strand of the phage (Deveau et al., 2008).

Further CRISPR diversification occurred in samples TS-H467-a and TS-H467-b (■ and ■■ in Figure 3), which were collected 2 and 4 transfers after TS-H467, respectively (Figure 1.1). 7 novel spacers were detected, 4 of which were immediately upstream of the fixed spacer. Notably, there are instances of a spacer independently acquired more than once, as evidenced by the same spacers found with different flanking spacer sequences (different locus context) (Figure 1.3, dotted box).

At the end of the experiment, all 454 FLX Titanium sequencing reads from sample TS-H467 were mapped to the phage 2972 genome sequence (average coverage 359x), and all positions examined for newly acquired polymorphisms. We found that all phage sequences recovered contained a synonymous mutation (SNP-i, Figure 1.4, Figure S1.1) in a proto-spacer region. The mutation occurs within a phage anti-receptor gene, 8 nucleotides from the PAM. The mutation is also 3 nucleotides from the nuclease cut site (Garneau et al., 2010), which is indicated by a dashed line in Figure 4. This mutation likely circumvents spacer1, which we infer to have conferred initial immunity based on inexact matching to phage 2972 (Figure 1.4, Table 1.1). Additionally, 88.2% of the phage reads contain SNP-ii (Figure S1.2, a synonymous mutation in the helicase gene, 6 nucleotides from the PAM and immediately adjacent to the nuclease cut site. This mutation likely thwarts a spacer in the +1 position (spacer32). Approximately 91.7% of phage reads contain SNP-iii (Figure S1.3), a non-synonymous substitution in the primase gene. Thus mutation occurs in the PAM of another spacer in the +1 position (spacer6) and likely inactivates it (Figure 1.2A). No phage mutation occurred in regions targeted by acquired CRISPR3 spacers.

Discussion and Conclusion

Streptococcus thermophilus was originally used to demonstrate CRISPR-based phage immunity (Barrangou et al., 2007). Arguably, it is currently the best characterized model system for studying CRISPR-phage interactions. To our knowledge, *S. thermophilus* DGCC7710 remains the only bacterium in which CRISPR immunity through spacer acquisition can be readily activated in the laboratory by phage exposure. Because CRISPR provides the primary phage immune mechanism, *S. thermophilus* is ideal for the study of processes that occur as co-evolving populations establish. Here, we took advantage of this capacity to conduct the first experimental case study in which metagenomic analyses were used to characterize the co-existing phage and host population structure. Given that we did not know how long it would take for mutations to approach fixation (when a mutation rises in abundance to become the only sequence type in the population) in the phage population, our approach was to conduct the

characterization after a period of interaction that would likely capture a small number of mutation events. This would allow the sequence of immunization and escape mutation events to be reconstructed. The results demonstrate that the time frame of the experiment was appropriate, laying the foundation for future time series dynamics experiments. The use of new generation sequencing technologies to deeply metagenomically- characterize diversifying populations provides a powerful general method for the study of CRISPR evolution *in vivo*.

In the challenge experiment, all sequences that sampled the CRISPR region adjacent to wild-type spacers indicate addition of spacer1. We infer this to mean that all surviving bacteria shared the same first new, non-wild-type spacer, spacer1, and that the colony selected for this experiment developed from a single CRISPR bacteriophage insensitive mutant (BIM) that initially incorporated spacer1. Significant CRISPR locus expansion and proliferation of multiple sub-strains of the initial CRISPR BIM occurred after only one week of co-incubation. This indicates very rapid diversification of population immune potential, as expected at the outset of an “arms race”. These changes were rapidly countered by mutations in the phage population, and after only one week, sequences of the wild-type phage were not detected. Even if the wild-type phage was present in levels undetectable by sequencing methods, it is clear that phage with the newly acquired mutations have become dominant in the population. It is notable that the phage mutations within proto-spacers are in a region close to the known site of CRISPR nuclease restriction (Garneau et al., 2010) as well as within the seed region predicted by Type I CRISPR systems (Semenova et al., 2011; Wiedenheft et al., 2011). Assuming phage mutation occurs randomly, the finding indicates very strong selection for phage with mutations close to, or within, the PAM, suggesting that perhaps the seed region also exists for the CRISPR1 locus of *S. thermophilus* (a type II-A CRISPR system)

Potentially, many phage mutation events could have circumvented immunity conferred by the early-acquired spacers. However, only one SNP type in each phage target region (proto-spacer and PAM) was fixed or nearly fixed (Figure 1.4). Thus, we infer that each of the phage mutations was associated with a powerful selection event (Figure 1.5). In fact, the implication of fixation of a single point mutation is that ultimately all phage particles present at the end of the experiment derived from a single phage replication event. This does not imply that the phage population size decreased that of a single burst. Rather, over a short period of time, only phage with the CRISPR spacer inactivating mutation could rise in abundance and ultimately dominate the population. Similarly, the predominance of cells sharing the same early-acquired spacer implies strong selection events relatively early in the establishment of co-existing, co-evolving populations. Based on the incomplete fixation of SNP-ii and SNP-iii, we infer that the severity of selection decreases as the phage and host populations diversify.

Given phage:bacterium ratios as high as 10^2 - 10^3 , it is surprising that some cells that persisted throughout the experiment had CRISPR1 loci only with spacers whose target sites have mutations in all or almost all phage. These hosts could have developed another form of immunity (e.g., envelope or restriction modification-based immunity). Alternatively, these cells may have been infected by defective, non-virulent phage, thus benefiting from superinfection immunity. More generally, infection by non-virulent phage may provide an extended opportunity to mount a CRISPR-encoded response.

In laboratory systems during the initial period of exposure, either the phage or host may be lost if, by chance, one becomes so immune or so infective that it wins the arms race. This may also occur periodically in natural systems. However, natural systems differ in that they are open, and re-colonization can occur by migration or re-immigration of phage or hosts from

surrounding regions. Our results suggest that it is during these re-colonization events that major diversity-purging events are most likely. Interestingly, if spacer acquisition occurs within a non-clonal host population, periodic introduction of new virulent phage could carry a single bacterial genotype to fixation. An implication is that traits unlinked to phage immunity, and themselves potentially not under strong positive selection, can be rapidly fixed. Similarly, introduction of an immune host could induce rapid genomic change in phage. Thus, a general implication is that CRISPR-based immune systems can be associated with major immunity-linked diversity purging events that can re-direct evolutionary trajectories of both host and phage populations.

Materials and Methods

The experiment was designed to track the CRISPR diversification of an initially phage-resistant isolate, escape of initial CRISPR-encoded immunity via phage genome mutation, and subsequent co-evolution of the host and phage populations. A *Streptococcus thermophilus* DGCC7710 strain ($4.10E4$ cfu/ml) (Barrangou et al., 2007) was challenged with the genomically characterized phage 2972 ($1.10E6$ pfu/ml) (Levesque et al., 2005) (Figure 1.1). The four CRISPR loci have been well characterized (Barrangou et al., 2007; Deveau et al., 2008; Horvath et al., 2008). Prior the experiment, the two functional CRISPR loci, CRISPR1 and CRISPR3 (Horvath et al., 2008), were checked with PCR amplification to ensure that the spacers present (wild-type) did not target the phage 2972 genome. Both CRISPR1 and CRISPR3 are type II-A CRISPR systems.

The challenge was conducted in 15 ml of sterile milk [10% (weight/volume) milk powder in water autoclaved for 20 min at 110°C] for 24 hours at 37°C . Plating to recover isolates with initial CRISPR-encoded immunity occurred on the surface of a FSDA medium (Huggins and Sandine, 1984) at 42°C . Colonies were allowed to interact with coexisting phage and diversify on the plate for 24 hours. One single colony was then picked, suspended in 250 μl of sterile milk, and incubated at 37°C for 16 hours. Subsequently, the co-culture was submitted to 2 successive transfers in sterile milk with a 0.5% (volume/volume) inoculation ratio incubated at 37°C for 16 hours. An additional transfer in 100 ml of M17 [5% (weight/volume) lactose] with a 1.5% (volume/volume) inoculation ratio, incubated at 37°C for 16 hours, provided sample TS-H467 (Figure 1.2A). The transfer to M17 medium was performed for ease of molecular work. Two samples from later time points, TS-H467-a and TS-H467-b, were subsequently recovered. After the genomic sample (TS-H467) was recovered, TS-H467-a DNA was collected following 2 further transfers in sterile milk and a single transfer to M17 medium. The second transfer culture was then re-inoculated into sterile milk, transferred one additional time to sterile milk, and then into M17 media for recover of DNA for sample TS-H467-b (Figure 1).

We conducted the initial challenge over a longer period (24 hours rather than 16 hours) than the other liquid cultures (sterile milk and M17) in order to allow enough time for CRISPR bacteriophage insensitive mutants (BIMs) to develop following phage exposure and to remove the large majority of the non-immune host population. The subsequent growth on a plate at an elevated temperature (42°C rather than 37°C) is optimal for CRISPR BIM development. The subsequent transfers, as well as the initial challenge, were all conducted in milk to mimic the natural habitat of the host and phage.

Total genomic DNA (bacteria and phage) extracted from sample TS-H467 was sequenced via 454 FLX Titanium (short read archive, accession number: SRA049615) at the W. M. Keck Center for Comparative and Functional Genomics (University of Illinois, Urbana-Champaign, IL). In addition, DNA was extracted from several colonies isolated from TS-H467

(* in Figure 1.2A), TS-H467-a (■ in Figure 1.3), and TS-H467-b (■■ in Figure 1.3) and subjected to CRISPR PCR amplification using primers targeting the leader end of CRISPR1, as previously described (Horvath et al., 2008). These amplicons were Sanger sequenced.

454-sequencing reads containing at least one ambiguous base were removed and read ends trimmed based on 20/15 neighborhood quality standard (NQS) (Altshuler et al., 2000), as described in (Brockman et al., 2008). Phred (Ewing and Green, 1998; Ewing et al., 1998) was used to trim the Sanger sequencing reads and Cross_match (developed by P. Green, University of Washington) was used to filter and remove vector sequence.

454-sequencing reads were mapped to the phage 2972 reference sequence with gsMapper from the Newbler package (using default parameters) and visualized with Consed (Gordon et al., 1998). The output from gsMapper containing “High-Confidence” differences was used to identify reads with single nucleotide polymorphisms (SNPs) occurring in the proto-spacer or PAM.

Custom Ruby scripts were used to extract spacers from individual 454 and Sanger reads. Results were checked by manual inspection. To remove effects of 454-sequencing artifacts, spacers were grouped via blastclust, with parameters: L = 0.85, S = 85, W = 8. Spacers across individual reads were arrayed and assembled for final presentation (Figure 2A). Proto-spacer positions were found using blastn against the phage 2972 sequence (GenBank accession number: AY699705), with parameters: F = F, G = 2, E = 1. The PAMs NNAGAAW (CRISPR1) and NGGNG (CRISPR3) were identified downstream of the proto-spacer (Horvath et al., 2008).

Figure 1.1. Overview of experiment and sampling time points.

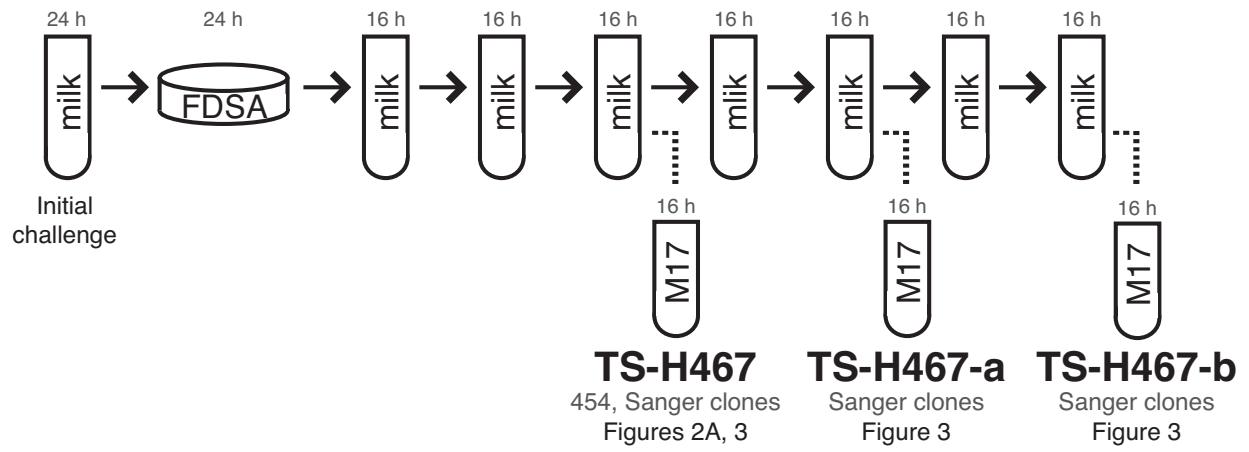


Figure 1.2. CRISPR spacers and single nucleotide polymorphisms (SNPs) in the phage 2972 genome acquired in the experiment. A. CRISPR1 reconstruction showing the leader sequence (L), spacers (boxes) that are wild-type (grey), fixed (bright red), and found in multiple (other colored) and single (white) individuals. Rows represent Sanger (•, see Methods and Materials) and single 454-sequences derived from sample TS-H467. Only sequences with new spacers are shown. White circles indicate spacers rendered ineffective due to phage mutation. B. Proto-spacer regions in phage 2972 targeted by newly acquired CRISPR1 spacers (blue lines on the right = sense strand, blue lines on the left = anti-sense strand) from sample TS-H467. Red lines indicate proto-spacers that contain mutations (SNP-i, SNP-ii, SNP-iii).

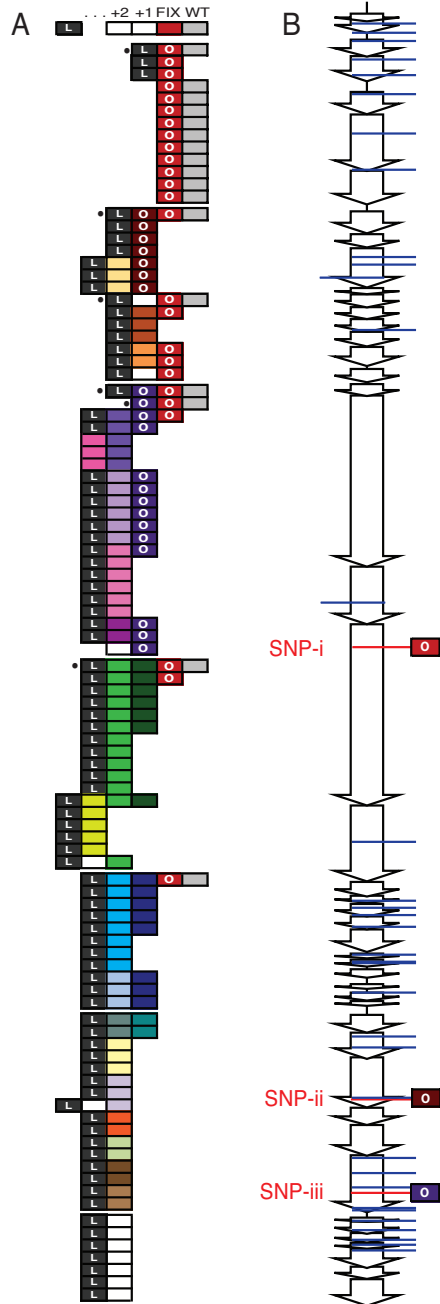


Figure 1.3. Reconstruction of CRISPR loci from multiple time points. Color scheme and symbols are as described in Figure 1. Rows represent single 454-sequences from TS-H467, Sanger sequences from TS-H467-a (■), and Sanger sequences from TS-H467-b (■). Patterned boxes represent spacers previously unique in TS-H467. Note the appearance of the same spacer sequence (indicated by shared color) in two different genomic contexts, indicating two separate spacer incorporation events.

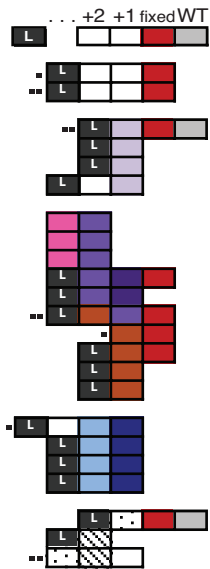


Figure 4. Location of SNPs relative to enzyme cleavage sites and PAMs. In this diagram of proto-spacers and spacers acquired in sample TS-H467, red bases indicate SNPs (top to bottom: SNP-i, SNP-ii, SNP-iii), dashed lines show the enzyme cleavage sites, and boxed bases are PAMs.

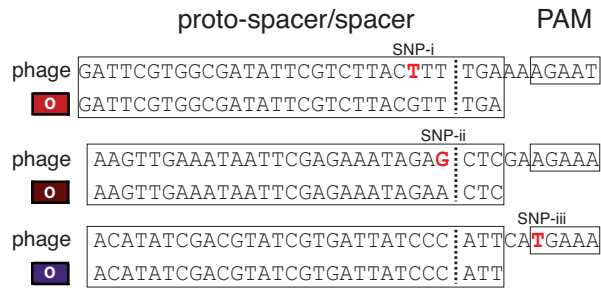


Figure 5. Schematic of theoretical phage population dynamics during co-incubation. Each line represents a phage genotype that is wild-type (WT) or that contains one or more of the following SNPs: SNP-i, SNP-ii, and SNP-iii.

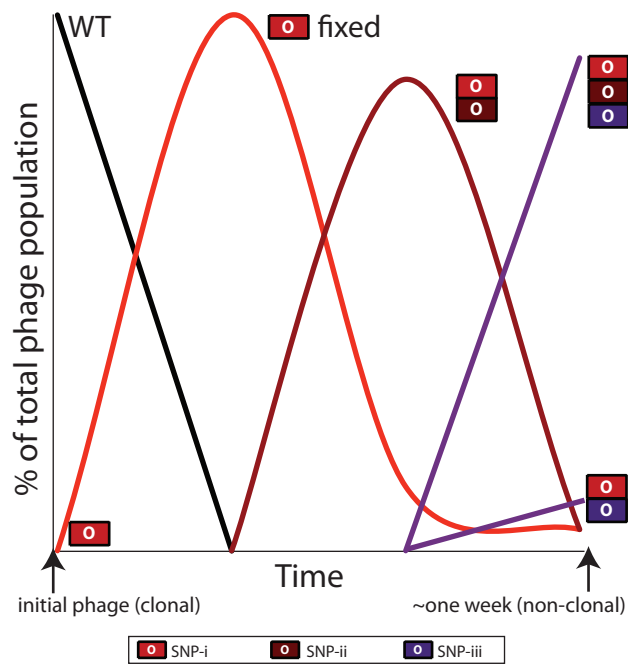














































Table 1. CRISPR1 (CR1) and CRISPR3 (CR3) spacer information. Sequences from CR1 and CR3, corresponding proto-spacer positions on the genome of phage 2972, and SNPs in the genome of phage 2972 acquired in the experiment are listed. In the ID column, symbols indicate imperfect matches to the proto-spacer target (*) attributed to sequencing error, spacers that match the proto-spacer region after mutation (°), and derivation from a Sanger sequence (¶ = TS-H467, ¶¶ = TS-H4670-a, ¶¶¶ = TS-H467-b). Spacer abundance may differ from Figure 2A as only reads with one spacer and leader or multiple spacers are shown in Figure 2A.

Locus	ID	Fig. 2A	Spacer Sequence	Count	Start	End	PAM	SNPs (phage 2972)
CR1	1		GATTCGTGGCGATATTCGTCTTACGTTTGA	32	17182	17211	AGAAT	17206 (SNP-i, G/T)
CR1	2		TACCGAAACGACTGGTTTGAAAAATCAAG	12	30803	30832	AGAAA	
CR1	3		AAATCAGTTTTTTTTCAGAACTTGTCT	6	25582	25611	AGAAA	
CR1	4		ATGCCATTCTTTAAAGAGGCTTTACTCGTT	3	6799	6828	AGAAA	
CR1	5		TCGTTAGAAGTGGATCAACATCTAGTACA	6	22395	22423	AGAAA	
CR1	6		ACATATCGACGTATCGTGATTATCCCATT	12	31709	31737	TGAAA	31740 (SNP-iii, A/T)
CR1	7		ATATCGTCCAGACTATCGCAGAATACTGAT	8	647	676	AGAAA	
CR1	8		TATAACTATTCTAAATTGAAAGGACGTATC	9	26380	26409	AGAAA	
CR1	9		GAACACGTAGGCAAAATATACCGACGAGGTA	2	4434	4463	AGAAA	
CR1	10		GCACTTGATCAAGTAGTGCCAGAAATGGTC	3	8732	8761	AGAAA	
CR1	11		AGATATTGATTATGGTGTAAAGCAGACCA	3	7020	7049	AGAAA	
CR1	12°		CACCAAGAGCGGTGTCCTCAAAGTCCTTG	2	32226	32255	AGAAA	
CR1	13		GTTAGGGATAAGAGTCAAGTGGCCGTCAGG	5	7383	7354	AGAAT	
CR1	14		CAGCTTGAAATGTTTATTGAAGCAGCAGTG	6	24624	24653	AGAAA	
CR1	15		AAAAGATAAAGAACCCTGGAATAATCAGC	6	23958	23987	AGAAA	
CR1	16		CCGATGTGTGGTCACGAAAAACAAACCGACA	2	29210	29239	AGAAG	
CR1	17		TGAAAAAACGAGGAGCACTCGTAGGAGTGG	2	32707	32736	AGAAG	
CR1	18		TGCATGGGAAACATCAGACCAATGGACAGA	1	27861	27890	AGAAT	
CR1	19		AAGACAAGTGGACGGCTTAGAAGATGTGG	1	24289	24318	AGAAA	
CR1	20		TGTTTTAAGTGGTATTATTATATTATCGAA	2	870	899	AGAAA	
CR1	21		TTAGAGACATTTGTGATTTACGACAACCTCA	1	3516	3545	AGAAG	
CR1	22		AAAGATTGCATCATCTAAGAGGTGTTAAAT	2	33108	33137	AGAAA	
CR1	23		TATTGGCATGATTTCAATTTTAATTTGGGAT	4	32135	32164	AGAAA	
CR1	24		CAATACCGTGCCAAGTCTGGTATAATAGTA	3	25373	25402	AGAAA	
CR1	25		CAATACGGAAACCTCCCTTGCCGTCAGCAG	2	16089	16060	AGAAA	
CR1	26		TGGAATTATCCAAGCTGGCTACATGTTAT	1	32503	32532	AGAAT	
CR1	27		TGGATCACTAAGAACTAGTACTGATTTTC	1	24129	24158	AGAAA	
CR1	28		TGGCACACGTGAAATCAAAGGAAGTAACGC	1	33286	33315	AGAAG	
CR1	29		ATTGTCTATTACGACAACATGGAAGATGAT	2	33044	33073	AGAAA	
CR1	30		AATTACTCTAAAACCTAGAGCTCATAAATTG	1	1060	1089	AGAAA	
CR1	31		GACAGCAAGATACACGTAGTAGATGAATTG	1	2446	2475	AGAAA	
CR1	32		AAGTTGAAATAATTCGAGAAATAGAACTC	6	29247	29275	AGAAA	29272 (SNP-ii, A/G)
CR1	33*		TATCTTAAGCGGGTACATCGTCAACGTCTAT	1	1642	1671	AGAAC	
CR1	34*		TAACCTCAGTACAATTTGAAACAGAAATTAGT	1	31204	31233	AGAAT	
CR1	35*		TAATCACTGGACTTTAACCAGACTACATTGG	3	1939	1968	AGAAT	
CR1	36*		TCGATAATCAGCCAAAGTATTAAGTGGTTA	1	27560	27589	AGAAA	
CR1	37*		AAAATCAAAGAAGGAAAATTCAGTTTTTTGT	1	25567	25596	AGAAA	
CR1	38 ^{¶¶¶}		TCAGTCGTTACTGGTGAACCAAGTTTCAAT	2	31583	31611	AGAAA	
CR1	39 ^{¶¶}		TTGGAATTCCTTGTGGAAGACCTTGCGAAT	1	29599	29570	GGAAA	
CR1	40 ^{¶¶}		ACTTCTGAGCGGCTGTGTAAGAGTCTGAC	1	22113	22084	AGAAA	
CR1	41 ^{¶¶}		TGCGTGGAACTGTGAGAACATAGTAGACTG	1	33769	33798	AGAAT	
CR1	42 ^{¶¶¶}		AGAACACTCACTAAATATATAATGGAAAC	1	34083	34111	ATGGA	
CR1	43 ^{¶¶¶}		AGCACAGAAATTTACCAGTTGAGTGGACTAA	1	32053	32082	AGAAA	
CR1	43 ^{¶¶¶}		TTCACTTCGATTCTGAGGCAAGAAGCTCGTG	1	1183	1212	AGAAG	
CR3	44	n/a	TGTCGACTTGTTAAAAAACTACTGAAGA	10	27975	28004	GGCG	
CR3	45*	n/a	ACCATACAATATCCTAGTACTCAACTGATAA	1	17788	17817	GGTG	
CR3	46	n/a	AGTGTCTGTGTGTAAGCATCTTTCCATA	2	22543	22514	GGCG	

Chapter 2

Persisting low-abundance viral sequences shape microbial CRISPR-based immunity

Abstract

Well-studied innate immune systems exist throughout bacteria and archaea, but a more recently discovered genomic locus may offer prokaryotes surprising immunological adaptability. Mediated by a cassette-like genomic locus termed Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR), the microbial adaptive immune system differs from its eukaryotic immune analogues by incorporating new immunities unidirectionally. CRISPR thus stores genomically recoverable timelines of virus-host coevolution in natural organisms refractory to laboratory cultivation. Here we combined a population genetic mathematical model of CRISPR-virus coevolution with six years of metagenomic sequencing to link the recoverable genomic dynamics of CRISPR loci to the unknown population dynamics of virus and host in natural communities. Metagenomic reconstructions in an acid-mine drainage system document CRISPR loci conserving ancestral immune elements to the base-pair across thousands of microbial generations. This ‘trailer-end conservation’ occurs despite rapid viral mutation and despite rapid prokaryotic genomic deletion. The trailer-ends of many reconstructed CRISPR loci are also largely identical across a population. ‘Trailer-end clonality’ occurs despite predictions of host immunological diversity due to negative frequency dependent selection (kill the winner dynamics). Statistical clustering and model simulations explain this lack of diversity by capturing rapid selective sweeps by highly immune CRISPR lineages. Potentially explaining ‘trailer-end conservation,’ we record the first example of a viral bloom overwhelming a CRISPR system. The polyclonal viruses bloom even though they share sequences previously targeted by host CRISPR loci. Simulations show how increasing random genomic deletions in CRISPR loci purges immunological controls on long-lived viral sequences, allowing polyclonal viruses to bloom and depressing host fitness. Our results thus link documented patterns of genomic conservation in CRISPR loci to an evolutionary advantage against persistent viruses. By maintaining old immunities, selection may be tuning CRISPR-mediated immunity against viruses reemerging from lysogeny or migration.

Introduction

Innate immune systems with built-in self/non-self recognition mechanisms have long been known to protect prokaryotic genomes against insertions of foreign DNA (Labrie et al., 2010). For example, well-studied restriction-modification systems often preserve genomic integrity by methylating prokaryotic DNA, enabling prokaryotes to recognize and cleave unmethylated foreign DNA (Wilson and Murray, 1991). Yet, the foreign DNA attacking prokaryotes includes the most abundant and rapidly diversifying members of the biosphere, viruses (Edwards and Rohwer, 2005). With viruses quickly evolving counter-strategies against prokaryotic immune systems, prokaryotes require immunological plasticity to keep pace. Here we computationally predict and directly document the evolution of an adaptive immune system that enables prokaryotes to serially acquire new immunities against diversifying viruses and plasmids. Importantly, the prokaryotic adaptive immune system is genomically encoded (i.e., heritable) and acquires new immune elements unidirectionally, making this adaptive immune system distinct from its eukaryotic analogues (Barrangou et al., 2007; Marraffini and Sontheimer, 2008).

The microbial adaptive immune system is mediated by a genomic locus termed Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR). CRISPR loci have been found in approximately 45% of sequenced bacteria and over 90% of sequenced archaea (Horvath and Barrangou, 2010; Marraffini and Sontheimer, 2010). Utilizing adjacently encoded CRISPR-associated (Cas) proteins (Makarova et al., 2006), CRISPR loci incorporate short 21–72 base-pair sequences from targeted regions in invading viruses and plasmids (Barrangou et al., 2007; Mojica et al., 2009; van der Oost et al., 2009; Garneau et al., 2010; Horvath and Barrangou, 2010; Marraffini and Sontheimer, 2010; Makarova et al., 2011). Once transcribed and processed into CRISPR RNAs, these viral and plasmid-derived sequences confer sequence-specific immunity by binding and cleaving cognate viral and plasmid regions during subsequent genomic invasions (Brouns et al., 2008; Manica et al., 2011).

The viral and plasmid binding sequences incorporated into host CRISPR loci are termed ‘spacers,’ reflecting their addition interspersing highly synonymous 23–47 base-pair sequences, termed ‘repeats’ (Barrangou et al., 2007; Deveau et al., 2008; Horvath et al., 2008). Correspondingly, the targeted viral and plasmid sequences are known as ‘proto-spacers’ (Barrangou et al., 2007; Deveau et al., 2008). With spacer immunity specific to a matching proto-spacer sequence, viruses can escape CRISPR targeting by mutating their proto-spacers or by mutating nearby proto-spacer adjacent motifs (PAMs), regions which likely act as recognition sites for the CRISPR/Cas machinery (Barrangou et al., 2007; Deveau et al., 2008). Natural selection favors the emergence of viruses with mutations in CRISPR-targeted regions, leading to a coevolutionary arms race (Dawkins et al., 1979) as hosts incorporate new spacers to combat viral adaptations (Makarova et al., 2006; Andersson and Banfield, 2008; Deveau et al., 2008). Coevolutionary arms races have been well-documented in other virus-microbe systems (Buckling and Rainey, 2002; Brockhurst et al., 2007; Paterson et al., 2010; Gómez and Buckling, 2011; Morran et al., 2011). Yet, unlike previously studied coevolutionary wars, CRISPR recorded arms races naturally differentiate current host adaptations from previous host adaptations. This is because new spacers are added unidirectionally, adjacent to a leader sequence at a single end of the locus termed the ‘leader-end.’ Previously acquired spacers are also commonly maintained, leaving a cassette-like recording of current (i.e., spacers closest to the leader-end) and past (i.e., spacers farther from the leader-end) adaptations. Partial timelines of coevolution can thus be constructed for host and viral species refractory to laboratory challenge experiments (Edwards and Rohwer, 2005).

Previously, we described one CRISPR recording through metagenomic reconstructions of the CRISPR loci sampled from floating microbial biofilms in an acid mine drainage (AMD) system (Denef et al., 2010a). The prime advantage of probing these generally closed, acidophilic environments is that they are dominated by relatively few species (Wilmes et al., 2008). Our AMD research targeted the extremophilic archaeon I-plasma (Andersson and Banfield, 2008). Growing in an AMD biofilm matrix at temperatures ranging from approximately 30° to 48° Celsius and pHs ranging from approximately 0.3 to 1.2, I-plasma is one of around 12 species in the acidophilic order Thermoplasmatales (Dick et al., 2009; Baker et al., 2010). Reconstructing the CRISPR loci of I-plasma, we noted that the newest, leader-end spacers emerged highly diverse and cell-specific. In contrast, the trailer-end spacers (i.e., the oldest spacers found farthest from the leader sequence) were highly clonal population-wide, matching earlier observations of trailer-end clonality in acidophilic *Leptospirillum* bacteria (Tyson and Banfield, 2008) and more recent observations in bacterial *Escherichia coli* and archaeal *Sulfolobus islandicus* (Held and Whitaker, 2009; Díez-Villaseñor et al., 2010).

Surprisingly, I-plasma's trailer-end spacers appeared conserved despite appearing to provide no immunity against current viruses (Figure S2.1). In reconstructions (~20-fold coverage) of the I-plasma locus in the AMD biofilm, only newly acquired leader-end CRISPR spacers matched currently sampled viruses, implying that previously targeted viral sequences had since evolved or disappeared. Similarly, laboratory challenge experiments (Barrangou et al., 2007; Deveau et al., 2008) document rapid viral evolution in the face of CRISPR targeting.

Here we sought to understand why trailer-end spacers are often conserved despite failing to confer immunity against current viruses. Trailer-end conservation is especially surprising in light of the genomic compactness of Bacteria and Archaea, whose genomes rarely exceed 13MB (Koonin and Wolf, 2008). Prokaryotes have also been shown to delete genetic material approximately ten times as frequently as they insert (Kuo and Ochman, 2009). With a bias toward genomic deletions, we hypothesized that bacteria and archaea would only preserve CRISPR's genetic material if natural selection favored it.

To find and probe the selection pressure driving the preservation of CRISPR trailer-ends, we combined metagenomic reconstructions of CRISPR loci across a multi-year period with a population-genetic mathematical model of virus-CRISPR dynamics in a natural system. Three previous studies have constructed mathematical models of virus-host dynamics in the CRISPR system (He and Deem, 2010; Levin, 2010; Haerter et al., 2011), but none were built to explain why CRISPR loci emerge with both trailer-end clonality and trailer-end conservation. Building a model in which CRISPR locus length is an emergent property of the model parameters, we probe whether tuning parameters to increase trailer-end conservation increases prokaryotic fitness even when viruses mutate rapidly. We further capture the dynamics through which the trailer-ends of CRISPR loci are purged of spacer diversity.

A population-genetic model (see Text S2.1 for the full algorithm) was built to analyze how the intracellular processes of CRISPR and virus mutation drive the long-term development of natural CRISPR loci captured via metagenomic analysis. For simplicity, the model restricts its study of host and viral genomes to monitoring host spacers and viral proto-spacers. All other elements in the genomes are ignored. Host and viral populations are then divided into 'strains': all hosts sharing the same ordered set of spacers are assigned to a single host strain while all viruses with identical proto-spacers are assigned to a single viral strain (Figure S2.2). Each strain's cumulative frequency is tracked across thousands of iterations, as mutations alter host immunity and viral infectivity.

The iterations of the model are not directly dependent on time. Each iteration is instead defined to be the period of variable duration in which a large, preset number of virus-host interactions occurs (Table 2.1). During each virus-host interaction, one of two possible outcomes generally occurs. If the host and viral strains share a spacer, the host survives and the virus is cleared. Conversely, if no spacer is shared, the virus kills the host and the virus survives. Of course, exceptions to both of these situations are allowed in the model. Hosts are given a small probability of surviving even when lacking spacers against an invading virus (Table 2.1). Further, CRISPR is given a small probability of failing to provide immunity even when a host spacer matches an infecting virus' proto-spacer (Table 2.1). This failure rate has been measured in viral plaquing assays conducted by two independent groups (Barrangou et al., 2007; Semenova et al., 2011).

With a large number of interactions per iteration, virus-host interactions are assumed to be well-mixed and distributed according to strain frequencies. Since viruses are most likely to encounter high-frequency host strains, this selects for the viral lines that can kill the dominant

hosts, resulting in negative frequency-dependent selection, a process termed ‘kill the winner’ in microbial ecology (Thingstad and Lignell, 1997). During some interactions, stochastic mutations create new host and viral strains, as hosts unidirectionally add spacers and viruses mutate random proto-spacers. Old host and viral strains are simultaneously depressed in frequency and driven extinct when no longer immune and infective, respectively. At the end of an iteration, the model takes a metagenomic snapshot of the surviving host and viral populations. We analyzed these snapshots across model iterations to capture patterns of CRISPR-driven immunity as they emerge.

Here we describe the main assumptions of the model; a more in-depth analysis of each model assumption can be found in the Supplementary Information (Text S2.2). First, the model assumes that virus and host populations do not go irreversibly extinct. With host and viral populations continually extant, in each iteration the model can simply wait until any preset number of virus-host interactions occurs. We can thus define iterations to be the variable duration period in which such a preset number of interactions occurs. Empirical support for assuming the long-run coexistence of virus and host in natural environments comes from two metagenomic studies. In the first study, Rodriguez-Brito et al., (Rodriguez-Brito et al., 2010) recovered consistently high amounts of virus and host genomes in four aquatic regions across a year-long period. Similarly, in the experimental part of our study, we reconstructed the relative abundances of CRISPR loci and viruses in an acid mine drainage system across the last two years of our six-year metagenomic time series experiment. In each sampling, both host and viral genomes were recovered.

Large microbial population sizes limit the effect of sampling noise in modulating the frequencies (genetic drift) of established strains in our model. But since new mutants arise at low frequencies, we incorporated demographic stochasticity in their ability to establish (i.e., avoid extinction due to a low initial frequency). We did so by allowing new mutants randomly distributed ‘emergence periods’ during which they were not subject to the model's clearance of low-frequency strains. All strains, excluding new mutants in their randomly-sized emergence periods, are cleared when their frequencies drop below a threshold, effecting mutation-selection balance and preventing the model from accumulating an uncontrollable number of strains as new mutants are created. Thus, without the randomness component, the emergence period allows new mutants a chance to reach ‘establishment frequencies,’ after which each mutant can compete in the model solely via its CRISPR-determined fitness.

By increasing the rate at which viable mutants establish, the emergence period increases competition between distinct spacer-adding lines (clonal interference). This promotes ‘kill the winner’ dynamics, making it harder for individual lines to sweep. Despite this increase in competition among beneficial mutants, below we capture losses of trailer-end diversity and rapid selective sweeps. To assure that these results also occur without the emergence period, we tested the model without an emergence period and found both trailer-end clonality and stochastic sweeps (Figure S2.3).

Results

Before analyzing the selective pressure responsible for trailer-end conservation in the single snapshot of CRISPR loci shown in Figure S2.1, we first sought to rigorously determine whether hosts actually preserve CRISPR trailer-ends across evolutionary timescales. To do so, we metagenomically tracked CRISPR spacer content and structure in a natural system over a six-year period. Our analyses focused on an archaeal G-plasma population and abundant viruses that

target it. Like I-plasma, G-plasma is a species in the order Thermoplasmatales (Banfield et al., 2005; Dick et al., 2009). Yet, G-plasma and I-plasma are sufficiently divergent at the rRNA gene sequence and amino acid level to be considered distinct genera (Yelton et al., 2011). Moreover, the lineages show limited genome synteny (Yelton et al., 2011).

To evaluate the extent to which G-plasma CRISPR locus spacers are conserved across time, we metagenomically reconstructed G-plasma CRISPR fragments seven times during the six-year study. In each sampling, the spacers in the CRISPR loci were aligned based on flanking genome sequences and paired read information (Methods). Notably, trailer-end spacers were conserved in both loci across the multi-year period (Figures 2.1 and 2.2).

Spacer preservation occurs despite deletions of single and multiple spacer-repeat units. Deletions of old spacers have also been observed in previous studies (Deveau et al., 2008; Horvath et al., 2008; Tyson and Banfield, 2008; Horvath and Barrangou, 2010). With new spacers more likely to provide immunity against current co-evolving viruses (Andersson and Banfield, 2008), we wondered why trailer-end CRISPR spacers are maintained. To probe whether natural selection conserves old spacers to maintain immunity against persisting viruses, we used the community genomic data across time to reconstruct putative viruses throughout the multi-year period (Methods and Methods). We previously noted that the first reconstructed virus, AMDV3, targets G-plasma. We inferred G-plasma targeting by detecting matches between G-plasma's CRISPR spacers and corresponding 'proto-spacer' sequences in AMDV3 (Andersson and Banfield, 2008). In the current study, a variant of AMDV3, denoted AMDV3b, was reconstructed and shown to also target G-Plasma. Importantly, each viral population is genomically heterogeneous due to single nucleotide polymorphisms (SNPs) and sequence insertions and deletions.

To test whether conserved trailer-end spacers may provide immunity to persisting viruses, we mapped G-plasma CRISPR spacers onto the reconstructed viral genomes (Methods). While most spacers shared between host and viral genomes were found at the new ends of G-plasma loci, several spacers with perfect identity to AMDV3b persist in older regions across all sampled times. The spacers matching AMDV3b are shown with black diamonds in Figure 2.2.

In addition to maintaining trailer-end spacers ('trailer-end conservation'), reconstructed CRISPR loci show far less spacer diversity at trailer-end positions than leader-end positions ('trailer-end clonality'). Unlike conservation, trailer-end clonality could have been expected from single time-point reconstructions, as have been reported previously (Tyson and Banfield, 2008; Held and Whitaker, 2009). Yet, previous analyses could not explain the dynamics through which trailer-end clonality emerges in natural CRISPR loci. In the I-plasma locus (Figure S2.1), all but the four newest spacer positions are clonal population-wide, indicating a recent selective sweep by an immune host lineage. Such a selective sweep is surprising in light of the cell-specific spacer diversity at the new ends of CRISPR loci. With a spacer addition rate high enough to enable numerous lines to acquire distinct beneficial spacers before any one line has swept (i.e., new-end diversity), one expects that competition between spacer-adding lines would prevent selective sweeps in a process known as clonal interference (Gerrish and Lenski, 1998). Further complicating the question of how trailer-end diversity is purged from CRISPR loci is the fact that the loss of trailer-end diversity does not have to occur via selection: it could result from the unidirectional nature of spacer addition. With spacers only incorporated at new-ends, trailer-end spacer diversity cannot increase once trailer-end positions have been filled, because no distinct spacers are incorporated there. Thus, as time progresses, all but one trailer-end lineage, the 'coalescent,' will necessarily go extinct even without selection, resulting in trailer-end clonality.

To ascertain whether selection drives losses of diversity at CRISPR trailer-ends despite high spacer addition rates (an average of eight spacer additions occur per iteration; see Table 2.1), we followed the spacer diversity of computationally reconstructed locus positions for thousands of iterations. We aimed to discover how rapidly locus positions evolved from highly polyclonal to clonal, using rapidity as a marker for sweeps. For simplicity, spacer deletions were removed from the model for this step, as we focused on the role of beneficial mutations (spacer additions) in driving losses of diversity.

As could be expected from the unidirectionality of spacer addition, after thousands of iterations, long-run model trajectories converge to the familiar pattern in which trailer-end spacers are clonal population-wide, while only polyclonal new-end spacers match co-evolving viruses (Figure 2.3 Left Panel). As in Figure S1, the majority of the locus is clonal (as noted on the figure, 128 clonal columns were removed for space conservation). Despite the eventual emergence of trailer-end clonality, CRISPR trailer-ends were initially highly diverse leader-ends (Figure 2.3, Right Panel and Figure S2.4). Interestingly, we reconstructed an intermediate stage in which the trailer-ends can be grouped into several sub-populations distinguished by their oldest spacers, indicating that gradual losses of diversity occur in the model (Figure 2.3 Middle Panel). Trailer-end sub-populations were similarly reported in metagenomic reconstructions from natural environments (Tyson and Banfield, 2008; Held and Whitaker, 2009). By tracking the frequencies of the top 14 spacers in one of the oldest CRISPR locus positions across thousands of iterations, we further verified that spacer fixations can require thousands of iterations (Figure S5).

Yet, in addition to gradual fixations, model results demonstrate rapid selective sweeps of individual host sub-populations. In order to identify sweeps, we created an algorithm that clusters CRISPR loci into an optimized number of sub-populations in any given iteration (Text S2.3). To decide on an ‘optimal’ number of clusters in an iteration, we utilized a machine learning cluster validation technique called the ‘silhouette width’ (Rousseeuw, 1987). We then captured iterations in which the predicted number of CRISPR sub-populations precipitously drops to one, indicating a sweep by a member of one ancestral sub-population (Figure 2.4A). To verify sweeps, we tracked the frequencies of all spacers in a new-end locus position through the period during which the clustering-predicted sweep occurs. Despite competition from numerous other spacers, a single spacer, unique to one diversifying host sub-population (Text S2.3), rapidly rises to high frequency in this position (Figure 2.4B). Importantly, the vast majority of virus-host interactions are immune during the sweep period (Figure 2.5), showing that the rapid loss of host diversity was due to a sweep by a highly immune host rather a bottleneck due to a lack of host fitness.

To understand how a sweep could occur despite model-implemented ‘kill the winner’ dynamics, we reconstructed the strain containing the sweeping spacer identified in Figure 2.4B. We noticed that the two subsequent spacers added on this strain targeted distinct viral sub-populations, immunizing the host against both dominant viral sub-populations (Figure 2.5). In this particular case, the viruses were unable to mutate both matching proto-spacers on a single line prior to the host sweep (Figure 2.5). Thus, while adding spacers that confer immunity to one viral sub-population is common in the model and results in clonal interference among similarly partially immune lines, rapidly acquiring immunity to all viral subpopulations is a rare, ‘multiple mutation’ event (Desai and Fisher, 2007), which leads to a uniquely immune line that can sweep. More generally, this captures how ‘kill the winner’ cannot maintain spacer diversity in CRISPR loci. Viruses cannot always make the requisite mutations needed to kill a host before that host

sweeps. Once trailer-end diversity is lost in even a single rare sweep, trailer-end diversity cannot be regained because distinct spacers are only added at the leader-end.

While unidirectional spacer addition alone explains the emergence of trailer-end clonality, it does not explain the more basic question of why trailer-end spacers are at all preserved despite rarely matching current viruses (Figures 2.1, 2.2). To probe the potential fitness cost associated with rapidly deleting CRISPR spacers, we introduced random spacer deletions into our *in silico* evolving system. Spacer deletion was implemented by allowing a preset fraction of spacer additions to occur with the loss of a randomly-sized, contiguous spacer block from a random starting point in the locus. A combined add/loss mechanism is consistent with experimental evidence indicating that spacer deletion occurs via homologous recombination (Garrett et al., 2011; Gudbergdottir et al., 2011) and data showing that losses often occur with simultaneous new-end spacer additions (Deveau et al., 2008; Palmer and Gilmore, 2010).

If selection played no role (i.e., spacers conferred no immunity) in CRISPR evolution, the equilibrium number of spacers in a strain's CRISPR locus would roughly be the ratio of spacer addition to loss rates. This is the steady state of the linear differential equation $dN/dt = a - d*N$, where N is the number of spacers, a , the spacer addition rate, and, d , the spacer deletion rate. Thus, even with selection extending the size of CRISPR loci to maintain spacer immunity, the long-run equilibrium lengths of CRISPR loci should be inversely proportional to their spacer deletion rates. By incorporating the deletion process into our model, we find that when only 5% of spacer additions occur with deletions, CRISPR locus lengths look qualitatively similar to model results with no deletions, with trailer-end conservation and clonality largely preserved (Figure 2.6A). Conversely, allowing 50% of spacer additions to result in deletions of random spacer blocks purges CRISPR trailer-ends entirely (Figure 2.6B). Given our experimental data showing that CRISPR loci conserve trailer-ends over time (Figures 2.1, 2.2), model results predict that the rate of spacer deletion is maintained below a threshold in many natural systems.

To understand why selection would maintain spacer deletion rates below a threshold, we compared the mean fitness of host strains across time under both low and high-loss regimes. Our measure of host mean fitness in an iteration is the fraction of virus-host interactions in which CRISPR provides immunity. While a low-loss rate (5%) produces consistently high levels of host immunity and thus fitness (Figure 2.6A Lower Panel), dramatic dips in host immunity are observed when the probability of spacer deletion is increased to 50% (Figure 2.6B Lower Panel). Troughs in host immunity predict rapid viral blooms due to the large number of productive virion producing interactions (Figure S2.7).

To understand why host immunity is depressed when CRISPR's spacer deletion rate is increased, we reconstructed host CRISPR loci from the time point at which the fraction of immune hosts is at a trough (iteration 768 in Figure 2.6B). During this predicted viral bloom, the few hosts immune to the top 300 viruses are surprisingly protected by two older spacers (Figure 2.6B Upper Panel). These older spacers were previously far more prevalent among hosts (Figure S2.8). Viral proto-spacer mutation eliminated the selection pressure maintaining the two spacers in the hosts, resulting in the rapid loss of the two spacers from most hosts due to the high spacer deletion rate. Viruses managing to preserve the targeted proto-spacers while avoiding extinction could then bloom, free from spacer-driven immunological control (Figure S2.8).

Importantly, the viral bloom is not monoclonal: a number of sub-populations can be found within the blooming viral population (Figure 2.7A). Further, the main viral sub-population, which contains the two older proto-spacers, is rife with new mutants containing polymorphisms in their proto-spacer sequences (Figure 2.7A). Blooming viral diversity matches

the host diversity evident from the CRISPR loci reconstructed during the bloom (Figure 2.6B). To quantify the correlation between virus and host polyclonalities, we superimposed virus and host strains onto a single matrix, with viral strains allayed along the rows and host strains allayed along the columns (Figure 2.7B). Each (row, column) entry of the matrix represents the number of shared spacers between the row's viral strain and the column's host strain (i.e., the level of immunity). This results in horizontal immunity vectors for each virus and vertical immunity vectors for each host. We then clustered the viral immunity vectors into an optimal number of viral sub-populations by maximizing the 'silhouette width' as above (Text S2.3) and analogously optimally clustered the column-wise host immunity vectors. Immunity clustering shows a clear pattern of specialization in which distinct host sub-populations coexist through immunity to distinct viral sub-populations in what could be termed 'cloud on cloud' immunity (Figure 2.7B). The presence of distinct immunological niches explains why only seven host strains matched the top 300 viral strains (Figure 2.6B); the other hosts survived through immunity to less frequent viruses (Figure 2.7B).

Matching model predictions of a deletion-induced polyclonal viral bloom, we used the community metagenomic data to capture a viral bloom of AMDV3b despite preexisting spacer immunity its host G-plasma population. We tracked the relative abundances of a number of host and viral species in the AMD consortium through a series of samples collected at a single AMD location between June 2006 and August 2007 (Figures 2.8 and S2.9). The G-plasma CRISPR loci from these samplings were shown in Figures 2.1 and 2.2 as reconstructions (3)–(7). Relative abundances of host and viral strains were determined by quantifying the number of reads showing high sequence similarity to the reconstructed composite sequences (Methods). While G-plasma was recovered from all samples across time, G-plasma was highly depleted in the August 2006 sampling, coincident with a bloom in the viruses shown to target it, AMDV3 and AMDV3b. Importantly, Figure 2.2 shows the preexisting presence of trailer-end spacers in G-plasma exactly matching AMDV3b (black diamonds), indicating a putative selective advantage to preserving old spacers and suggesting that spacer deletion between samplings may have driven the rapid proliferation of AMDV3b.

Further supporting model predictions, the viral bloom is polyclonal with a number of sub-populations clearly recognizable (Figure 2.9). A monoclonal rather than polyclonal bloom is the expected outcome when viruses out-mutate host immunity (i.e., the successful viral mutant alone blooms), indicating that the bloom was not the result of a recent viral mutation but instead due to CRISPR failing against a wide range of extant viral sequences. Correspondingly, there is no evidence of diminished CRISPR diversity among bloom-surviving G-plasma hosts. In fact, two G-plasma sub-populations, differentiated by distinct trailer-end spacers, precede and survive the crash (Figure 2.2) as occurs in model simulations in which the deletion rate is high enough to prevent the formation of clonal trailer-ends (Figure 2.6B).

Discussion and Conclusion

Here we metagenomically track virus and host populations across time in a natural environment and use a mathematical model to reconstruct the dynamics through which CRISPR loci could evolve between these snapshots. We first capture surprising selective sweeps through which highly diverse CRISPR 'leader-ends' become clonal 'trailer-ends' across time. Our results also explain why CRISPR loci maintain trailer-end immunities for thousands of microbial generations (immunological memory). Both model and metagenomic data capture blooms of persisting viral sequences against which hosts had preexisting spacer immunity. The model

directly shows how accelerated spacer deletions drive these blooms, with precipitous drops in host fitness occurring when spacer deletion is increased. Without viral persistence as a selection pressure favoring memory in CRISPR loci, genomically compact prokaryotes would be expected to purge trailer-end spacers given documented genomic deletion biases and the eventual cost of maintaining excess genomic material (Kuo and Ochman, 2009).

Of course, the genomic cost is likely not significant for each short spacer added. Yet, if CRISPR loci grew without bound, at some point there would be a cost associated to maintaining and transcribing enormous loci. Evidence for a genomic length cost emerges in two recent studies. An elegant analysis noted that highly expressed eukaryotic genes possess significantly shorter introns than less expressed genes, a fact attributed to the ATP-cost of transcribing even short DNA regions (Castillo-Davis et al., 2002; Carmel and Koonin, 2009). In *Salmonella*, Kuo and Ochman (Kuo and Ochman, 2010) noted that bacterial pseudogenes are deleted faster than they would be by drift alone (which has exponential waiting times), pointing to selection as a driver of genomic compactness (Kuo and Ochman, 2010). As in the eukaryote study, the few enduring bacterial pseudogenes in (Kuo and Ochman, 2010) appear to be less expressed. Interestingly, elongated CRISPR loci may have an answer to the transcriptional cost problem: they appear to disproportionately produce CRISPR RNA at the leader end (Marraffini and Sontheimer, 2010). An intriguing possibility is that CRISPR loci could bet-hedge (Cohen, 1966; Beaumont et al., 2009), with selection tuning the level of trailer-end spacer transcription to scale with the probability of encountering matching viral sequences.

In pinpointing blooms of persisting viruses as the selection pressure favoring CRISPR memory, we noted a surprising polyclonality in both virus and host in the natural system. Had this been the expected, laboratory-observed bloom in which a virus simply mutates around host immunity (Barrangou et al., 2007; Deveau et al., 2008), the result would have been a monoclonal bloom of the viral variant for which the hosts were not able to acquire spacers in time. For a polyclonal bloom to occur, rather than a single lucky viral mutation, host immunity must fail against a large swath of viruses. There are thus two possibilities for how this polyclonal bloom occurred: either the CRISPR system did not provide any immunity at all (i.e., spacers are not immunogenic), or, as in the model, the hosts prematurely deleted key spacers allowing diversified viruses sharing these key old spacers to resurge and bloom. While we cannot entirely dismiss the first possibility, we did simulate the model under the null hypothesis in which spacers are not immunogenic. In that case, when CRISPR loci evolve neutrally, simulated loci emerge with few spacers and no trailer-end clonality. In contrast, naturally sampled G-plasma loci contain tens of spacers and exhibit dichotomous patterns of trailer-end clonality and new-end diversity. More generally, because the rate of neutral fixation of trailer-end spacers scales inversely with the effective population size (Barrett et al., 2006), large microbial populations make genetic drift an unlikely driver of observed CRISPR locus patterns.

Three previous models have been built to study questions surrounding CRISPR-based immunity. Haerter and colleagues studied how viral diversity is maintained against CRISPR, but their model did not track and reconstruct spacer patterns within CRISPR loci (Haerter et al., 2011). Levin (Levin, 2010) focused on the fundamental question of why CRISPR loci are found in some but not all microbes, but did not include virus and host mutational processes. It thus could not capture the long-run evolution of CRISPR loci within microbes that do maintain CRISPRs. To model this long-run evolution, He and Deem (He and Deem, 2010) elegantly applied an HIV-derived differential equation model (Nowak and May, 2001). Yet, in using an HIV model, He and Deem assumed that CRISPR-immunized Bacteria and Archaea control viral

abundances in the same way that cytotoxic CD8⁺ T cells target HIV virions. Thus, viral populations surprisingly decline in their system if all host strains (the viral growth source) are increased by a constant factor, as roughly occurs after an influx of resources. Further, in assuming pre-stipulated locus lengths in which each leader-end spacer addition occurs with a corresponding trailer-end spacer deletion, the model in (Nowak and May, 2001) could not probe whether reducing spacer deletions to increase CRISPR locus lengths is an evolutionarily beneficial strategy.

In protecting against blooms of old viral sequences, model predictions and metagenomic data suggest that CRISPR's immune memory makes it suited for environments in which viruses persist for long periods or remigrate from adjacent regions. CRISPR-based immunity may thus be more prevalent in biofilms than in dilute ocean environments (Sorokin et al., 2010). Immunity against persistent viruses may also explain CRISPR's presence in 90% of sequenced Archaea, which have disproportionately been sampled from extreme environments where viruses tend not to lyse their hosts (Prangishvili et al., 2006; Manica et al., 2011).

More generally, proviral latency is a viral persistence strategy and a clear barrier to eradicating pathogens. A fascinating study recently showed that of the 132 spacers matching viruses in CRISPR loci reconstructed from *Pseudomonas aeruginosa* hospital populations, all spacers matched lysogenic but not lytic viruses (Cady et al., 2011). And while these spacers do not appear to block lysogenization, the same group and others have demonstrated CRISPR-mediated control on inserted lysogens, apparently preventing lysogenic induction and infectious spread across susceptible populations such as biofilms (Zegans et al., 2009; Edgar and Qimron, 2010; Cady and O'Toole, 2011; Palmer and Whiteley, 2011). A potential explanation for the demonstrated connection between CRISPRs and lysogenic viruses could be CRISPRs immunological memory. By maintaining old immunities, CRISPR may have evolved to safeguard against reemergences of ancestral viruses from lysogenic dormancies.

Materials and Methods

For the 2006–2007 time series study, biofilms were sampled from the acid mine drainage solution – air interface at the C +75 m location in the Richmond Mine (Iron Mountain, CA - 40°40' 38.42" N and 122° 31' 19.90" W (Elevation ~3,100')) in June, August, and November 2006, as well as May and August 2007. Environmental parameters of this site at the times of sampling have been reported previously (Denef et al., 2010b). Samples were transferred to dry ice on site and stored at –80°C.

As described in detail previously (Lo et al., 2007), for each biofilm collected, high molecular weight DNA was extracted from a 1 g subsample using phenol-chloroform isoamyl. To further remove contaminating extracellular polysaccharides, the DNA was subsequently run on a gel and purified via a QIAquick Gel Extraction Kit (Qiagen, Venlo, Netherlands). Preparation of shotgun metagenomic libraries and pyrosequencing using the 454 Genome Sequencer FLX-Titanium system were performed at the W. M. Keck Center for Comparative and Functional Genomics (University of Illinois, Urbana-Champaign, IL) according to manufacturer's instructions (454 Life Sciences, Branford, CT) (Margulies et al., 2005). Signal processing and base calling were performed using the bundled 454 Data Analysis Software version 2.0.00.

Sequencing reads from the five libraries were co-assembled using Newbler (GSAssembler v. 2.0.01, Roche) using default parameters except for a 95% nucleotide identity and 40 nt minimum overlap requirement. Replicated reads were identified using a previously described

protocol based on CD-HIT clustering (Gomez-Alvarez et al., 2009) (>95% identity, >five identical bases at the start of the read, no equal length requirement). Within each CD-HIT cluster, reads that shared the same start position on the assembled contigs were identified and removed except for the longest read. Additional filtering of reads containing ambiguous bases, resulted in a total of 990,386 reads (~350 Mbp). A second assembly, using identical parameters, was performed using this filtered reads dataset.

For community profiling, read assignment to previously identified genomic sequence bins was performed by blastn analysis (e-value cutoff of e^{-20}) using a database of contigs previously assembled and binned from four other Richmond Mine biofilm samples: 5-way, collected in March 2002 (Tyson et al., 2004; Simmons et al., 2008), UBA and UBA filtrate collected in June 2005 (Lo et al., 2007; Baker et al., 2010), and UBA-BS collected in November 2005 (Dick et al., 2009).

Contigs representing virus genome fragments were identified based on (a) similarity to previously identified virus contigs recovered from the same system, (b) extreme high depth of sequence coverage (in the case of AMDV3b), (c) assembly curation into genome fragments with detectable sequence similarity to the known viruses, and (in all cases) (d) targeting of the genome sequence fragments by CRISPR spacers. Viruses were determined to replicate in specific hosts based on extensive targeting of their genomes by spacers from host-specific CRISPR loci. Curation of contigs containing reads identified as viral was carried out using Consed (Gordon et al., 1998). Contigs were then imported into GSMapper and extended manually and joined, where appropriate, so that regions fragmented by elevated sequence divergence could be condensed. Cases of extreme divergence were treated as separate contigs. Locations where genomic datasets were fragmented by gene content differences were noted, and the information used as part of the binning procedure. Viral genomes related to the previously studied AMD viruses but that assembled separately were distinguished. For example, the deeply sampled AMDV3b genome is related to the previously reported AMDV3 population and also to a shallowly sampled AMDV3c population (results not shown) that is also present in the C +75 m dataset.

Strainer (Eppley et al., 2007) was used to visualize single nucleotide polymorphism patterns and other forms of variation. This made use of the “.ace” file generated by GSMapper and read re-mapping step that corrects for homopolymer errors during import into Strainer.

CRISPR spacer analysis was performed on individual sequencing reads rather than contigs generated from an automated assembly. Sanger reads (mate-paired ~800 bp sequences from each end of an ~3 kb clone) from the 5-way, UBA, UBA-BS, and UBA filtrate datasets, and 454 reads from the C +75 m series, were used in the reconstruction of both G-plasma CRISPR loci (data are separated by time points in Figures 2.1 and 2.2). Any 454 reads containing at least one ambiguous base (“N”) were removed. Using a custom Ruby script, the ends of each 454 read were trimmed until a base passed 20/15 NQS (neighborhood quality standard) (Altshuler et al., 2000), with a variation described in (Brockman et al., 2008). Cross_match (developed by P. Green, University of Washington) was used to remove any remaining B adaptor sequences (from library construction). Phred (Ewing and Green, 1998; Ewing et al., 1998) was used to trim the Sanger sequencing reads and Cross_match was used to filter vector sequence.

Sequencing reads that sampled the CRISPR loci were identified based on the presence of specific repeat sequences (see below). Custom Ruby scripts were used to extract CRISPR spacer sequences from 454 and Sanger sequencing reads. We allowed for variation in the repeat

sequences to avoid omitting spacer sequences due to errors in sequencing (e.g., homopolymer runs). Spacers were grouped using blastclust (using parameters of 85% identity and 85% length overlap) to remove duplication of groups due to sequencing error. Custom Ruby scripts were used to array CRISPR spacers back onto sequencing reads. Assembly of each locus was manually performed in Microsoft Excel based on overlapping spacer patterns and sampling of the flanking genome on part of the read or its mate pair (in the case of Sanger reads). Where possible, 454 reads were arrayed so that patterns of sequential spacers matched locus regions defined based on Sanger reads. For data presentation in Figs. 2 and 3, unique patterns defined by multiple overlapping 454 reads were condensed to report the longest possible sequence of spacers.

Spacer matches were detected using blastn, with parameters for short sequences ($G = 2$, $E = 1$, $F = F$). Perfect matches signify exact matches (100% identity across entire length of spacer) while imperfect matches require at least 85% identity across at least 85% of the spacer. The databases used in the blast searches were composed of AMDV3b sequences recovered in this study. While the database used to detect imperfect matches only contained contig sequences, the database used to detect perfect matches also included the individual sequencing reads that comprised each of the contigs.

For each individual sample, each read was assigned to a sequence bin (organism or virus type) based on blastn analysis (cutoff $< e^{-20}$). The unassigned category indicates similarity to contigs in the AMD sequence database with unknown affiliation. Note that, as described previously (Denef et al., 2010b), changes in solution pH occurred at the sampling site over the time period studied. This altered the overall community composition, particularly the relative abundances of Bacteria and Archaea.

The mathematical model—see Text S2.1 for the complete algorithm—was programmed and simulated in MATLAB (version 7.7). Model simulations recorded the spacers in all CRISPR loci across iterations, storing distinct spacers as distinct numbers. Images of CRISPR loci (i.e., spacer patterns) were then produced in R (version 2.11). The R ‘Cluster’ package was used to track the evolution and diversity of CRISPR lineages across time.

Figure 2.1. Trailer-end conservation and clonality documented in G-plasma CRISPR loci #1. Metagenomic reconstructions of the first CRISPR locus of a G-plasma population sampled in 2002 (1), 2005 (2), June 2006 (3), August 2006 (4) November 2006 (5), May 2007 (6) and August 2007 (7). In each sampling, the CRISPR spacers (boxes) are aligned horizontally according to their ordering in the metagenomic reads, with CRISPR repeats removed for compactness. Overlapping 454 spacer patterns are also condensed (Methods). The left-ends are the leader-ends, where new spacers are unidirectionally incorporated. Boxes filled with the same color represent identical spacers, with two exceptions. Black-filled boxes show flanking genetic material and white-filled boxes denote cell-specific spacers found only once in the dataset. White gaps reflect unsequenced regions in the metagenomic reconstructions. When separated spacers can be linked via paired reads, the intervening region is shown as a grey bar. Boxes containing a black ‘X’ indicate probable spacer deletions. When spacers match reconstructed AMDV3b viral sequences, diamonds are inserted, with filled diamonds showing perfect matches and open diamonds reflecting imperfect matches. Trailer-end conservation (presumed immunological memory) and clonality are pronounced in this locus, with large numbers of matching spacers preserved across the six-year period. Another example of trailer-end conservation and clonality—in the CRISPR loci of archaeal I-plasma—is shown in Figure S2.1.

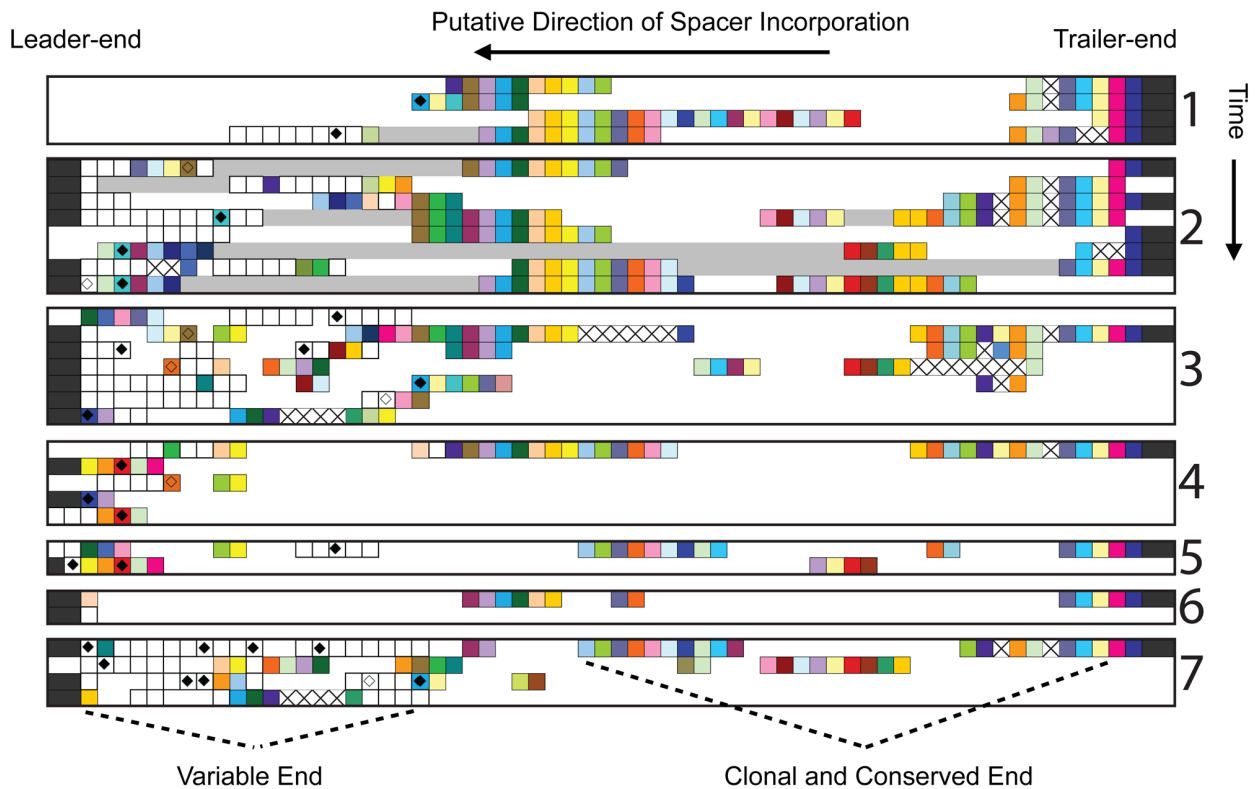


Figure 2.2. Trailer-end spacers of G-plasma CRISPR locus #2 match AMDV3b viral regions across the six-year period. Metagenomic reconstructions of the second CRISPR locus of G-plasma at the seven sampled time points. Notably, several trailer-end G-plasma spacers match reconstructed AMDV3b across all time points (filled diamonds).

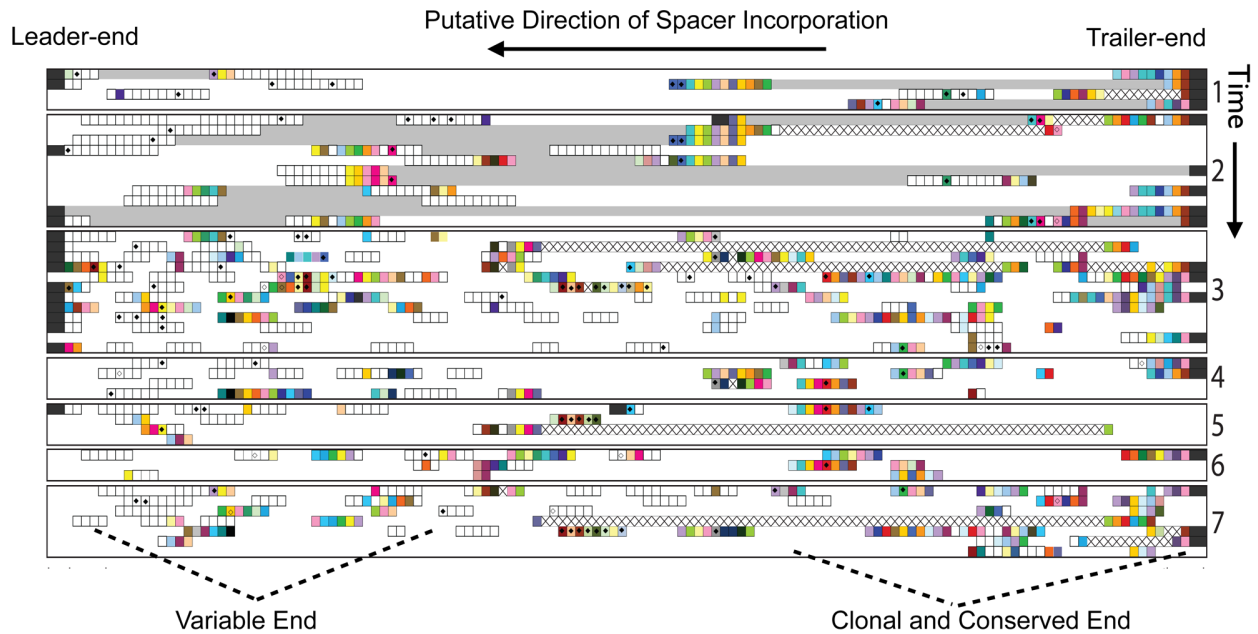


Figure 2.3. Model captures the emergence of trailer-end clonality in CRISPR loci. Computational reconstructions show the loss of trailer-end diversity from CRISPR loci. Reconstructions show the 45 most frequent host strains at the 100th, 500th and 7000th iterations of a representative simulation without spacer deletion. In each panel, the rows show distinct host strains, with their spacers allayed across the columns from right to left as in Figures 2.1, 2.2 and S2.1. Circles indicate spacers perfectly matching any of the 300 most frequent viral strains in that iteration. To preserve space, 128 clonal columns are removed in iteration 7000 prior to the divergence of sub-populations from a common ancestor (arrow). Notably one ancestral population still at low frequency (~ 0.007 as shown in Figure S4) in the 100th iteration is the common ancestor of all surviving strains.

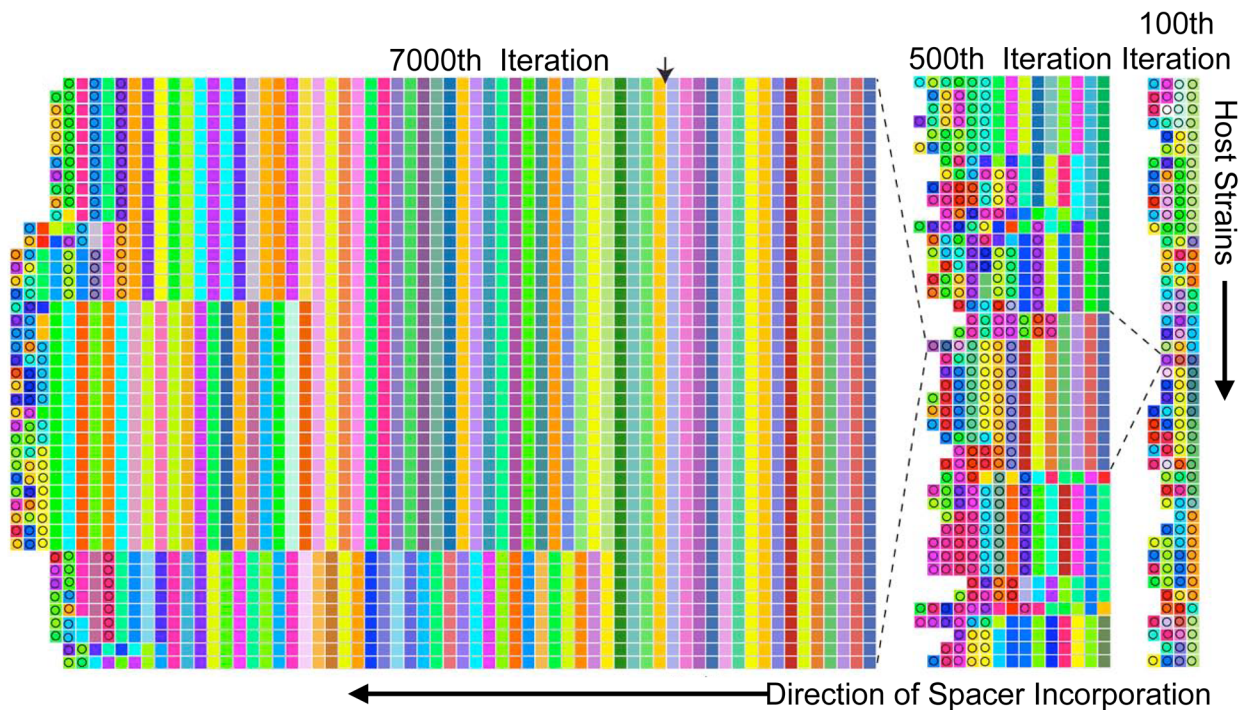


Figure 2.4. A selective sweep of spacer diversity. (A) Hosts CRISPR loci from the simulation in Figure 2.3 are clustered (Text S2.3) into distinct sub-populations every 100 model iterations to capture how trailer-end clonality emerges. Cluster heights represent the cumulative frequencies of all strains in a given cluster, cluster widths show the number of distinct strains in that cluster, and the combined height of all clusters in an iteration reflects the fraction of virus-host interactions that is immune (i.e., host mean fitness). A marked loss of host diversity occurs prior to iteration 3800 (\rightarrow), after which the sweeping sub-population diversifies through distinct leader-end spacer incorporations (Figure S6). (B) The frequencies of all host spacers at a single leader-end column are tracked during the clustering-predicted sweep. A single spacer (shown in black) rapidly rises in frequency before iteration 3800 as predicted by the clustering. Subsequent ‘kill the winner’ oscillations occur before all competing hosts go extinct. A second sweep purges the remaining diversity at this locus position.

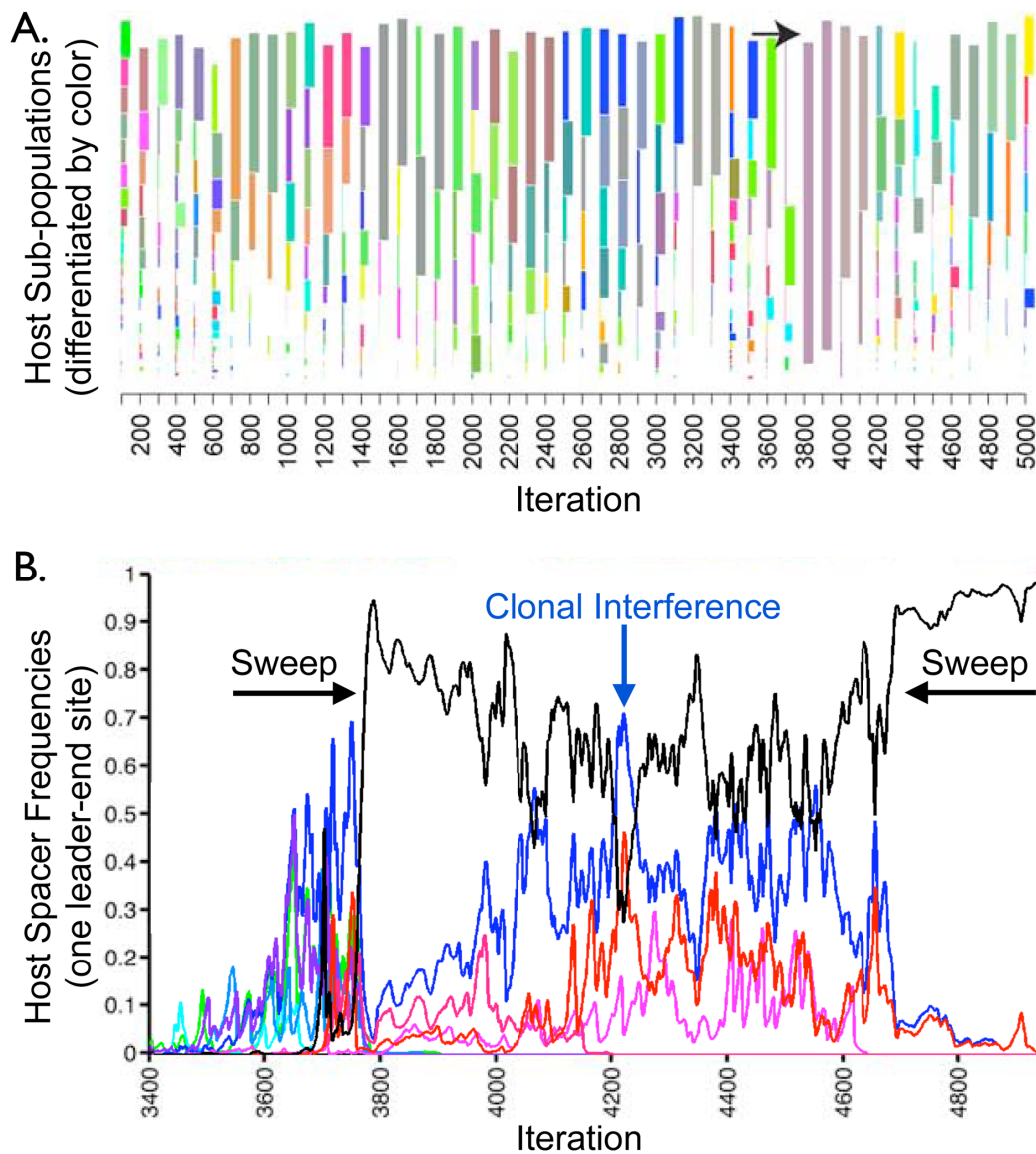


Figure 2.5. Sweep driven by spacer-mediated immunity against multiple viral sub-populations. In the upper panel, the frequency of the sweeping spacer identified in Figure 2.4B is again shown in black. Also tracked, are the two adjacent spacers added by the black spacer's successful host line. The frequencies of these adjacent spacers in their respective locus columns are shown in red and blue. In green, we track the fraction of immune virus-host interactions. The lower panel shows the frequencies of the three corresponding proto-spacers in the viral population. The inverse fluctuations in viral proto-spacer frequencies show that the viruses fail to lose all three proto-spacers on a single line until just prior to iteration 3800, after the sweep. The host line thus sweeps due to immunity to both viral sub-populations.

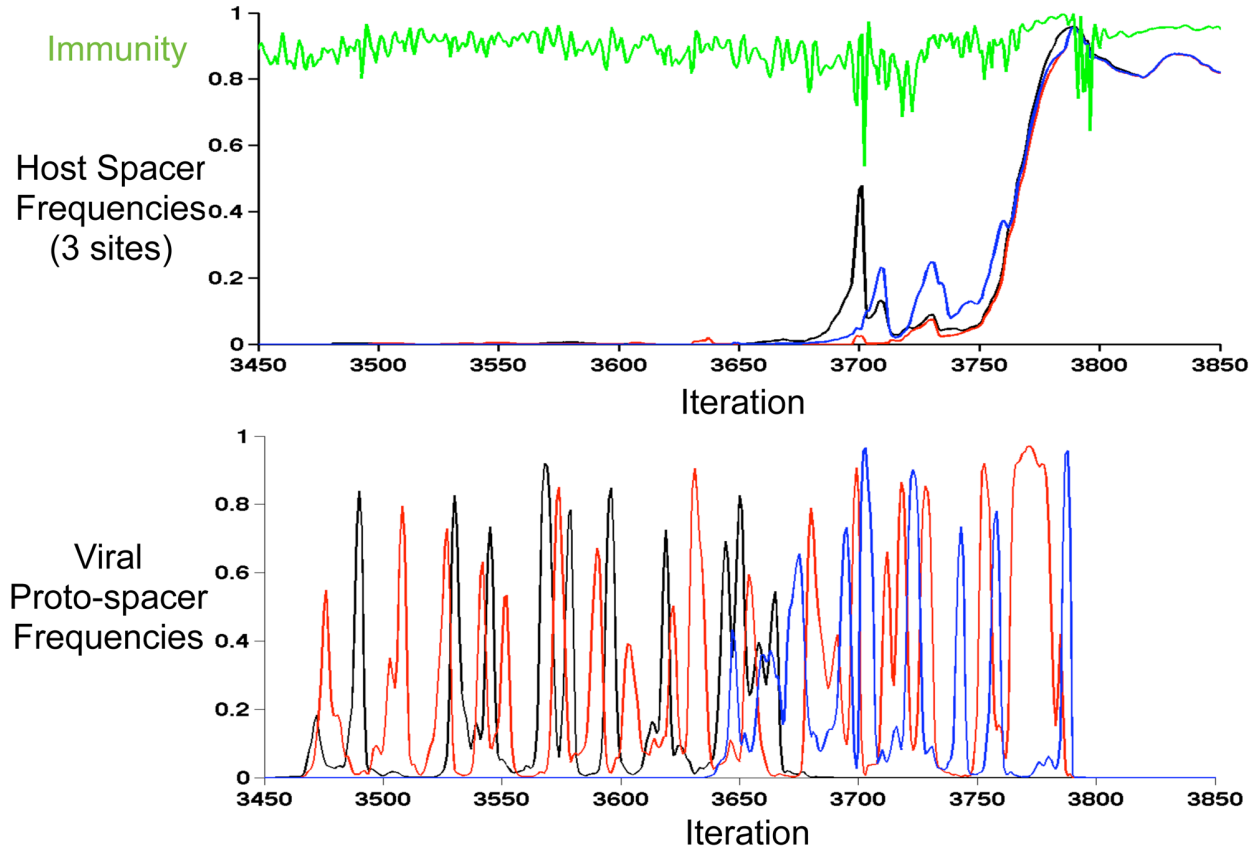


Figure 2.6. Model shows trailer-end conservation protecting hosts against blooms of old viral sequences. The model is extended to allow a parameterized fraction of (single) spacer additions in host CRISPR loci to occur with deletions of randomly-sized blocks of spacers from random locus positions. The lower panels in (A) and (B) plot host immunity (blue) against maximum viral strain frequency (red) in each iteration. (A) When 5% of additions occur with deletions, trailer-end memory and clonality are preserved. Only new-end spacers target current viruses and CRISPR's antiviral immunity is maintained at high levels across thousands of model iterations. (B) When 50% of spacer additions occur with deletions, trailer-end memory and clonality are purged. Depletions in host immunity occur (lower panel), indicating viral blooms due to the large fraction of interactions in which CRISPR fails to provide immunity (*i.e.*, host and virus do not share spacers). During the predicted bloom at iteration 768, immunity against the top 300 viral strains is conferred by two older spacers, which are lost from most host lines prior to the bloom (Figure S2.8).

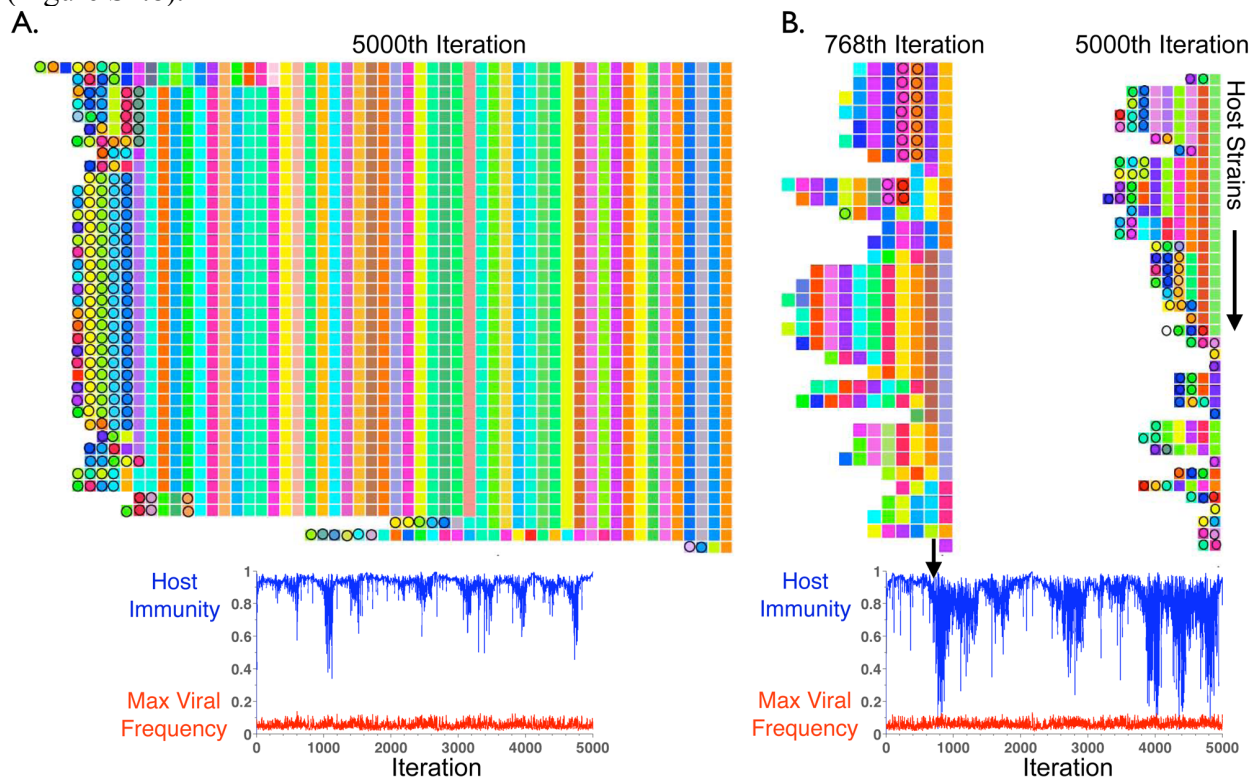


Figure 2.7. Clustering by immunity reveals diversity during viral bloom. (A) The 651 viruses at the model predicted bloom (iteration 768 in Figure 2.6B) are shown along the rows, with the virus' 50 aligned protospacers shown along the columns. Distinct proto-spacers are colored differently. Strains are then clustered based on proto-spacer relatedness. The two proto-spacers providing immunity at the bloom (Figures 2.6B and S2.8), are shifted to the two rightmost columns for clarity. Five distinct viral sub-populations are observed in the mosaic, with the largest blooming sub-population characterized by closely related mutants sharing the two critical proto-spacers in their rightmost columns. (B) Viral and host sub-populations at the bloom are superimposed on one another to reveal ‘cloud on cloud’ immunity at the bloom. The rows contain viral strains and the columns show host strains. Each entry of the heat-mapped ‘immunity matrix’ shows the number of shared spacers between the respective host and viral strain. Pale yellow color represents no shared spacers (susceptibility), yellow one shared spacer, orange two shared spacers, and red three shared spacers. The silhouette width (Text S2.3) was maximized to cluster both hosts (columns) and viruses (rows) into an optimal number of sub-populations based on immunity profiles. Distinct host sub-populations possess immunity to distinct viral sub-populations.

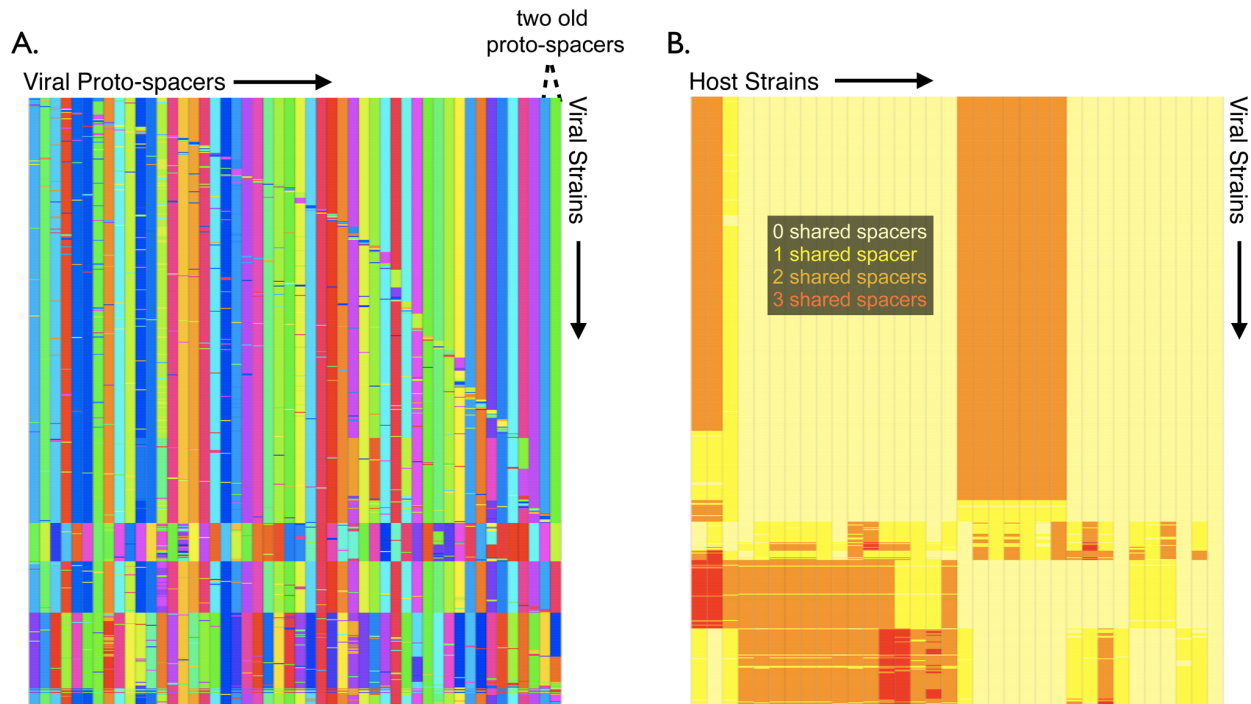


Figure 2.8. Metagenomic sampling across time captures a natural viral bloom. Number of sequencing reads of G-plasma and its viral populations, AMDV3 and AMDV3b, calculated from the community genomic data at a single location across five time points in 2006–2007. Relative abundances of all archaeal, bacterial, viral and plasmid genomes reconstructed from this community during 2006–2007 are shown in Figure S2.9. Both Figures 2.8 and S2.9, capture a bloom of AMDV3b virus (bright red) at the second time point, August 2006, coincident with the depletion of its archaeal G-plasma host (bright green). Notably, the G-plasma CRISPR loci from these time points were reconstructed in samplings (3)–(7) of Figures 2.1 and 2.2. G-plasma contained several spacers exactly matching the blooming AMDV3b sequences prior to the August 2006 bloom (black diamonds in Figure 2.2).

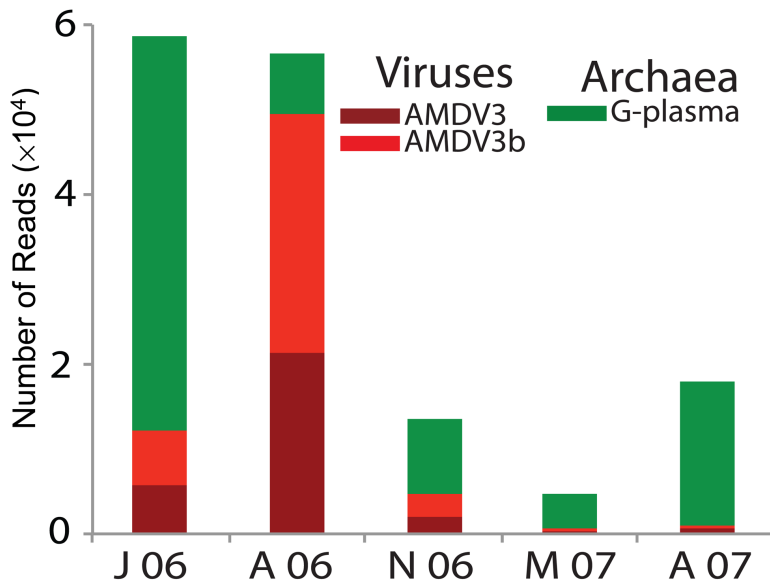


Figure 2.9. Natural viral bloom is polyclonal. Sequence variation within a gene of the blooming AMDV3b viral population (345 bp field of view). The top bar of the figure represents an 18 kb contig sequence of AMDV3b, with predicted genes shown as boxes. Below the contig, is a close-up view of sequence variation within a single gene. White bars represent aligned sequencing reads, while colored bars indicate SNPs relative to the composite sequence. The black region is a large deleted sequence block in one individual viral genome. Distinct viral sub-populations are captured during the bloom, each sharing common SNPs. Also shown in the figure are regions of the AMDV3b contig that match G-plasma spacers: closed circles in the contig represent perfect matches and open circles represent imperfect matches. When the match between G-plasma spacer and AMDV3b protospacer occurs within a predicted gene, the circle is placed inside the gene box; matches to intergenic regions are shown below on the contig line.

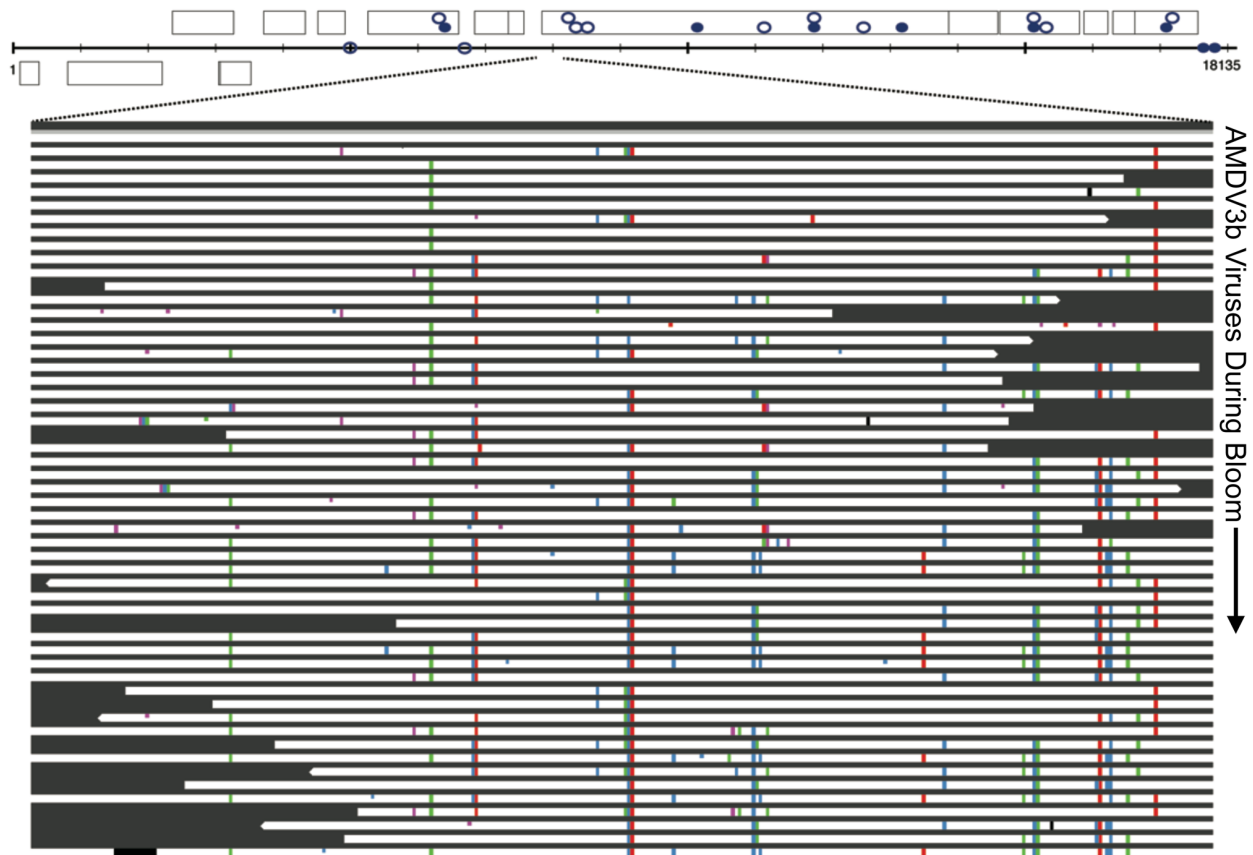


Table 2.1. Table of parameters used in model.

Symbol	Value (Range Probed)	Description
K	10^6 (10^5 – 10^8)	Interactions per iteration.
S	50 (1–300)	Fixed number of proto-spacers per viral genome.
P_{v_mut}	.003 (10^{-4} – $3 \cdot 10^{-3}$)	Probability that viruses mutate a random proto-spacer in an interaction. For bacteria and DNA-based viruses this has been measured at \sim .003 mutations per genome per replication [72].
P_{b_add}	$8 \cdot 10^{-6}$ (10^{-6} – 10^{-4})	Probability that hosts unidirectionally add a random spacer in an interaction, as measured in CRISPR laboratory experiments [10]. With 10^6 interactions per iteration, numerous (<i>e.g.</i> , 8) strains add new spacers per iteration, causing clonal interference (“kill the winner”) and multiple-mutation driven sweeps.
P_{b_lose}	0 (0–1)	Expected frequency of spacer additions in which hosts delete a random spacer block.
$f(n)$	$10^{(-4+n)}$ $n > 0$ $1 - 10^{-9}$ $n = 0$	Given n shared spacers, the probability a virus-host interaction is productive (<i>i.e.</i> , virus lives and host dies). When $n = 0$, f is set to an extremely small but still positive number to prevent host extinction.
i_b	0.1 (.01–0.5)	Fraction of parent strain’s frequency that each host mutant is initialized with. Because CRISPR immunity is genetic, fitness is inherited from parent strains.
i_v	0.1 (.01–0.5)	Fraction of parent strain’s frequency that each viral mutant is initialized with.
G	3 (0–3)	Average of Poisson-distributed clearance-free emergence iterations given to each new host and viral mutant strain.
V_{min_freq}	10^{-6} (10^{-8} – 10^{-3})	Frequency threshold below which viral strains beyond their emergence iterations are cleared.
B_{min_freq}	10^{-6} (10^{-8} – 10^{-3})	Frequency threshold below which host strains beyond their emergence iterations are cleared.
V_{list_max}	300 (100–5000)	Maximum number of surviving viral strains beyond their emergence iterations.
B_{list_max}	300 (100–5000)	Maximum number of surviving host strains beyond their emergence iterations.

doi:10.1371/journal.pcbi.1002475.t001

Chapter 3

Metagenomic reconstructions of bacterial CRISPR loci constrain population histories

Abstract

Bacterial CRISPR-Cas systems can provide insight into recent population history because they rapidly incorporate, in a unidirectional manner, short genome fragments (spacers) from coexisting infective populations into host chromosomal loci. Immunity is achieved by sequence identity between transcripts of these spacers and their targets. Here, we developed new bioinformatics methods to analyze the type I-E CRISPR-Cas locus of *Leptospirillum* group II bacteria in biofilms sampled over five years from an acid mine drainage (AMD) system. Despite recovery of 452,686 spacers, rarefaction curves of distinct spacer types show no approach to saturation. The vast repertoire is attributed to retention of old spacers, despite rapid evolution of the targeted phage genome regions (proto-spacers). The oldest spacers are conserved for at least five years, and 12% of these retain perfect or near-perfect matches to proto-spacer targets. The majority of proto-spacer regions with a perfect or near-perfect match to a CRISPR spacer have the AAG proto-spacer adjacent motif (PAM). Spacers throughout the locus target the same phage population (AMDV1), but there are blocks of consecutive spacers without AMDV1 targets, and only newer spacers target the recent dominant plasmid populations. Results suggest long-term coexistence of *Leptospirillum* with AMDV1, periods when this phage was less evident, and time-varying plasmid population. Metagenomics can be applied to millions of coexisting cells in a single sample to provide a vast spacer inventory, allow identification of phage and plasmids, and enable analysis of previous phage and plasmid exposure. Thus, the approach can define aspects of the prior bacterial environment.

Introduction

The biology of natural ecosystems is shaped by interactions between microorganisms and their phage (Chibani-Chennoufi et al., 2004). However, cultivation has usually been required to determine phage host range and to study the interaction dynamics (Hyman and Abedon, 2010). Cultivation-independent genomic methods provide new approaches to these problems and can provide insight into the impacts of phage on population and community structures (Allen and Banfield, 2005). Genomic analysis can also elucidate the roles of phage and mobile elements in genome evolution (Allen et al., 2007). When applied to time series samples, these methods may also be able to constrain the rates of evolutionary processes (Denef and Banfield, 2012).

Many bacterial and archaeal genomes encode one or more CRISPR locus, named for the clustered regularly interspaced short palindromic repeats that separate spacer sequences that are transcribed and processed into small interfering RNAs (crRNAs) to confer immunity to phage, plasmids, and transposons [reviewed extensively in (Horvath and Barrangou, 2010; Karginov and Hannon, 2010; Marraffini and Sontheimer, 2010)]. New spacers are incorporated at the leader end of CRISPR loci (where transcription begins), whereas older spacers accumulate at the trailer end (Barrangou et al., 2007; Tyson and Banfield, 2008). The crRNA silencing requires identity with targeted sequences, and immunity may be lost by mutation in either the target region, or in an associated proto-spacer adjacent motif (PAM) that is apparently required for CRISPR function [reviewed in (Karginov and Hannon, 2010)]. While some mechanistic aspects remain unclear, cells that contain at least one CRISPR spacer that perfectly matches a region of the invading phage or plasmid with a flanking PAM will be immune (Deveau et al., 2008; Horvath and Barrangou, 2010). However, recently, studies in a certain type of CRISPR-Cas

system have shown that mutations in the proto-spacer, nearest the PAM, allows phage to escape while mutations in other regions of the proto-spacer have no impact on immunity (Semenova et al., 2011; Wiedenheft et al., 2011).

An important challenge in ecological studies is to detect and recover genome sequences from uncultivated phage and to link phage to their hosts. The CRISPR locus provides a means to address these two issues. First, spacer sequences extracted from CRISPR loci can be used to identify phage genome fragments in sequence datasets and initiate phage genome reconstruction (Andersson and Banfield, 2008). Second, assuming that hosts only incorporate spacer sequences from phage that infect them, CRISPR spacer sequences can be used to define host range (Andersson and Banfield, 2008). Community metagenomic datasets are a powerful way to approach these tasks because they simultaneously sample host CRISPR loci and the mobile elements they target.

Community metagenomic datasets provide inventories of spacer sequences for a population. Mapping these onto their targets can provide insight into the variety of phage and plasmids that target the host. Community metagenomic datasets also provide inventories of sequencing reads from the CRISPR locus. These can be used to compare the spacer complements of different host cells (e.g., to determine if different cells have the same immune potential). Importantly, the spacer complements of loci from coexisting individuals can be compared in a position-specific way to provide insight into population history. For example, conserved spacer sequences and spacer order in the older part of the locus may be interpreted to indicate origin and derivation from a common ancestral population or locus transfer (Tyson and Banfield, 2008), and transition from clonal to non-clonal loci has been suggested to indicate a recent selective sweep (Andersson and Banfield, 2008).

A limitation with metagenomic studies that target whole community DNA is that the sequencing is spread over entire genomes, so the number of reads recovered from a specific CRISPR locus may be insufficient to fully sample the spacer diversity of each host population. This problem is particularly significant if the loci are large and rapidly expanding. One approach to building a deeper inventory of CRISPR spacer sequences is to amplify the loci using PCR primers that target the repeat sequences (Pride et al., 2011). This relies upon knowledge of the repeat sequence, and can provide only very limited information about locus structure. Here, we combined metagenomic reconstruction with the PCR-based approach so as to take advantage of both methods. Using whole community metagenomic sequence, we reconstructed CRISPR locus geometry from natural populations of bacteria from the genus *Leptospirillum* and recovered sequences for their dominant phage. In addition, we used high-throughput sequencing to sample the spacer inventory of *Leptospirillum* deeply enough to assess population diversity and evaluate the phage/mobile elements they target. The analysis targeted biofilm samples collected over a five-year period. The results show that population-level analyses of CRISPR loci can provide insight into phage-host interaction dynamics and the recent history of bacteria in natural systems.

Results

We studied 9 microbial biofilm communities sampled from the air-water interface in the Richmond Mine (40° 40' 38.42" N and 122° 31' 19.90" W) at the 5way, A Drift (UBA), and C Drift (C75m) locations (Table 3.1 and Figure S3.1). Analysis included whole community genomic Sanger sequencing data from the 5way (March 2002) and UBA location (June and November 2005) [previously discussed in (Lo et al., 2007; Tyson and Banfield, 2008)]. Whole community genomic 454 FLX sequencing was applied to DNA extracted from five samples from

the C75 location (June 2006, August 2006, November 2006, May 2007, and August 2007) [previously discussed in (Denef and Banfield, 2012)].

Previous analysis of the *Leptospirillum* group II spacer complement indicated high levels of spacer diversity, especially in the 5way population (Tyson and Banfield, 2008), but sampling depth was insufficient to assess this inference in detail. Thus, for 5way (March 2002) and UBA (July 2005) samples, the entire *Leptospirillum* group II CRISPR locus was amplified with specific CRISPR primers and sequenced via 454 FLX (see Materials and Methods). From the community genomic datasets, we recovered spacers and associated Cas protein sequences from CRISPR regions in *Leptospirillum* group II and III genomes (Tyson et al., 2004; Simmons et al., 2008; Goltsman et al., 2009). The two closely related species of *Leptospirillum* group II (5way and UBA type) both have a single type I-E CRISPR system whereas *Leptospirillum* group III encodes two systems—one type I-E and one similar to type III (not discussed due to a lack of sufficient sequence information).

Spacers were extracted *in silico* from single amplicon and community genomic sequencing reads based on the detection of the *Leptospirillum* group II and III repeats (see Materials and Methods). The average G+C content of the *Leptospirillum* group II and III spacers is 55% and 56% respectively, similar to G+C content of the genomes (55% and 58%). The average length of *Leptospirillum* group II spacers is 32.7 ± 1.2 nucleotides while 33.1 ± 0.4 is the average for *Leptospirillum* group III spacers. We detected 452,686 total and 18,187 unique *Leptospirillum* group II spacer sequences, and 457 total and 318 unique *Leptospirillum* group III spacer sequences (Table 3.1). High error rates in the individual 454 sequencing reads inflate the unique spacers count. Thus, the spacer sequences across all datasets were clustered into groups using blastclust, with the parameters of 85% length and 90% identity within each group. We identified a total of 3,933 unique groups from *Leptospirillum* group II and 296 unique groups from *Leptospirillum* group III across all samples (Table 3.1).

For the two deeply sequenced *Leptospirillum* group II CRISPR amplicons datasets for the 5way and UBA samples, rarefaction curves were generated using spacer counts for each group found within each sample (Table S3.1). Both curves demonstrate no approach to saturation, despite deep sampling, implying a large diversity of spacers in each CRISPR locus (Figure 3.1A). Most of the unique groups occur in the dataset only once or twice, e.g., 35% of unique *Leptospirillum* group II groups are only found once across all datasets (Figure 3.1B). A few spacers occur over 1,000 times (Figure 3.1B).

We determined the order of spacers and thus, the local locus spacer arrangement, by identifying two to five sequential spacers in individual reads from the genomic and amplicon datasets for *Leptospirillum* group II and group III CRISPR loci (Material and Methods). By arraying overlapping patterns of spacers in different reads, we reconstructed the CRISPR loci geometry. The spacer sequence pattern was highly conserved at the end of the locus distant from the leader sequence (Figure 3.2).

The loci reconstructed from 454 data were compared to loci reconstructed from Sanger community genomic datasets from samples collected in 2002 and 2005 from the 5way and UBA locations, respectively (Tyson and Banfield, 2008). Figure 3.2 shows the trailer end of reconstructed *Leptospirillum* group II CRISPR loci is conserved in all samples collected between 2002 and 2007. Consequently, spacer abundance correlates strongly with spacer position in the locus (Figure S3.2). Within the shared block, excision events of a single or multiple spacers are evident. There are locus specific spacers found in the block of conserved shared spacers. These

spacers may have been present in ancestral strains, but lost from all but one locus through excision.

We detect variation in *Leptospirillum* group II in the complement of spacers in coexisting individuals as well as variation in spacer complement in samples collected at different times. Lineage variants are defined by the spacer content at the locus leader end. The CRISPR locus reconstructed from the 454-sequenced 5way sample shows evidence of excision events all along the locus when compared to the Sanger sequenced sample, though both samples were collected simultaneously. These may be real excisions, but the possibility that missing spacers are PCR artifacts cannot be ruled out, so this is a potential limitation of the amplification approach.

In the two UBA samples used for Sanger and 454 sequencing only collected a month apart, we recognized several CRISPR lineages (locus-specific series of spacers, colored in shades of blue). Interestingly, there are different patterns of spacer loss, as well as differences in sub-strain (defined by spacer content) abundance, between the samples. For example, the third UBA strain (pale blue in Figure 3.2) is essentially unrepresented in this sample collected one month later. We do not consider it likely that this is a PCR artifact.

Spacers in the trailer end are also generally conserved in reconstructed *Leptospirillum* group III CRISPR loci (Figure S3.3). Interesting, the number of spacers shared between CRISPR loci of *Leptospirillum* group III in the 5way (2002) and UBA (2005) samples is much lower than the number shared among the *Leptospirillum* group II loci across all time points (Figure 3.2).

To search for the PAM for *Leptospirillum* group II and III CRISPR loci, we compared short stretches of sequence immediately flanking the targeted proto-spacer region of non-CRISPR read sequences (likely phage, plasmids). Flanking sequences of proto-spacers that perfectly matched spacers were compared using WebLogo. For *Leptospirillum* group II, we detected a conserved tri-nucleotide ‘AAG’ immediately prior to the 5’ end of the proto-spacers. For *Leptospirillum* group III, we identified the conserved di-nucleotide ‘AA’, also at the 5’ end. We conclude that these are the PAM sequences required for CRISPR sampling and function.

We examined the frequency with which PAMs could be identified adjacent to proto-spacer regions targeted by spacers with perfect and imperfect matches. For *Leptospirillum* group II, we found that 76% of the 7,643 spacers with perfect matches and 62% of the 27,559 spacers with imperfect matches have a PAM (Figure 3.4). The percentages increase to 86% and 83%, respectively, with the allowance of one polymorphism in any position in the tri-nucleotide for *Leptospirillum* group II (Figure 3.4). For *Leptospirillum* group III CRISPR loci, 94% of the 1046 spacers with a perfect match and 85% of the 1079 spacers with imperfect matches have a PAM.

We evaluated host genome targeting by the *Leptospirillum* group II and group III spacers, considering both perfect (100% identity across entire length) and imperfect matches (90% identity over 85% length). We limited the analysis to host genome regions with the PAM. This analysis used all the spacers within a group, not a representative sequence for the group. *Leptospirillum* group III spacers had only one match (perfect) to the host genome, and this targeted an intergenic region. For *Leptospirillum* group II, the majority of genes targeted by spacers with PAMs are transposases, hypothetical genes, and other phage or plasmid genes (Table S3.2).

Overall, there are 6 genes in the *Leptospirillum* group II 5way-type genome and 26 genes in the UBA-type genome that have spacer matches (Table S3.2). Spacers derived from the *Leptospirillum* group II 5way-type more often exactly match genes and intergenic regions of the UBA-type genome (27 matches) than its own genome (4 matches) (Figure 3.4). Similarly, spacers derived from the UBA-type more often exactly match the 5way-type genome (33

matches) than itself (26 matches) (Figure 3.4). However, this trend is not seen with the imperfect matches to genes or intergenic regions (Figure 3.4). Notably, one spacer group (group3548) is responsible for 90% of all matches of UBA-type CRISPR spacers to intergenic regions in both genomes types.

It is likely reads with perfect or imperfect matches to spacers that are neither CRISPR nor host genome involve phage, plasmids, or other mobile elements. We examined spacer matches to all reads in this study without clustering spacers into groups because, although clustering removes common sequencing errors, it also hides real sequence variants. The results are summarized in Figure 3.5; for full details, refer to Table S3.4. For *Leptospirillum* group II, there were 35,564 matches (representing 7,659 unique spacers from 1,792 unique groups) to non-CRISPR, non-host genome read sequences. For *Leptospirillum* group III, there were 2,125 (representing 199 unique spacers from 188 unique groups) (Table S3.4). We categorized perfect and imperfect matches to non-CRISPR, non-host genome reads into four categories: perfect spacer matches with a PAM, imperfect spacer matches with a PAM, perfect spacer matches without a PAM, and imperfect spacer matches without a PAM (Figure 3.5).

For each *Leptospirillum* group II CRISPR locus in each sample, the relative abundance of spacer matches is fairly consistent across the different match categories. The same is true for *Leptospirillum* group III CRISPR loci, though the patterns in *Leptospirillum* group II and group III differ. Imperfect matches with a PAM represented the most abundant category for *Leptospirillum* group II whereas perfect matches with a PAM and imperfect matches with a PAM were the most abundant match types for *Leptospirillum* group III (Figure 3.5). Thus, the main difference between *Leptospirillum* group II and group III match types is the higher proportion of perfect matches with PAMs for *Leptospirillum* group III spacers. Notably, regardless of the spacer match type, there are consistently more matches with PAMs than without PAMs.

In order to determine the extent to which older spacers can silence phage and mobile elements, we tested for perfect and imperfect matches (black and grey boxes in Figure 3.2) as a function of spacer position within the CRISPR loci. In Figure 3.2, ‘imperfect’ matches’ include every match except perfect spacer matches with a PAM. In *Leptospirillum* group II and III loci (Figure 3.2), we found that shared conserved spacers (found in more than one time point in same locus location in both genome types) have either no match or imperfect matches, with one exception (found in 5way *Leptospirillum* group II locus). This exception involves a spacer that occurs in multiple different locus contexts (attributed to sampling of the proto-spacer region in independent events). In contrast, there are numerous matches of locus-specific spacers (shown as colored boxes) to putative mobile elements.

We also examined the relative abundance of the four different matches types as a function of spacer position within the *Leptospirillum* group II locus (Figure 3.6). The first panel shows match types regardless of locus position (Figure 3.6A), and is included for comparison with match types associated with the old (Figure 3.6B) and new (Figure 3.6B) end spacers. As noted above and in Figure 5, when including all spacers in the analysis regardless of their locus position, the most abundant category is an imperfect match with a PAM while the least common is a perfect spacer match without a PAM (Figure 3.6A). For the old end, which features spacers conserved across time, the most abundant type is imperfect matches without a PAM (Figure 3.6B). When only examining the spacers closer to the leader end (spacers not shown in Figure 3.2), the trend resembles that for all spacers (i.e., Figure 3.6C resembles Figure 3.6A). To highlight subtle differences between Figures 3.6A and 3.6C, we plotted the ratio of the

abundances (Figure 3.6D) and found a slightly elevated level of perfect matches with PAMs in the leader end. The relatively small degree of elevation in perfect matches with PAMs is somewhat surprising, because newer spacers are more likely to target co-existing phage and mobile elements.

The large number of spacer matches obtained for *Leptospirillum* group II (35,664) allowed us to test whether the relative frequency of mutations in the spacer sequences and PAMs is predicted by random mutation. If random, a simple expectation is that the ratio of mutation frequency in spacers vs. PAMs should be predicted by the ratio of the lengths of the spacer and PAM (for *Leptospirillum* group II, this ratio is 10.4). Across the entire *Leptospirillum* group II CRISPR locus (Figure 3.5A), spacers with mutations (and a perfect PAM) are 9.2 times more common than PAMs with mutations (associated with a perfect spacer).

To seek evidence for the persistence of the only well defined *Leptospirillum* group II phage, AMDV1, in the AMD ecosystem, we investigated the location of spacers within the CRISPR locus that target this phage population. The analysis included spacers with perfect and imperfect matches to phage reads. Across all time points, spacers with matches to AMDV1 occur, and are associated with spacers found throughout most of the loci (black boxes in Figure 3.7). However, the oldest spacers shared among all loci do not contain detectable matches to any sequences (Figure 3.2). Within the locus specific spacers (colored boxes in Figure 3.7), there are several blocks of spacers within the first and second UBA June 2005 as well as C75 composite loci that do not contain matches to phage AMDV1. Notably, only newer spacers (closer to the leader end and not depicted in any loci in Figure 3.2) have matches to a plasmid population.

Discussion and Conclusion

In this study, we analyzed the targeting of the phage and mobile element populations by CRISPR spacers from *Leptospirillum* bacteria over a five-year period. This enabled us to detect changes in immune potential and in the effectiveness of spacers, and provided insight into the usefulness of locus reconstruction for recovery of information about population history.

The trailer-end *Leptospirillum* group II spacers were largely conserved over the five-year study period. Given low observed rates of old end change over the study period, the oldest spacers in the earliest sample were probably incorporated long before our first sampling, so the locus could potentially record information about phage community composition for well over five years. Old end conservation might indicate shared ancestry, although locus lateral transfer can complicate this interpretation (Tyson and Banfield, 2008). Similarly, the spacer complement can distinguish populations sampled only months apart, though changes may be due to environmental proliferation of strains with different CRISPR loci, spacer addition, or both.

Most spacer diversity occurs at the leader end, as expected. Amplification and sequencing of the CRISPR region uncovered a vast variety of immune potential in one population. In fact, rarefaction curves derived from the recovered spacers show a lack of saturation, despite the unprecedented depth of sequencing. This finding provides support for prior speculation that, in some cases, most cells can contain different CRISPR loci (Tyson and Banfield, 2008). For loci with highly variable spacer complements, it may be inferred that the population has not experienced a strong bottleneck recently. In this circumstance, spacer sequences can be used to evaluate the diversity of coexisting phage and mobile elements. Two possible cases illustrate the potential utility of this approach for analysis of the recent growth environment of a host: Case one: a locus has spacers that target a single clonal phage; Case 2: a locus has spacers that target a diverse phage population, as well as many different phage and plasmid types. In the first

instance, we might infer recent growth in a simple environment, such as a laboratory culture, in the second we might infer growth in a more complex, diverse natural system. When attempting to recover information about the recent environment of an unknown bacterial strain population of medical or other significance, the effectiveness of CRISPR-based analyses will be higher if a large database of known phage types is available. Analysis of metagenomic sequences from coexisting phage can greatly augment this database.

Locus reconstruction provides a way of increasing the power of the CRISPR locus to provide information about recent population history. Specifically, if the spacer sequences can be classified into groups of new, older, and old (based on where they occur on the locus), the targets for each group could be evaluated separately. In the current study, we find that older (but not very old) and new spacers target essentially the same phage population, a result that points to the persistence of *Leptospirillum* in an environment with the same phage population over the time period represented by the locus (> 5 years). Notably, absence of targeting of AMDV1 by some mid-locus spacer blocks suggests short periods of fluctuation in phage exposure. Similarly, multiple strains (distinguished based on their CRISPR locus reconstruction) may record different exposure patterns. For example, the block of consecutive spacers in the UBA locus of sub-strain 2 without targets in our dataset, flanked by blocks with many targets, may record a period of time when that strain was exposed to a phage/mobile element pool not detected within the five years of our study. The targeting of plasmids (reconstructed from metagenomic datasets) only by new end spacers may indicate recent their immigration. Lack of AMDV1 targeting by the oldest spacers may be due to virus evolution rather than the absence of the ancestral AMDV1 population. Regardless of the explanation in this case, phage evolutionary rates rather than spacer retention timescales, may determine the timespan for useful CRISPR-based tracking.

Because we generated a large dataset of CRISPR spacers, we could evaluate factors that determine the total spacer pool. The relevant parameters are the diversity of the phage/mobile element target populations and constraints on phage regions that can serve as spacer sources (proto-spacers). The most important consideration is apparently the requirement for a PAM. For phage AMDV1, there are 1445 detected PAMs for *Leptospirillum* group II (implying 1445 potential spacer sequences). Despite this, we found 3,933 spacer groups. The great excess relative to the predicted spacer inventory, combined with the evidence of single mutations in PAM and spacer sequences, indicates that some spacer diversity is the result of resampling of rapidly evolving phage populations. Even after considering this effect, other targets (e.g., as yet undetected other phage and plasmids) are likely required to explain the size of the inventory.

A few CRISPR spacers have matches to the host genome sequence. Notably, these almost only target mobile elements integrated into the host genome, not core functional genes. Self-targeting (chromosomal proto-spacer sequences with PAMs), regardless of the target type, should be a problem for the host if, as expected, the Cas machinery targets DNA. It is possible that this finding indicates that the target in this system is RNA. Alternatively, the spacer and genomic target may not coexist in the same genome. This is plausible, because many comparative genomic studies of closely related strains have shown that gain and loss of mobile element genes is a major contributor to divergence of coexisting individuals (Allen and Banfield, 2005). This cannot be resolved in the current study because analyses involve short reads from innumerable coexisting individuals. However, if RNA targeting could be ruled out, apparent self-targeting or targeting of a gene recognized in other strains/species may indicate loss of the targeted gene (and the location of the spacer in the locus may distinguish recent from more gene ancient loss).

The correspondence between spacer sequences and their targets (or lack of it) can provide information about the factors that shape locus evolution. Generally, spacers in the ‘leader end’ exactly match co-existing targets while those in the ‘trailer end’ match imperfectly or even contain no detectable match at all. We infer that old end (inherited) spacers specific to only one population (e.g., at one time point) were lost from all other populations sampled at other time points. In other words, there is pressure to maintain useful old end spacers if the element it target is present, so detection of sample-specific old end spacer might imply the presence of the target only in that sample.

In addition to locus position-dependence of the degree to which spacers in a population match to co-existing targets, (Figure 3.6), there are differences in the likelihood that a spacer will target a proto-spacer region with a PAM. Interestingly, the spacer region tends to mutate before the PAM, with frequencies approximately as expected for random mutation, and only spacers close to the old typically lack matches to sites with PAMs. This suggests that, on average, spacers transition through the locus until the probability is high that they are ineffective. This, and the balance between the spacer addition rate and phage mutation rate, may be important determinants of locus length.

Currently, sequencing of CRISPR repeats and analysis of spacer order in isolates is used for strain tracking (Schouls et al., 2003; Grissa et al., 2008). The present study illustrates that community metagenomic sampling of natural communities of bacteria, phage, plasmids, and other mobile elements provides additional information. Specifically, comparison of loci in coexisting individuals provides insight into population diversity. In addition to uncovering evidence for recent bottlenecks, spacer inventory analysis can constrain the complexity of the current and past environments in which a bacterial population has grown. When applied to environmental populations, spacer matches to mobile elements provide insights into the recent history of exposure of host strains to the phage and plasmid pool.

Materials and Methods

The CRISPR-associated (*cas*) genes were retrieved from previously reconstructed genomes of *Leptospirillum* group II and III (Simmons et al., 2008; Tyson and Banfield, 2008; Goltsman et al., 2009). To determine the CRISPR-Cas type of these organisms as well as to rename the predicted Cas proteins via the new classification system (Makarova et al., 2011), we examined all predicted *cas* genes flanking the CRISPR loci. After transcription of the genes, the predicted proteins were searched to examine the family or superfamily of proteins to which they belonged. The identify of the *cas* genes and the structure of the *cas* suite helped to determine the exact type of CRISPR-Cas system (Makarova et al., 2011).

Extraction of community DNA and sequencing of biofilms sampled from the 5way (March 2002), UBA (June 2005 and November 2005), and C75 (June 2006, August 2006, November 2006, May 2007, and August 2007) locations in Richmond Mine, Iron Mountain, CA (Table 1 and Figure S1) have been previously described in (Tyson et al., 2004; Lo et al., 2007; Andersson and Banfield, 2008). For the 5way (March 2002) and UBA (July 2005) samples, community DNA was extracted with methods described in (Tyson et al., 2004). Primers were designed to target the entire CRISPR locus in 5way- and UBA-type *Leptospirillum* group II genomes: 5'-GCTCTTTCAGCCAAGATGGT-3' and 5'-TGGGGACCCTCCTTAGAAAT-3'. CRISPR loci were amplified with the *Leptospirillum* group II CRISPR locus specific primers using the Hot Start Herculanase (Stratagene). Agarose gel visualization of amplicons from both samples revealed a smear of differently sized fragments. Replicate PCR reactions were combined

for 454 GS FLX sequencing, which was completed by the Joint Genome Institute (Walnut Creek, CA).

In addition to the standard quality clipping of the 454 GS FLX read sequences performed by JGI, the SFF files were rescored using *sffrescore* (from the Genome Sequencer FLX System off-instrument software package) to generate the new phred-like quality scores. Because analysis of CRISPR spacers was conducted on the read level without assembly, extra filtering was performed to ensure good quality sequence. Reads containing at least one ambiguous base (“N”) were automatically removed (Huse et al., 2007). In addition, the ends of each reads were trimmed until a base passed 20/15 NQS (neighborhood quality standard) (Altshuler et al., 2000), with a variation described in Brockman et al 2008 (Brockman et al., 2008). The program *Cross_match* (developed by P. Green, University of Washington) was used to remove any remaining B adaptor sequences (from the 454 library construction process) from the trimmed reads.

Prior to *in silico* spacer extraction from sequencing reads, we further screened the community genomic data to remove reads that do not contain a CRISPR repeat. For individual Sanger reads, we required at least one instance of exact *Leptospirillum* group II (repeat) or group III (repeat) repeat sequence. For individual 454 reads, we required at least one instance of a *Leptospirillum* group II or group II repeat sequence, allowing for homopolymer errors in each position.

We developed a suite of tools for analyzing CRISPR sequences from both CRISPR amplicons and community genomic sequencing reads. Initially, sequencing reads are processed using software that recursively scans the read for a repeat sequence. Starting at the 5’ end of the sequence and working towards the 3’ end, the software looks for the repeat, and if found, breaks the sequence on the repeat, creating a pre and a post fragment. The pre fragment is analyzed to see if it is of minimum length to be a spacer sequence as well to determine if it contains a significant partial match to the repeat sequence. If it does, the pre fragment is inventoried as a “matching spacer.” Alternatively, if the pre fragment is too short or does not contain a partial repeat sequence, it is inventoried as an “unknown.” The post fragment is then treated as a new sequencing read and analyzed again, but without the partial repeat scanning (since the original sequence was broken on a legitimate repeat sequence). This recursive analysis continues until the input read has been completely processed. The last fragment of the process is treated in a similar manner as the initial fragment and scanned for partial repeats. If a partial repeat is found, the spacer is inventoried. Otherwise, if the last fragment is too small to analyze, it is treated as an “unknown.”

Every step in the process is inventoried in a mysql database. This allows us to reconstruct the processing events of every read as it gets analyzed by the repeat matching software. Additionally, it simplifies report generation for the final step in the analysis. After processing all the sequencing reads for an amplification experiment, a report is generated that contains a summary of the unique spacers and every configuration they are found in the data set.

To prevent an overestimation of CRISPR spacer diversity by accounting for 454 GS FLX sequencing read errors, the extracted spacers were grouped via *BLASTclust*, with parameters of 95% length overlap and 90% identity. Each group of spacers was considered one species and the abundance of each group was considered the abundance of each species, calculated from the total of the abundance of each spacer within the group. Rarefaction curves were created using *Analytic Rarefaction 1.3* (developed by S. M. Holland, University of Georgia; program freely available at <http://www.uga.edu/strata/software/>).

Spacer orders for each sequencing read (that contained at least two spacers) were obtained by listing the sequential order of each spacer (while ignoring the repeat sequence). The spacer order was then converted into a group order (all spacers were converted into groups), reducing the total overall amount of data used to assemble each CRISPR loci. The group orders were imported into Microsoft Excel and arrayed manually. Notably, the sequence of each spacer was listed as well to resolve any ambiguities.

Spacer matches against the host genome and non-CRISPR containing sequencing reads were determined by using *blastn* to detect perfect (100% match across 100% spacer length) and imperfect (90% match across 85% spacer length) nucleotide matches. For host genome matches, spacers were searched against the *Leptospirillum* group II and *Leptospirillum* group III genome sequences. For further analysis, only reads with proto-spacer sequences flanking an accurate PAM sequence (see below) were considered. For non-CRISPR containing sequencing reads, the spacers were searched against all community genomic read datasets listed in Table 3.1, with both *Leptospirillum* group II and *Leptospirillum* group III CRISPR reads removed.

PAM sequences for both *Leptospirillum* group II and *Leptospirillum* group III were identified by obtaining all the 5' and 3' flanking sequences flanking a proto-spacer sequence that matched perfectly to a spacer. These flanking sequences from both ends were then used as input for the program Weblogo to examine the frequency of each nucleotide for every position. Notably for *Leptospirillum* group II, the 5' flanking sequence contained three conserved nucleotides ('AAG') and for *Leptospirillum* group III, the 5' flanking sequence contained two conserved nucleotides ('AA'). Both PAM sequences were immediately flanking the proto-spacer sequence. For all perfect and imperfect spacer/proto-spacer matches, the two and three bases immediately upstream of each proto-spacer sequence was obtained in order to determine the frequency of accurate and inaccurate PAM sequences for *Leptospirillum* group II and *Leptospirillum* group III, respectively.

Figure 3.1. CRISPR spacer diversity in *Leptospirillum* group II. **A.** Rarefaction curve for spacer groups recovered from the 5way March 2002 samples (black line) and UBA July 2005 sample (grey line) datasets. Note that neither curve is approaching saturation, despite deep sampling. **B.** Rank abundance graph for the 5way CRISPR showing that only a few spacer groups were highly sampled (> 1000 counts).

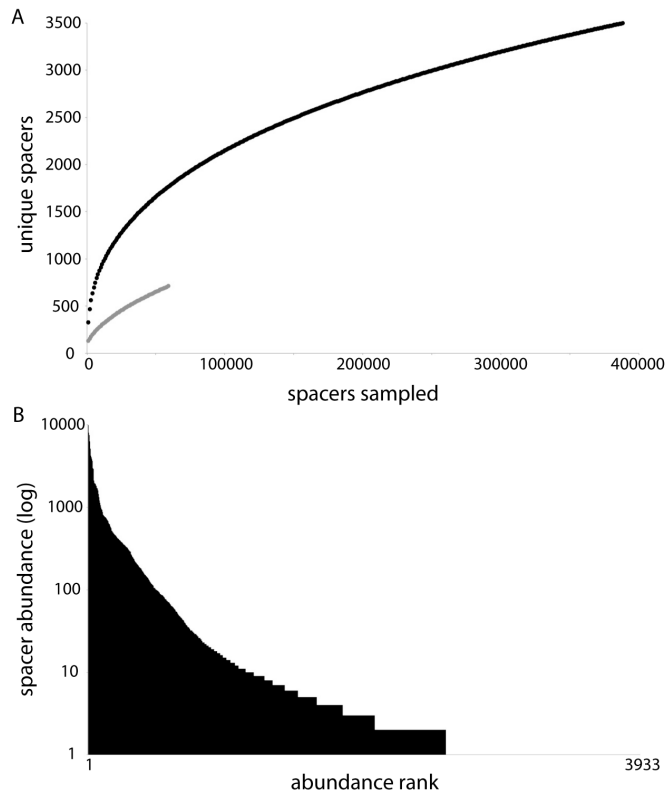


Figure 3.2. Reconstruction of *Leptospirillum* group II CRISPR loci from 5way, UBA, and C75 datasets. Loci are shown vertically from trailer to leader end, with spacers represented as wide rectangles. White rectangles represent spacers shared between at least two loci while colored rectangles represent spacers unique to a specific locus. Stripped lines in the loci show spacer loss. Note that the 5way loci are shown split in half due to space constraints. In the 8 columns left of each reconstructed locus, the placement of squares indicates the sample that contained the matching mobile element sequence. The 8 columns represent the following samples (from left to right): 5way-Mar 2002, UBA-Jun 2005, UBA-Nov 2005, C75-Jun 2006, C75-Aug 2006, C75-Nov 2006, C75-May 2007, and C75-Aug 2007. Perfect spacer matches with a PAM are shown as black squares while perfect spacer match with a PAM and imperfect spacer match with or without a PAM are shown as grey squares.

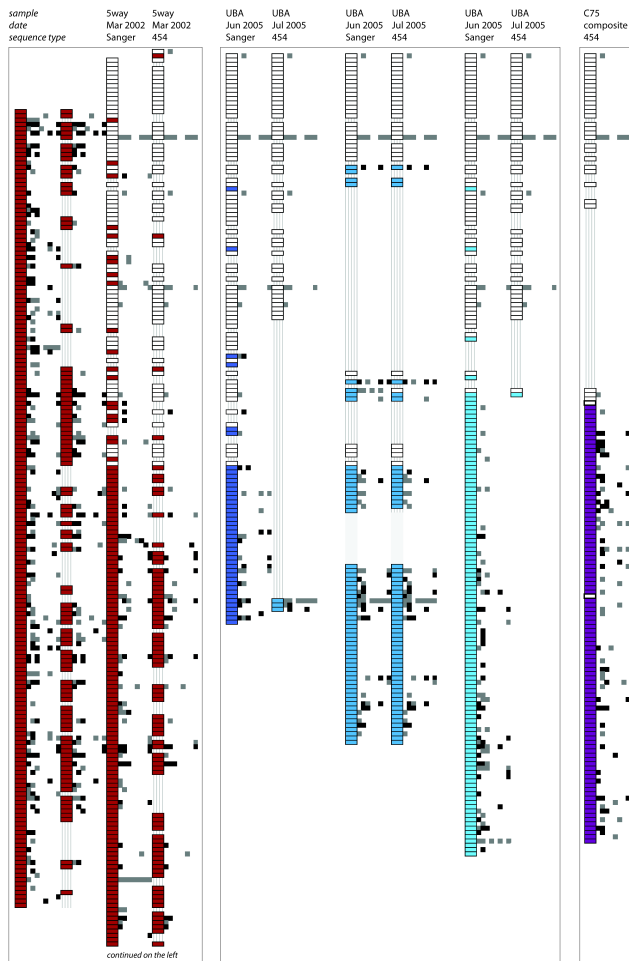


Figure 3.3. Frequency of tri-nucleotide sequences in the PAM position of proto-spacers found across all samples. Only the accurate *Leptospirillum* group II PAM sequence ('AAG') and imperfect PAMs (allowing for one polymorphism in any position) are shown. Relative abundances of perfect (black) and imperfect (grey) matches are shown.

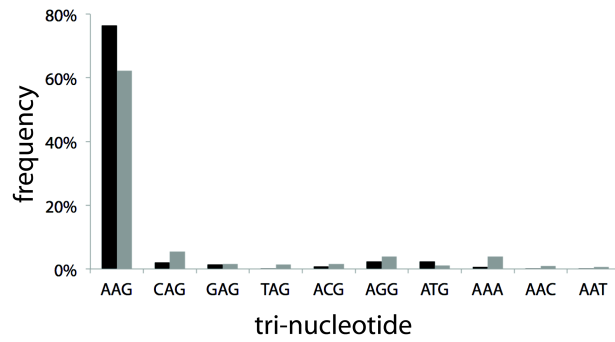


Figure 3.4. Abundance of *Leptospirillum* group II spacers from 5way and UBA samples matching to non-CRISPR host genomic regions. Perfect (solid) and imperfect (striped) spacer matches to intergenic and intragenic regions in the 5way type *Leptospirillum* group II genome (black) and in the UBA type *Leptospirillum* group II genome (grey).

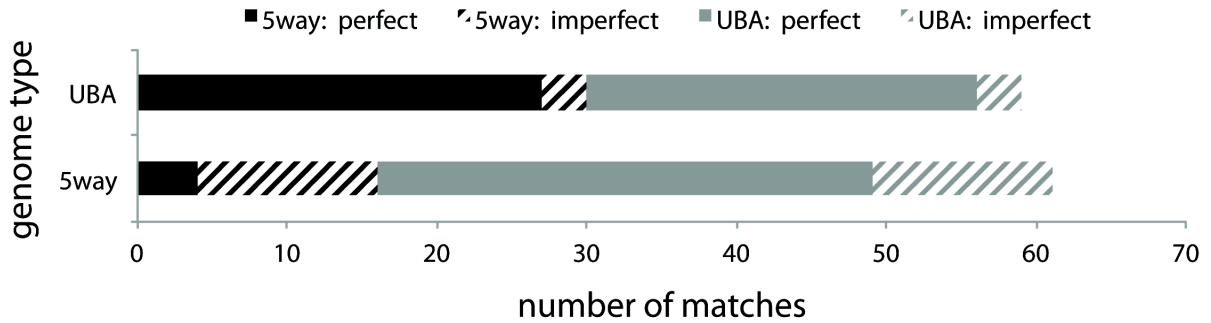


Figure 3.5. Summary of *Leptospirillum* group II and group III spacer matches to non-CRISPR, non-host genome reads across all datasets. Matches are separated into four categories, as listed in the legend.

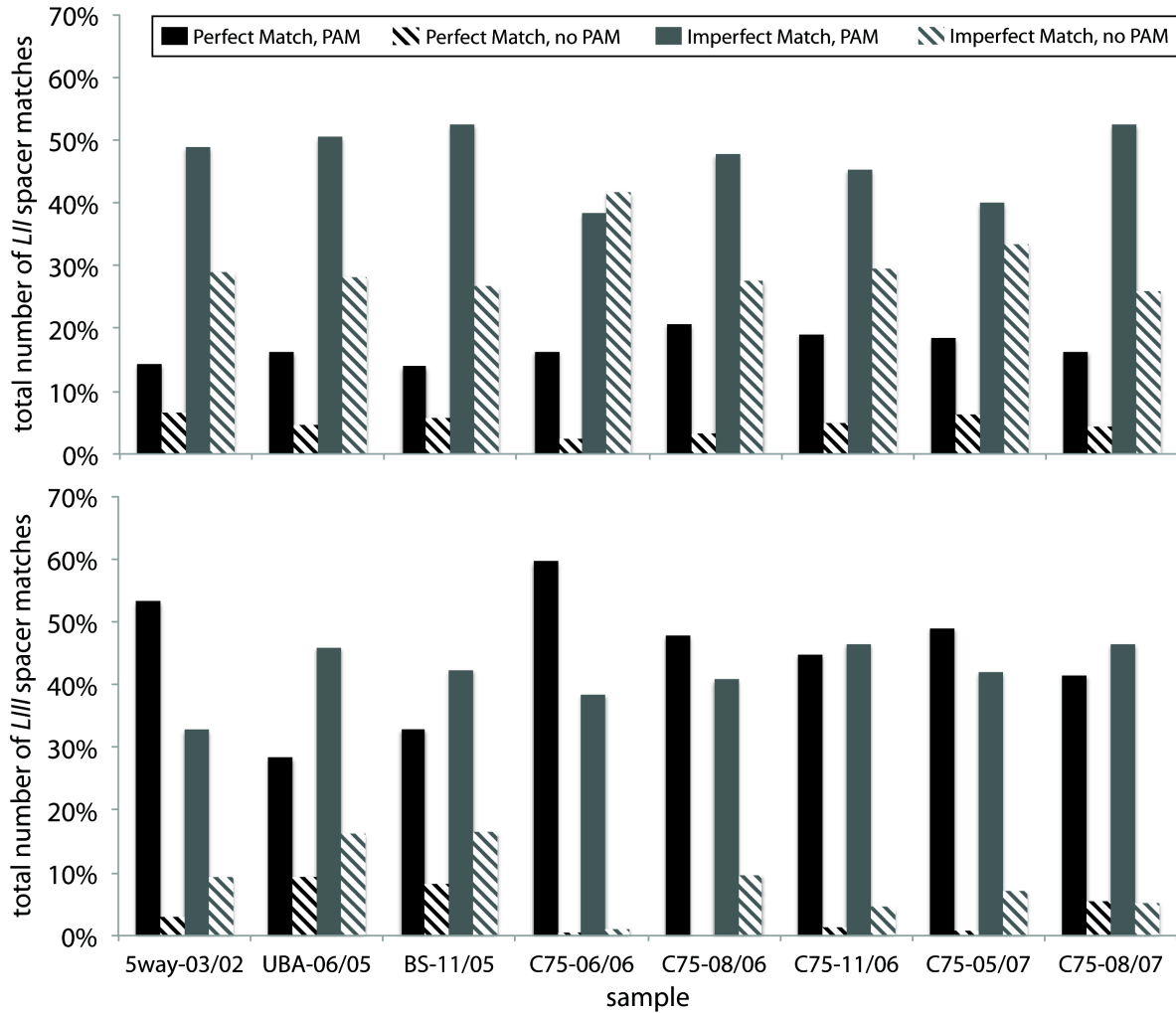


Figure 3.6. Different types of *Leptospirillum* group II spacer matches to all targets in the community genomic datasets (excluding CRISPR and the host genome). The four types of matches include: perfect spacer matches with a PAM, imperfect spacer matches with a PAM, perfect spacer matches without a PAM, and imperfect spacer matches without a PAM. **A.** Plot shows relative abundance of matches from all spacers. **B.** Plot shows relative abundance of matches from all spacers from the trailer end, limited to the region containing shared spacers (**Figure 2**). **C.** Plot shows relative abundance of all matches from spacers from the leader end, limited to the spacers not shown in **Figure 2**. **D.** The ratio of matches from spacers at the leader end relative to matches from all spacers.

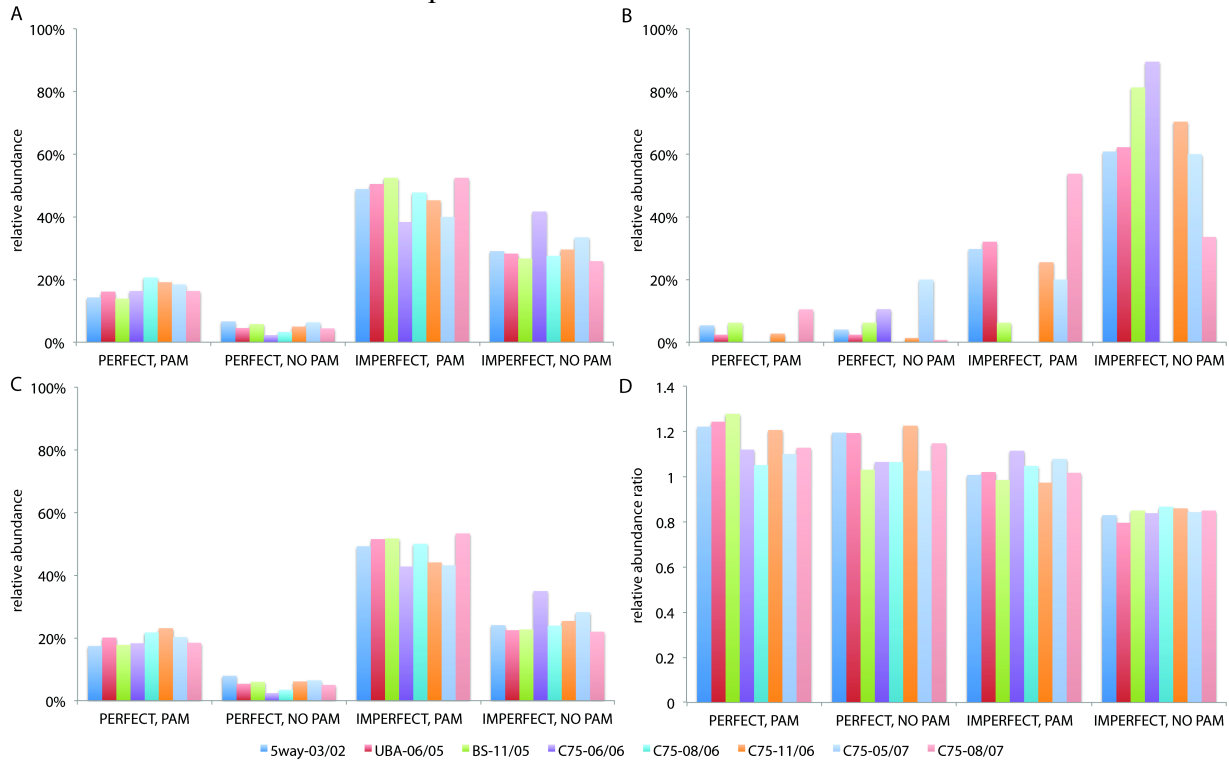


Figure 3.7. Spacers with matches to phage AMDV1 in *Leptospirillum* group II CRISPR loci from 5way, UBA, and C75 datasets. Reconstructed loci are represented in the same manner as in Figure 2. In the column left of each reconstructed locus, the placement of black squares indicates the spacer has a perfect or imperfect match to phage AMDV1.

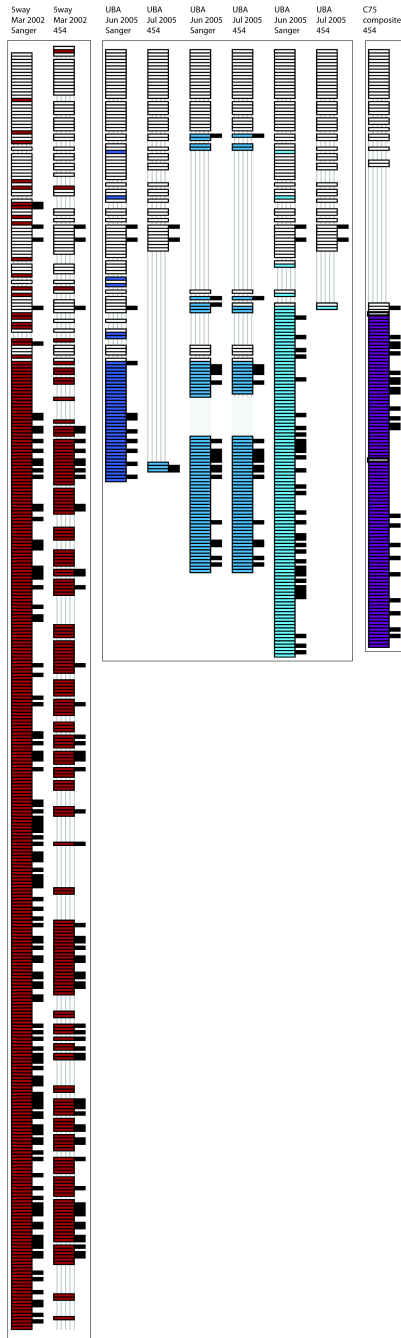


Table 3.1. Sampling conditions, sequencing information, and the number of CRISPR spacers and CRISPR spacer groups recovered from *Leptospirillum* groups II (*LII*) and III (*LIII*) for each AMD biofilm collected for this study. ‘Type’ refers to the source of the sequencing, either community genomic dataset (CG) or CRISPR amplicon dataset (PCR). In certain cases, spacers and spacer groups may be shared across multiple samples.

<i>Date</i>	<i>Location</i>	<i>pH</i>	<i>Temp.</i>	<i>Type</i>	<i>Platform</i>	<i>Seq. Data</i>	<i>LII Spacers</i>	<i>LII Groups</i>	<i>LIII Spacers</i>	<i>LIII Groups</i>
Mar 2002	5way	0.83	42.0° C	CG, PCR	Sanger, 454	136, 52 MB	389991	3516	157	91
Jun 2005	A Drift (UBA)	1.10	41.0° C	CG	Sanger	114 MB	1871	277	237	187
Jul 2005	A Drift (UBA)	1.23	38.0° C	PCR	454	13 MB	59764	716	n/a	n/a
Nov 2005	A Drift (UBA)	1.50	38.0° C	CG	Sanger	106 MB	41	41	48	44
Jun 2006	C Drift (C75)	0.70	43.0° C	CG	454	80 MB	143	75	0	0
Aug 2006	C Drift (C75)	1.00	43.0° C	CG	454	88 MB	79	61	0	0
Nov 2006	C Drift (C75)	1.18	42.7° C	CG	454	90 MB	261	115	8	8
May 2007	C Drift (C75)	1.17	44.4° C	CG	454	96 MB	304	100	7	7
Aug 2007	C Drift (C75)	1.12	40.2° C	CG	454	95 MB	232	99	0	0

Chapter 4

Stability and factors impacting phage and viral diversity over a multi-year period in a model microbial ecosystem

Abstract

Viruses of Bacteria and Archaea shape microbial ecosystem structure and dynamics and impact microbial evolution. However, much remains to be learned about the number of virus populations per bacterial or archaeal host, the timescales of variation in virus diversity, and the importance and function of CRISPR-based immunity in natural ecosystems. Here, we studied targeting of bacteriophage by CRISPR loci of bacterial *Leptospirillum* Group II populations and targeting of archaeal viruses by loci in E-plasma and G-plasma populations in a well-defined, low diversity acid mine drainage (AMD) ecosystem. Whole microbial community metagenomic data were collected over a span of about eight years, allowing investigation of virus sequence diversity, CRISPR spacer retention, and the linkages between them. Based on analysis of reconstructed non-clonal CRISPR loci in the bacterial population and time series information about phage diversity in the system, we infer a strong connection between CRISPR-based immunity of the host genotype and the coexisting phage strains. We documented persistence of phage populations over multiple years, and seasonal reproducibility in phage sequence type, suggesting host-driven viral selection. By analysis of E-plasma CRISPR spacer targeting of archaeal virus populations in samples collected two years apart, we detect persistence of virus populations, yet single nucleotide polymorphisms in many of the spacer target regions provide evidence for evolution driven by host immunity. Based on spacer matches to identified viral types, we estimate that E-plasma is the host for substantially more than six viral populations. However, this approach may underestimate virus diversity, as it requires matching of spacers to putative viral sequences. To assay overall virus diversity and to estimate of the fraction of viruses and phage targeted by CRISPR, we prepared and directly sequenced a concentrate of virus-like particles. The majority of putative viruses lacked corresponding spacer sequences, suggesting that they are silenced by other mechanisms. Our results indicate high virus richness despite low host species richness, persistence of populations over multiple years, and suggest the importance of additional mechanisms of phage and virus immunity.

Introduction

Bacteriophage and archaeal viruses (collectively referred to here as viruses) have the ability to alter the composition and function of bacterial and archaeal communities through predation and horizontal gene transfer, which drives genetic diversity (Fuhrman, 1999; Weinbauer and Rassoulzadegan, 2004; Pal et al., 2007). Despite their small size, viruses can have a far-reaching impact on global cycles in ecosystems (Zehr and Ward, 2002; Arrigo, 2005; Giovannoni and Vergin, 2012; Stocker, 2012) and on industrial applications such as food biotechnology (Leroy and De Vuyst, 2004), bioremediation (Furukawa, 2003), and bioleaching (Bosecker, 1997). Thus, it is critical to understand the dynamic interactions between microbial and viral populations. While the relationship between Bacteria and Archaea and their associated viruses has been well described in host-virus model systems in a laboratory environment [e.g., (Bohannan and Lenski, 2000; Buckling and Rainey, 2002; Lythgoe and Chao, 2003)], less is known about their intricate relationship within naturally occurring microbial communities due to the lack of knowledge about host range and to the complexity of most natural systems.

One layer of the intricate host-virus relationship includes the immunity of the host and the infectivity of the virus. Multiple mechanisms exist within host cells that allow for defense against foreign invaders [reviewed in (Labrie et al., 2010; Stern and Sorek, 2011)]. For example, some mechanisms operate prior to viral DNA injection into the cell, such as those preventing viral adsorption. Other systems function after DNA has been injected into the cell (Labrie et al., 2010). For example, restriction-modification systems target and degrade foreign DNA while abortive infection systems allows host cells to trigger cell death to avoid further infection of the community. Another type of immunity, conferred by the CRISPR-Cas system, allows host cells to evade viruses (and plasmids) via DNA or RNA silencing.

The CRISPR-Cas immune system is comprised of a CRISPR (clustered regularly interspaced short palindromic repeats) locus and CRISPR-associated (Cas) proteins [reviewed extensively in (Horvath and Barrangou, 2010; Karginov and Hannon, 2010; Marraffini and Sontheimer, 2010)]. The locus consists of repeat sequences interrupted by unique sequences called spacers that are derived from virus and plasmid genomes (Horvath and Barrangou, 2010). Proto-spacers are regions within the virus and plasmid genomes that became incorporated as spacers in CRISPR loci (Mojica et al., 2009). The entire CRISPR locus is transcribed from the leader end and processed into short crRNAs [reviewed in (Horvath and Barrangou, 2010)]. The crRNAs silence DNA or RNA of viruses and plasmids based on sequence identity. Flanking the loci, a suite of *cas* genes encode Cas proteins that aid in the procurement and integration of spacers into the CRISPR loci, the processing of crRNAs, and the conferment of immunity (Makarova et al., 2011). Notably, not every fragment of virus or plasmid sequences can be a proto-spacer. Each CRISPR-Cas system has a specific proto-spacer adjacent motif (PAM), which flanks and identifies proto-spacer sequences in viruses and plasmids (Mojica et al., 2009).

CRISPR-Cas systems are distinct from the other defense mechanisms because spacers provide a record regarding the identity of cellular invaders. Because spacers are added to the CRISPR locus, the locus records unsuccessful infection events by viruses and plasmids. Since the incorporation of new spacers from viruses and plasmids into CRISPRs occurs only at the leader end (Barrangou et al., 2007), CRISPRs can be interpreted as historical timelines of virus and plasmid exposure (see Chapter 3). It can also provide a method of linking a microbial host to its associated viruses and plasmids (Andersson and Banfield, 2008). Spacers that have sequence identity to co-existing viral or plasmid sequences typically are mostly located near the leader end of CRISPR loci have (Andersson and Banfield, 2008). Due to mutations that occur in virus or plasmid genomes over time that disrupt the sequence identity to spacers, and thus spacer effectiveness, the older spacers (at the trailer end) generally have few matches to targets (Andersson and Banfield, 2008; Tyson and Banfield, 2008), and ultimately, these spacers are lost. Thus, there is a delicate balance, as the hosts with active CRISPRs strive to acquire new spacers while the viruses evade via sequence mutations. There is also a balance between spacer loss (to relieve the burden of replication of unneeded genome sequence) and the risk associated with loss of immunity in the face of the return of an earlier present viral strain (see Chapter 2).

The majority of CRISPR research involves examining the mechanism of the CRISPR-Cas system in model organisms [e.g. (Haurwitz et al., 2010; Wiedenheft et al., 2011)]. While this has provided valuable knowledge regarding the biochemistry and function of the CRISPR-Cas system, it has left the area of how CRISPR loci function in nature largely understudied. For example, in the laboratory, bacteria and archaea are typically challenged with only a single virus or plasmid. However, evidence shows that CRISPR loci defend against multiple viruses and plasmids (Andersson and Banfield, 2008) as a single locus can contain spacers with different

target populations. In addition, current models of CRISPR-based virus-host interactions lack information about natural viral population diversity, the rate of evolution of CRISPR immune systems driven by the virus-host “arms race”, and the resulting population structures over space and time.

In natural systems, levels of archaeal virus and bacteriophage diversity, timescales of variation, and factors that impact diversity levels, are difficult to assess. Ecosystems with limited species richness are an obvious target for such investigations. Of particular value are systems in which CRISPR loci play a prominent role in virus/phage immunity because CRISPR spacer sequences, which confer immunity, enable linking of virus/phage to their host populations. For these reasons, we sampled biofilm communities from the acid mine drainage (AMD) system at Iron Mountain, CA, USA. In this relatively low diversity ecosystem, almost all bacterial and archaeal host populations have at least one CRISPR locus and seasonal fluctuations in community composition occur in response to geochemical changes. Whole microbial community metagenomic data were collected over a span of about eight years, allowing investigation of virus/phage sequence diversity, CRISPR spacer retention, and the linkages between them.

Results

We studied microbial communities in acid mine drainage (AMD) biofilms that grow at the air-solution interface in the Richmond Mine (40° 40' 38.42" N and 122° 31' 19.90" W). Biofilms were sampled at different locations within the 5way, and A, B and C Drifts between 2002 and 2010 (Table 4.1). Assembly of the community genomic sequence data allowed reconstruction of virus/phage and microbial genomes, including the CRISPR loci. Using spacer sequences, spacer targets, inferred to be phage, viral, or plasmid sequences, were identified (Table S4.1). Previously undetected phage and archaeal viruses (viruses) were identified from a database of contigs derived from an assembly of multiple samples sequenced with Illumina technology and a database of reads from a virus concentrate generated by multiple density gradients (Table 4.1). These putative viruses and plasmids were detected via searching for sequences with perfect and imperfect matches to CRISPR spacers from bacteria (*Leptospirillum* group II and group III) and archaea (G-plasma and E-plasma). Using data from a combination of sequencing technologies (Table 4.1), we identified *Leptospirillum* group II and E-plasma CRISPR spacers with targets (either putative or known viruses) and determined their location within the CRISPR loci structure. Additionally, sequenced amplicons of a region within the bacteriophage AMDV1 genome provided information on the spatial and temporal diversity of AMDV1.

To investigate the timescale over which spacers become ineffective and the pattern of spacer loss, we studied the dynamics and structure of *Leptospirillum* group II CRISPR loci in biofilms collected from the C75 location over 5 years: June 2006, August 2006, August 2007, November 2007, June 2008 (two spatial samples), September 2010, November 2010, and December 2010 (Table 4.1, Figure 4.1). Briefly, CRISPR spacer identity and order within the loci were determined by identifying the sequencing reads containing CRISPR repeats and assembling these reads into a dominant path with Velvet (see Materials and Methods). By searching all recovered spacers against all non-CRISPR containing reads from each sample, we detected all possible spacer targets (phage and mobile elements).

Genome analyses of *Leptospirillum* group II show specific genotypes (that differ by only by a handful of SNPs across the entire genome) dominate at different time points (Denef and Banfield 2012). When comparing the CRISPR locus type to the dominant *Leptospirillum* group

II within a given sample, the same *Leptospirillum* Group II genotypes had the same dominant CRISPR loci. Comparison of the CRISPR loci from all 11 samples reveals a similar structure, consisting of 4 main blocks of spacers that are either shared or missing amongst the samples (Figure 4.1). Not surprisingly, CRISPR loci share large blocks of spacers at the trailer end in all samples, whereas the leader ends have unique blocks of spacers (Figure 4.1). There is evidence for at least two excision events within the loci that removed large blocks of spacers but also two genes (Figure 4.1). One of these genes has similarity to a hypothetical gene while the other is annotated as a transposon.

Further analyses of spacer targets were completed using the spacers recovered from CRISPR loci reconstructed from Illumina sequencing reads (biofilm sampled between November 2007 and December 2010). Targets were successfully identified for 59% of the *Leptospirillum* group II spacer sequences. Some CRISPR spacers are inferred to remain effective, based on perfect sequence identity, across three years. Interesting, some of the oldest spacers (deep red in Figure 4.2) still have matches to co-existing target sequences (i.e., viruses). There is evidence for the disappearance and reappearance of virus populations targeted by CRISPR spacers, as their sequences were only detected at specific times (Figure 4.2). Interestingly, disappearance of certain blocks of spacers would have had no impact on immunity (those targeted sequences do not co-exist in the sample with missing spacers). For example, the orange and green blocks of spacers only matched targets (including viruses) in earlier collected samples. Given that the amount of sequencing is finite and unequal across samples, the lack of a targeted sequence indicates that the strain is, at best, at low abundance in the sample.

To further evaluate recombination and variation in the phage AMDV1, we amplified a ~650 nucleotide region of the genome using primers targeting relatively low variation regions. Primers were applied to 9 biofilm samples collected in November 2007, June 2008, October 2008, and June 2009 (4 spatially resolved samples) in A Drift and in February 2008 (2 spatially resolved samples) in B Drift (Table 4.1). The translated sequence of the targeted region has similarities phage tail sheath proteins. The amplified material was sequenced with Sanger technology and the resulting Sanger sequences were aligned and compared using the Strainer program (Figure 4.3). The alignment shows conservation of small sequence polymorphism patterns across space and over a three-year period, as well as linkage patterns consistent with homologous recombination amongst closely related genotypes.

In order to better evaluate the virus genotypic variation over time for the AMDV1 populations, we compared the pools of all potential *Leptospirillum* group II spacer sequences (i.e., the set of predicted spacers, based on the AMDV1 sequences). This required identification of the set of all possible proto-spacer sequences, predicted based on the AMDV1 genome sequence, the previously identified PAM sequence (AAG) and the average spacer length for the *Leptospirillum* group II CRISPR locus (33 nucleotides). Briefly, all PAMs were identified in the AMDV1 amplicon sequences and the potential spacers (proto-spacers) flanking the PAMs were extracted *in silico* (see Materials and Methods). In most cases, the positions of PAMs on the reads were identical between sequences. However, in certain cases, there was either a missing or additional PAM sequence due to viral genome polymorphisms, which resulted in either a missing or additional spacer, respectively. There were also differences in the spacers themselves. In fact, in a single potential proto-spacer region, there were up to 12 different spacer sequences, usually only differing by one single nucleotide polymorphism. Thus, differences in the pool of potential spacers from each sequence result from mutations either in the PAM or in the proto-spacer sequence.

The complete set of predicted *Leptospirillum* group II spacers was evaluated against the AMDV1 amplicons sequences to determine the fraction of individual viral particles with sequences with an exact match to a predicted spacer and an intact PAM in each sample. These individual virus particles should be unable to infect the *Leptospirillum* group II host. The fractions of matches per spacer type were used in the hierarchical clustering of the 9 samples to determine how changes in the PAM or proto-spacer sequence affects CRISPR immunity. This allowed us to compare infectivity profiles for samples separated by time and space. The spatial samples from A and B Drift cluster closely (Figure 4.4); two samples from A Drift (November 2007 and October 2008) cluster together and cluster somewhat closer to the B Drift time samples, rather than the samples collected from the same site at a different time of year (summer samples; Figure 4.4). From this analysis, it is apparent that the infectiveness of AMDV1 populations is more similar for samples collected in the same season than for samples collected at the same site at different times. A potential explanation for this is seasonally-driven changes in host immunity profiles.

Illumina datasets of AMD communities (Table 4.1) greatly augmented the previously reported group of available viral sequences (Andersson and Banfield 2008) that were assembled only from datasets sequenced with Sanger and 454 technologies and produced more contig sequences than previously retrieved from single Sanger or 454 dataset alone. A total of 20,147 CRISPR spacers, recovered from 2 CRISPR loci in E-plasma, 2 CRISPR loci in G-plasma, 1 CRISPR locus in *Leptospirillum* group II, and 1 CRISPR locus from *Leptospirillum* group III (Table S4.2), were searched against all contigs derived from the Illumina AMD datasets to identify a more complete set of phage, virus, and plasmid sequences in the AMD system.

We recovered 1,820 contigs from the new assembly that have a match to at least one CRISPR spacer, roughly 8% of total assembled contigs (23,265 assembled contigs) (Table S4.1). The average contig with a CRISPR spacer match size is ~18 kb, and the average read depth for each contig is 128,000x. Spacer matches to prophage or plasmid-integrated regions found within bacterial and archaeal genomes likely increased the average contig size. When the 1,820 contigs with spacer matches were searched against known AMD viruses, plasmids, and microbial hosts at the nucleotide and translated nucleotide level, we found that the majority of sequences were identical or similar to previously detected viruses and plasmids. 952 contigs have low or no detectable matches to known viruses and plasmids, and are likely novel viruses or plasmids not previously found due to low abundance, low sequence coverage, or spatial and seasonal variations. 1.3% of the total number of reads used in the assembly belong to one of these 952 virus- or plasmid-like contigs. This finding indicates that the viruses and plasmids comprise of at least 1.3% of the community genomic DNA, which is most likely an underestimation as viral genomes are significantly smaller than microbial and fungal genomes.

Open reading frames determined for each of the 952 candidate virus/plasmid contigs and the resulting 3,260 predicted proteins were searched against InterPro Scan to detect any matches to protein signatures and to evaluate support for viral or plasmid origin. Of the 3,260 predicted proteins, we found matches for 968 predicted proteins (Table S4.3), many of which were similar to viral- or plasmid-like proteins. Included in this list are structural proteins, such as baseplate assembly, hexon, tail tube, and tail sheath as well as DNA binding and packaging proteins, such as primases, helicases, DNA breaking-rejoining, delivery proteins, polymerases, and intergrases (Table S4.3). Also, there were matches to proteins found in bacteriophages and archaeal viruses, such as Acidianus two tailed virus, bacteriophage A500, HK97, PRD1, T4, phiE125, Sulfolobus virus STSV, and Ralstonia phage RSS1 (Table S4.3). In addition, plasmid proteins were detected

in high abundance, such as CopG, TraR, ParB-like nuclease, plasmid recombination, plasmid replicase, and conjugation related proteins (Table S4.3). Other common plasmid encoded proteins were detected, including those that involve antibiotic resistance (such as bacilysin, arsenic, and penicillin), restriction-modification, methylation, and toxin/anti-toxin systems (Table S4.3). These predicted proteins suggest spacer sequences enabled an enrichment or selection of viral- and plasmid-like sequences.

Notably, there is a putative phage sequence, referred to here as AMDV10 (scaffold_1127), that has 142 and 7 exact matches to *Leptospirillum* group II and group III CRISPR spacers, respectively, throughout the 14 kb contig (Table S4.1). It is the largest contig retrieved that does not have any detected matches any known host or virus in the AMD system. Surprisingly, it had not been previously detected in any individual dataset and it has little or no similarity to the known *Leptospirillum* group II and group II bacteriophage, AMDV1. The majority of the 27 predicted genes were unknown, but a few were similar to DNA delivery, DNA-directed polymerases, virus hexons, and bacteriophage PRD1, P3 (Table S4.3). These viral-like genes, in conjunction with the presence of *Leptospirillum* group II and group III CRISPR matches, support the hypothesis that AMDV10 is a newly described virus detected that targets these bacterial populations.

We examined the location of CRISPR spacers with matches to AMDV10 within the CRISPR locus to infer details of the history of this virus population history (Figure 4.5). Interestingly, the virus is predominately targeted by 5way *Leptospirillum* group II CRISPR spacers. The shared older spacers have no targets to AMDV10 and only certain small blocks of spacers have targets in the 5way locus towards the leader end (Figure 4.5).

Another novel sequence, referred to here as AMDV11 (scaffold_1537), has 7 and 6 exact matches to *Leptospirillum* group II and group III CRISPR spacers, respectively, throughout the 15 kb sequence (Table S4.3). While some fragments are similar to known AMD viruses, this single contiguous piece is has not been previously detected. Although it is generally difficult to determine if certain phage or viruses are lytic or lysogenic, the presence of predicted proteins on AMDV11 with similarity to Cro and lambda repressor (responsible for regulating the transition between lytic and lysogenic life stages) suggests that this is a lysogenic virus.

In order to augment the inventory of viruses and mobile elements identified based on CRISPR targeting, we developed a protocol in order to concentrate virus-like particles (see Materials and Methods). Briefly, multiple biofilm samples from the UBA location (BS type biofilm) were collected in November 2007 and concentrated with PEG. The concentrate was put through a sucrose gradient and the fractions containing the most viral like particles (as seen under SYBR gold staining) were further purified through a sucrose cushion. DNA was extracted from the resulting concentrate and whole genome amplified due to minimal genetic material prior to 454 sequencing. We utilized electron microscopy to view representative morphologies of VLPs (Figure 4.6), which revealed a variety of structures that support presence of a wide diversity of viral genomes in DNA extracted from the virus concentrate.

After filtering, sequencing of the viral concentrate produced 442,940 contigs, with an average read length of 289 nucleotides (a total of ~129 MB of read sequences). We detected only 19 reads with similarity to 16S rRNA and only 1 read with similarity to 18S rRNA, indicating only a very low level of microbial or fungal contamination of the sample. A search of the translated reads against known AMD proteins reveals that only a minor subset of sequences has similarity to bacterial and archaeal sequences (Figure 4.7a). Further, only 6% of sequences have matches to host microbial genomes. These findings support the notion that the concentrate is

highly enriched for VLPs. We detected a protein sequence match (blastx) for roughly 50% of all reads. Of the reads with sequence matches to proteins previously identified in genome sequencing of AMD biofilms, about 60% have matches to proteins on fragments not assigned to microbial genomes (potentially viruses/plasmids). Only 25% of reads have matches to previously reported viral sequences (Andersson and Banfield, 2008) (Figure 4.7b). The majority of reads with matches to AMDV viruses have similarity to known archaeal viruses (Figure 4.7b).

Both population-level variation and artifacts arising from whole genome amplification largely precluded the assembly of sequences from the virus concentrate sample. The 14 contigs that were assembled (Figure S4.1) do not have high similarities to any sequences in the NCBI database, but two have some similarities to plasmid proteins in the ACLAME database, which contain phage, viruses, prophage, and plasmids sequences.

Due to the low level of assembly achieved for the concentrate sample, our analyses primarily used unassembled reads. Of the total 443,940 reads from the virus concentrate dataset, 26% (115,845) have matches to a known AMD virus sequence. Of the remaining sequences, 11% (35,660) have matches to a sequence in the ACLAME database. We then searched sequences with any known virus matches (292,435) for matches to spacers from *Leptospirillum* group II, *Leptospirillum* group III, E-plasma, and G-plasma (Table S4.2). From the 292,425 sequences without similarity to known AMD viruses or sequences in the ACLAME database, 7% (21,640) have matches to at least one spacer (Table S4.4). After filtering sequences that match AMD host organisms (26,098) or 16S rRNA (5), there are 244,692 sequences that are unknown, or roughly 55% of the total dataset. Interestingly these 244,692 can be clustered into 23,808 non-redundant sequences (at 90% identity). This finding indicates that the same genomes were sequenced multiple times, and rules out the explanation that lack of assembly was due to very low abundance levels. Searches of these sequences against the nr database produced only 10,755 with matches. Thus, after all searches, roughly 53% of the total dataset have detectable matches.

We also investigated CRISPR-based virus-host interactions in Archaea. The archaeal E-plasma genome contains two different CRISPR loci (first identified in Andersson et al 2008). Spacers were extracted from both loci and the loci assembled using the spacer order on the Sanger reads (Table 4.1, Figure 4.8, Figure 4.9). Of the 538 total spacers recovered from E-plasma, 48% have a perfect or imperfect match to a known archaeal virus. There is a clear transition through the locus from spacers without matches, spacers with inexact matches, and spacers with perfect matches to AMDV2, especially with E-plasma CRISPR #1 (Figure 4.8 and Figure 4.9). We also considered the distribution of spacers in the E-plasma locus with matches to other viral sequences (Figure 4.10). Spacers throughout both loci target AMDV2 and AMDV4; a few spacers target AMDV3, and other viruses have significantly fewer matches.

In order to evaluate the magnitude of sequence variation in the AMDV2 population, we mapped reads from six different 454-sequenced datasets to the previously identified AMDV2 consensus sequence [constructed from Sanger sequences derived from UBA, and UBABS samples, (Andersson and Banfield 2008)]. The six datasets included the virus concentrate from UBA-BS and five samples in a time series sampled from June 2006 to August 2007 at C75 (Table 4.1). AMDV2 is only present in the virus concentrate (at ~120x depth) and the August 2007 sample (at ~5x depth). Interestingly, some SNPs in the August 2007 sample were also found in sequences in the virus concentrate (Figure 4.11).

To examine how polymorphisms in genomes of members of the AMDV2 virus population impact CRISPR spacer immunity over time, we determined the effectiveness of E-plasma spacers (via sequence identity) against AMDV2 sequences retrieved from 2005 and

2007. As a result of sequence polymorphisms that distinguish the 2005 and 2007 sequences, the E-plasma populations had slightly different CRISPR immunity levels. Notably, the overall pool of spacers with matches to AMDV2 remained constant between time points. Between 2005 and 2007, mutations arose in detectable levels in AMDV2, and these caused shifts in the type of matches (perfect and imperfect). For example, 4 spacers from E-plasma CRISPR #1 shifted from having perfect matches to imperfect matches while 11 spacers had the opposite transition. In the E-plasma CRISPR #2, 4 spacers changed from having perfect matches to imperfect matches while 7 spacers experienced the opposite change. Overall, the frequency of spacers with perfect matches to AMDV2 sequences increased for both loci from 2005 to 2007. Table S4.5 shows blast search results, which reveal that spacers have perfect matches in one sample but not in the other, either due to new mutations or shifts of different viral strains over time.

Discussion and Conclusion

In this study, we investigated virus population structure and variation in the complement of CRISPR immune potential in the associated host populations over time in a model natural system, where relatively deep, comprehensive analysis is possible due to low diversity microbial communities. The analyses also uncovered sequences of novel phage and viruses, which have the potential to shape host population structure and function, including biomass and carbon turnover in the system. A particular advantage of the studied system is that it was possible to examine the viral populations and CRISPR loci derived from biofilm communities in different locations within the AMD system across a multi-year period. This enabled investigation of the composition and structure (organization of spacers, linked to their targeted virus, with information about effectiveness due to spacer sequence identity) of CRISPR loci, CRISPR-mediated host-virus dynamics, and the diversity of viruses to reveal insights into the stability of virus populations over time as well as the role and importance of CRISPR-Cas immune systems in natural ecosystems.

We interpret the spacer excision events in CRISPR loci found in *Leptospirillum* Group II as evidence for the selection of hosts with specific CRISPR spacer content based on the presence or absence of spacers and coexisting viral or plasmid sequence targets (Figure 4.1). The correspondence between loss of immunity to specific virus sequence types and disappearance of that virus suggest host selection based by virus immunity (Figure 4.2). While fluctuations in the presence or absence of spacer targets across the sampling period may result from differences in sequence depth, these differences more probably result from variations in abundances and diversity of the targeted viruses and plasmids. Interestingly, some trailer (old) end spacers seem to retain effectiveness (Figure 4.2), supporting the hypothesis generated by the model discussed in Chapter 2 that older spacers are retained as viruses that lose infectivity can remain at low levels in the system, and potentially re-infect if immunity is lost.

The interaction dynamics of E-plasma and its various viruses was analyzed using the two CRISPR loci of E-plasma. The spacer content is highly variable within the archaeal population (the loci is highly non-clonal), suggesting rapid host-virus co-evolution. In previous work, we determined that archaeal virus AMDV2 was capable of E-plasma infection based on E-plasma CRISPR spacer matches to sequences from the AMDV2 genome (Andersson and Banfield, 2008). The comparison of the two E-plasma loci from 2005 and 2007 shows that AMDV2 was fairly consistent in sequence variation levels, without any drastic changes in sequence type (the main differences in CRISPR targeting result from shifts between perfect matches and imperfect matches, a consequence of SNPs) rather than a complete change in spacers with targets). If

AMDV2 sequences varied drastically, the pool of spacers with targets would be extremely different between time points.

Tracking over time and space of *Leptospirillum* Group II's AMDV1 phage population using PCR-based virus sequencing (Figure 4.3) indicates general stability in virus genome sequence over a multi-year period, although spatial variation in the form of single nucleotide polymorphisms implies some variation in host strains targeted (Figure 4.4). The high sequence diversity of proto-spacers with the same associated PAM indicates rapid phage mutation to escape CRISPR targeting. The findings do not indicate rapid selection events (bottlenecks), as seen in Chapter 1, when a phage for which the host has no immunity was introduced in single isolate system. Rather, the data support a model in which host and virus genetic "clouds" (with differing resistance/infectivity levels) shift in composition over time/space (Banfield and Young 2009). Notably, disappearance and appearance of PAMs in the AMDV1 amplicon sequences due to SNPs indicates the 2-year timescale over which proto-spacer types are created or destroyed in the virus genomes.

The history of a host population (e.g., CRISPR locus clonality or clonal to non-clonal regions) and its past exposure to virus and plasmids (location of spacers targeting specific viruses/plasmids in the locus) can be investigated via CRISPR analysis (Chapter 3). Notably, *Leptospirillum* group II spacers with matches to AMDV10 were not found throughout the locus, and are absent near the trailer end, suggesting that the AMDV10 population did not consistently co-exist with *Leptospirillum* group II (at least not in high abundance). The AMDV10 population may be transient. Similarly, for archaea, multiple viruses (AMDV2, AMDV3, AMDV4, AMDV6, AMDV7, AMDV8), have E-plasma in their host range based on spacer matches yet the pattern of spacers suggests that while AMDV2 and AMDV4 have long periods of co-existence with E-plasma, the other archaeal viruses come and go.

When all E-plasma CRISPR spacer matches were considered against known AMD viruses (Figure 4.8 – 4.10), 52% spacers lack identified targets. The existence of spacers without exact or inexact matches implies other viral or plasmid types are capable of infection, but not yet identified. Overall, we identified for six different virus types that replicate in E-plasma. Considering the spacers lacking identified targets, this archaeon may be the host for many more than six different virus types. While there are estimates to the number of virus particles to host cells (10:1) (Weinbauer, 2004), there is little or no prior information regarding the number of virus types capable of infecting a single host strain. For *Leptospirillum* Group II, we identified three distinct virus populations, with the expectation that it hosts many more, given the absence of matches for many spacers (Chapter 3, Figure 4.2 and 4.5).

Previously known plasmid and viral sequences in AMD biofilms (Andersson and Banfield 2008) were likely only detectable because the plasmids and viruses were present within host cells, since the extraction process was optimized for retrieval of microbial DNA. Small virus-like particles (VLP) would most likely be removed along with the supernatant after the first round of centrifugation, thus likely would not be detected in sequencing. The viral concentrate studied here provided more comprehensive insight into the diversity of plasmids and viruses in the system. In the virus concentrate, sequences not targeted by CRISPR appear to largely derive from viruses and plasmids, a finding that indicates reliance of other viral defense mechanisms for certain viral types. Only one tenth of the concentrate sequences that were targeted by spacers could be associated with previously identified AMD viruses and plasmids, although most previously known viruses were detected in the concentrate. This finding indicates substantial

virus/phage diversity in the ecosystem, despite community dominance by a relatively small number of host populations.

To date, the CRISPR loci from hosts in the AMD microbial community found at Richmond Mine, CA have the highest reported percentage of spacers with matches to targets, whether the target be genomic, virus, or plasmid. This is likely due to a number of factors, with the strongest being the relatively low diversity and relatively deep sequencing, which improved our ability to find virus and plasmid sequences via spacer matches. Interestingly, the vast majority of bacteria and archaea in these communities have CRISPR loci, perhaps because the microbial cells are in close contact with their viruses in the high-density AMD biofilms. However, the observation of viruses (detected only in the concentrate) that may replicate in CRISPR-containing hosts but are not targeted by CRISPRs warrants more research. The hosts generally have spacers that target the virus throughout the viral genome, a phenomenon also reported in a study of targeting of phage 2972 by CRISPR spacers in its host *S. thermophilus* population (Paez-Espino et al., 2013). This observation of many spacers per viral target suggests that the failure to detect CRISPR spacers targeting the novel viral concentrate strains cannot be attributed to insufficient sequencing.

There could be several reasons for the observation that some viruses are apparently not targeted by the CRISPR-Cas system. While CRISPR spacers may have initially targeted these viruses, their genomes could have mutated much faster than occurred in other viruses for which spacer matches were identified, thus making the CRISPR-virus association unrecognizable. Despite this variation in virus mutation rates, it is unlikely that the CRISPR loci did not contain even a single spacer with an inexact match to the viruses. Absorption mechanisms or uptake blocks (Hyman and Abedon, 2010) in hosts may prevent certain viruses from injecting their genetic material, precluding spacer sampling by the Cas proteins. However, this cannot explain the results we observe, as these viruses could not exist if they were unable to successfully infect their hosts at some time. Thus, perhaps the most likely explanation is that some viruses in the AMD system (sampled in the virus concentrate) have a mechanism to elude or disable the CRISPR-Cas system, so that proto-spacer regions are never incorporated into the host genome as spacers. For example, there is evidence of viral genes that can inactivate the CRISPR-Cas immune system in *Pseudomonas aeruginosa* (Bondy-Denomy et al., 2013).

Even small changes, such as single nucleotide polymorphisms, can impact CRISPR-Cas immune potential and viral population structure. While diversity in both CRISPR loci and viral populations is apparent, temporal and spatially resolved biofilm samples reveal a surprising level of stability of virus and spacer sequences over a multi-year period in the AMD system. While *Streptococcus thermophilus* and phage 2972 system serve as a model to provide in-depth examination of short-term dynamics between a single host and phage, the results from AMD communities can be applied microbial and viral populations, such as the effects of viruses on microbial populations used the bioleaching and the resulting metal recovery process. In particular, information regarding the number of viruses per host, the stability of viral populations, and the relative importance of CRISPR-Cas system is invaluable in studying the dynamics and co-evolution of host and virus populations in natural ecosystems from which the majority of knowledge is solely derived from sequencing data.

Materials and Methods

Table 4.1 lists the complete set of samples used, which includes biofilm collected from 5way, A drift, and B drift. All datasets have been previously described and published (Tyson et al., 2004; Lo et al., 2007; Andersson and Banfield, 2008; Deneff and Banfield, 2012), with the exception of the virus concentrate dataset and the AMDV1 amplicon sequences. The AMD assembly (created with Velvet) was comprised of sequences from a number of sites and time points (Table 4.1).

Biofilm collected from the UBA location in the AMD site (description in Chapter 3) in November 2007 (Table 4.1) was used to concentrate and enrich for virus particles. The set of biofilm samples (and the associated AMD solution) were initially mechanically disrupted via mixing with a sterile syringe and mixed with 4.0 mm glass beads on a vortex. The supernatant was centrifuged at 8,000xg for 20 minutes in order to remove large particulate and cellular debris and the resulting supernatant was used to repeat the process with gradually smaller glass beads (1.0 mm and 0.1 mm). The supernatants were combined with Solid PEG 8000 (final concentration of 10% w/v) in order to reduce the solution volume and further concentrate the virus particles. After incubation at 4°C overnight in the dark, the samples were centrifuged at 13,000xg for 30 minutes. The pellet was reconstituted in 8 ml of 0.02 µm filtered Storage Media (SM) buffer (50 mM Tris, 10 mM MgSO₄, 0.1 M NaCl, pH 7.5) (Sambrook and Russell, 2001).

The virus particles in the sample were further enriched using cesium sulfate-sucrose step gradients [using a protocol modified from (Satyanarayana et al., 2004)]. Steps gradients of 1 ml each were created with 0%, 22.5%, 30%, and 37.5% (w/v) cesium sulfate in 10% (w/v) sucrose-SM. The step gradient containing the supernatant from the previous step was centrifuged at 200,000xg in a Beckman ultracentrifuge with a SW40Ti rotor for 2.5 hours at 8°C in order to separate the virus particles based on density. Different fractions, as defined by bands, were collected via syringes. SYBR Gold staining and epifluorescence microscopy were used to ensure the purity of the fractions (Chen et al., 2001; Shibata et al., 2006; Patel et al., 2007) and to visually confirm the presence of virus-like particles (VLPs). The second fraction from the bottom, approximately 1.5 ml of liquid, was selected for additional purification and concentration via a sucrose cushion [protocol modified from (Stang et al., 2005)]. After completing the selected fraction volume to 10 ml with SM buffer, the sample was layered onto a cushion of 2 ml of 30% (w/v) sucrose-SM and then subsequently centrifuged for 3 hours in a SW40Ti rotor at 30,000 rpm.

Following the sucrose cushion, transmission electron microscopy was used to observe the different morphologies of virus particles. A sample of the virus concentrate (2 µl) was placed onto Silicon Monoxide Type-A, Removable Formvar, 300 mesh, Copper grids (Ted Pella, Inc., Redding, CA, USA). After drying overnight, each grid was rinsed in 0.02 µm filtered distilled water in order to wash away any formed crystals. After staining the grid with uranyl acetate (1% w/v) for 20 seconds at room temperature and washing twice with 0.02 µm filtered distilled water. The dried grids were examined on a Hitachi model 7100 Transmission Electron Microscope.

Using a formamide extraction method, DNA was obtained from the viral concentrate sample. Briefly, the sample was combined with 0.1 volume of 2 M Tris-Cl (pH 8.5)/0.2 M EDTA, 0.01 volume 0.5 M EDTA, 50 µg of Glycogen, and 1 volume of formamide. After incubation at room temperature for 30 minutes, 2 volumes of 100% ethanol was added. The solution was inverted several times to mix and incubated at -20°C overnight. Samples were centrifuged at 16,000 xg for 20 minutes at 4°C in order to collect a pellet. Following the decantation of the ethanol supernatant, the pellet was washed twice by with cold 70% ethanol

and centrifuging at 16,000 xg for 5 minutes at 4°C. After decanting all the supernatant, the pellet was incubated overnight at room temperature to evaporate the residual ethanol and resuspended in 100 µl of sterile PCR quality water.

In order to confirm the purity of the virus concentrate sample, we amplified a region of the 16S rRNA of known AMD bacteria and archaea (Bond et al., 2000). As a control, PCR was performed on the biofilm before and after purification in 30 µl reactions, which contained 2 µl of virus concentrate DNA extract, 1X Sigma REDTaq PCR Reaction Buffer (100 mM Tris-HCl, pH 8.3, 500 mM KCl, 11 mM MgCl₂, and 0.1% gelatin), 1 µM each primer, 0.2 mM dNTP, and 1 U Sigma REDTaq DNA Polymerase (1 U/µL in 20 mM Tris-HCl, pH 8.0, 100 mM KCl, 0.1 mM EDTA, 1 mM DTT, stabilizers, 0.5% IGEPAL[®] CA-630, inert dye, 50% glycerol). The PCR reactions were completed in a MasterCycler[®] gradient thermocycler (Eppendorf, Germany), with the following program: initial step of 94°C for 12 minutes, 30 cycles of 94°C for 1 minute, an annealing step of 45°C for 45 seconds, and extension of 72°C for 1.5 minutes, and a final extension step at 72°C for 10 minutes.

Only low concentrations of DNA could be obtained for the virus concentrate sample. As this amount was determined to be insufficient to be directly used for sequencing, the extracted DNA was amplified using isothermal strand displacement with random hexamer primers and Phi29 DNA polymerase via the GenomiPhi V2 DNA Amplification Kit (GE Healthcare, Buckinghamshire, UK), with no modifications to the manufacturers' protocol. This amplification method is known to bias towards small, circular, single stranded genomes in mixed communities by two to three orders of magnitude (Haible et al., 2006; Kim et al., 2008). Thus, in an effort to reduce amplification bias, the amplification was completed in triplicate and combined.

To confirm the presence of viral sequences in the amplified, extracted DNA, a region from the AMDV1_1 gene was amplified with PCR primers (CS14435_1869F 5'-GTCTGAATGAAGCGAGCTG-3' and CS14435_2730R 5'-CTTGTCAGTCTCAATCCTGCAC-3'), designed using the known AMDV1 sequences (Andersson and Banfield, 2008). PCR was performed on the biofilm before and after purification in 50 µl reactions including 3 µl of template DNA extract, 1X Sigma REDTaq PCR Reaction Buffer (100 mM Tris-HCl, pH 8.3, 500 mM KCl, 11 mM MgCl₂, and 0.1% gelatin), 1 µM each primer, 0.2 mM dNTP, and 1 U Sigma REDTaq DNA Polymerase (1 U/µL in 20 mM Tris-HCl, pH 8.0, 100 mM KCl, 0.1 mM EDTA, 1 mM DTT, stabilizers, 0.5% IGEPAL[®] CA-630, inert dye, 50% glycerol). Reactions were carried out in an MasterCycler[®] gradient thermocycler (Eppendorf, Germany), with the cycle consisting of the following steps: 94°C for 12 minutes, 30 cycles of 94°C for 1 minute, an annealing step of 58°C for 45 seconds, and extension of 72°C for 1 minute, followed by a final extension step at 72°C for 10 minutes.

Multiple samples (Table 4.1) were used to analyze *Leptospirillum* group II CRISPR loci from C75 between 2006 and 2010. The reads in the Illumina datasets were trimmed based on quality score. The first five samples collected were sequenced with 454 technology and the trailer ends of these CRISPRs were assembled previously (Chapter 3, Figure 3.2). The remaining samples were sequenced with Illumina technology (Table 4.1). For Illumina reads, spacers were extracted after assembly due to the short length of each individual read. Bowtie was used to find reads with three or less mutations to the perfect repeat sequence (GTATTCCCCACGTTTCGTGGGGATGAACCG). Subsets of reads with CRISPR repeats and their corresponding mate pairs were created for each time point in C75. Each subset of reads was assembled with Velvet. The majority of all reads (~85% to 95%) within each dataset was used and assembled into a single contig. Custom Ruby scripts were used to extract the spacer

sequences in order for each contig. These spacers were viewed in Excel and aligned with against each other and with *Leptospirillum* group II CRISPR loci from other time points.

Sanger sequences derived from biofilm samples at the A drift (UBA and UBA-BS) were used to analyze the two CRISPR loci in E-plasma. Spacers from this dataset were extracted in order per read, unlike the Illumina dataset (from which spacers were extracted from assembled contig sequences). This method is similar to those employed in Chapter 3. Briefly, raw 454 reads were screened for the presence of ambiguous bases (i.e., “N”). Reads for CRISPR analysis were identified via sequence similarity to a repeat from either locus #1 (repeat=TTTCCATGACTGAAAAGTCATGGCTCCATTGAAG) or locus #2 (repeat=TATCAATCTCTCTAGGAGTTAGACTTTTA). Spacers were extracted via a custom Rub script. All spacer sequences were extracted *in silico* via the location of flanking repeat sequences. Spacers were also detected at the read ends if partial repeats were detected. Spacers from two CRISPR loci from G-plasma were also used in this study (Chapter 2, materials and methods).

AMDV1 amplicons were screened for vector with crossmatch and quality screened with phred (Ewing and Green, 1998; Ewing et al., 1998). The reads were then aligned with MUSCLE (Edgar, 2004) and the alignment was imported to and visualized in Strainer (Eppley et al., 2007) (Figure 4.3). Using the *Leptospirillum* group CRISPR locus PAM, all potential proto-spacer sequences were identified at the read level. Then, using these potential proto-spacers, we examined the relative abundance of all reads (from the alignment) that match the potential proto-spacers. Note that some proto-spacers were similar due to single nucleotide polymorphisms. The relative abundance values for each potential proto-spacer were used for hierarchical clustering, generating a diagram using black/red values to show abundance.

Reads from the virus concentrate and all C75 454 samples were mapped against the AMDV2 genome sequence (Andersson and Banfield, 2008), similar to the method used in Chapter 1. Only the virus concentrate and August 2007 C75 sample contained reads that mapped to AMDV2. Briefly, gsMapper was used to identify single nucleotide polymorphisms and generate an .ace file to allow viewing of the assembly in Consed (Gordon et al., 1998).

Prior to any searches of reads or contigs for spacer matches, the sequences containing the corresponding CRISPR repeats were removed from analysis in order to detect phage, viruses, and plasmids rather than CRISPR loci themselves. Similarity searches were complete with either blast (Altschul et al., 1997) or usearch (Edgar, 2010). A perfect match is considered a spacer that matches 100% across the entire length while an imperfect or an inexact match is defined as a spacer match having a threshold of 80% identify over 80% length.

Firstly, spacers were searched against all contigs assembled from all available Illumina datasets. This subset of contigs with at least a single match to a spacer was used to search against known AMD viruses, AMD organisms, ACLAME database (Leplae et al., 2010), and the NCBI nr database. Sequences with many spacer matches, without any matches to known AMD sequences, or those that have matches to sequences within the ACLAME database were screened through InterPro scan (Zdobnov and Apweiler, 2001)(results shown in Table S4.3).

Multiple searches were completed for the virus concentrate read database. All reads from the virus concentrate sample were searched against all known AMD organisms (at the nucleotide level) to determine the relative abundance of reads that match to each organism (Figure 4.7). Then, reads were searched for similarity to known AMD viruses. Remaining reads were then screened for the presence of perfect or imperfect spacer matches. Searches to translated DNA (proteins or predicted proteins) from the ACLAME database and subsequently to searched against all known AMD organism genomes. Then, the reads were also searched against the

SILVA 16s rRNA database (Quast et al., 2013) to screen out any remaining 16s sequences in the dataset. Finally, all remaining reads that could not be identified was searched for any similar sequences within the non-redundant NCBI database.

All *Leptospirillum* group II spacers from C75 between November 2007 and December 2010 were searched against all 454 databases in order to find perfect and imperfect target sequences. E-plasma spacers were the searched against all reads from Sanger dataset and virus concentrate reads that comprise AMDV2 as well as other archaeal viruses (search was completed in a similar method as in Chapter 2). E-plasma and *Leptospirillum* group II CRISPR spacers with matches to their associated viruses were identified and the location of spacers within the locus was determined (Figures 4.5, 4.8-4.10).

Figure 4.1. Comparison of reconstructed CRISPR loci in *Leptospirillum* Group II between June 2006 and Dec 2010 from the C75 site. Each thick horizontal bar represents the reconstructed CRISPR locus (leader on left end). Colors represent blocks of spacers shared between time points. Horizontal stripes indicate missing/excised spacers while white space indicates a lack of sequencing depth at that location. Gene insertions are indicated.

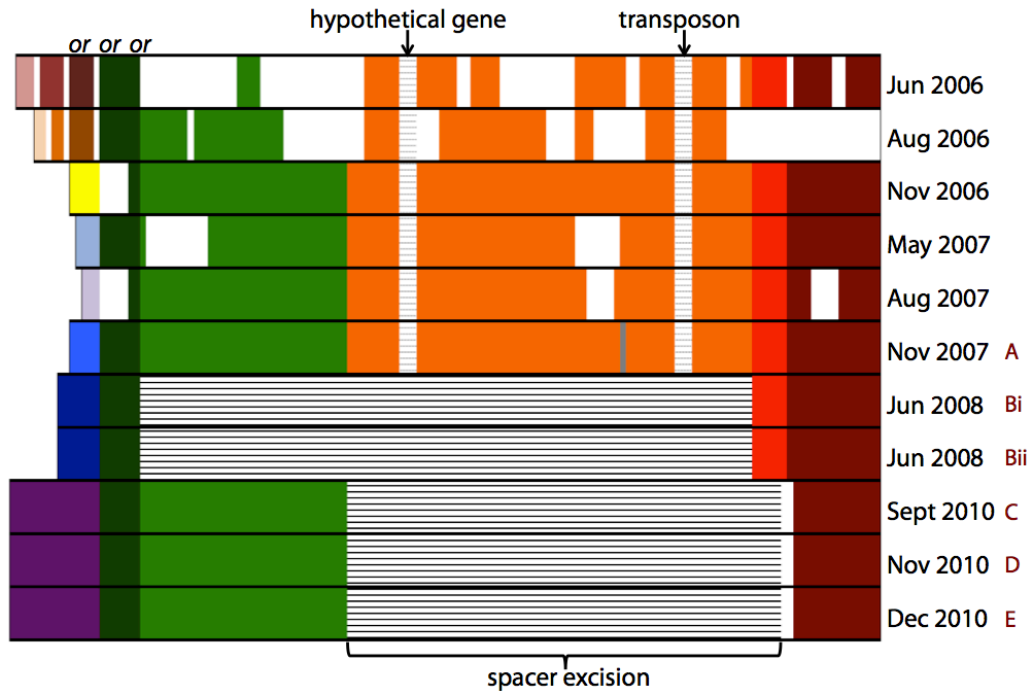


Figure 4.2. Comparison of reconstructed CRISPR loci and corresponding spacer targets in *Leptospirillum* Group II between June 2006 and Dec 2010 from the C75. Representation of CRISPR loci similar to loci in Figure 4.1. Boxes underneath spacers represent matches to targets in the corresponding dataset (x-axis). Black boxes indicate perfect matches while grey shows imperfect matches. Legend: A=November 2007, Bi=June 2008, Bii=June 2008, C=September 2010, D=November 2010, E=December 2010.

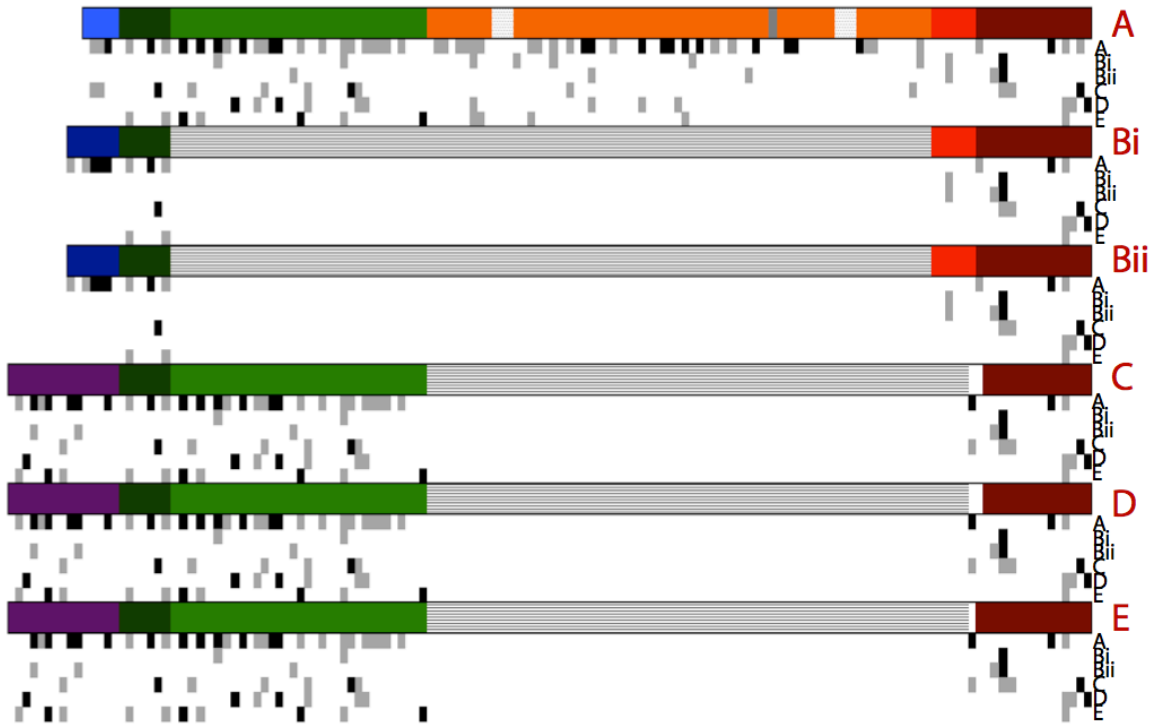


Figure 4.3. Sequence alignment of an amplified region in bacteriophage AMDV1 across time and space. Image of MUSCLE-aligned sequences viewed in Strainer. Each light gray horizontal line represents a single Sanger sequence. Different colored vertical bar represent different nucleotide changes.

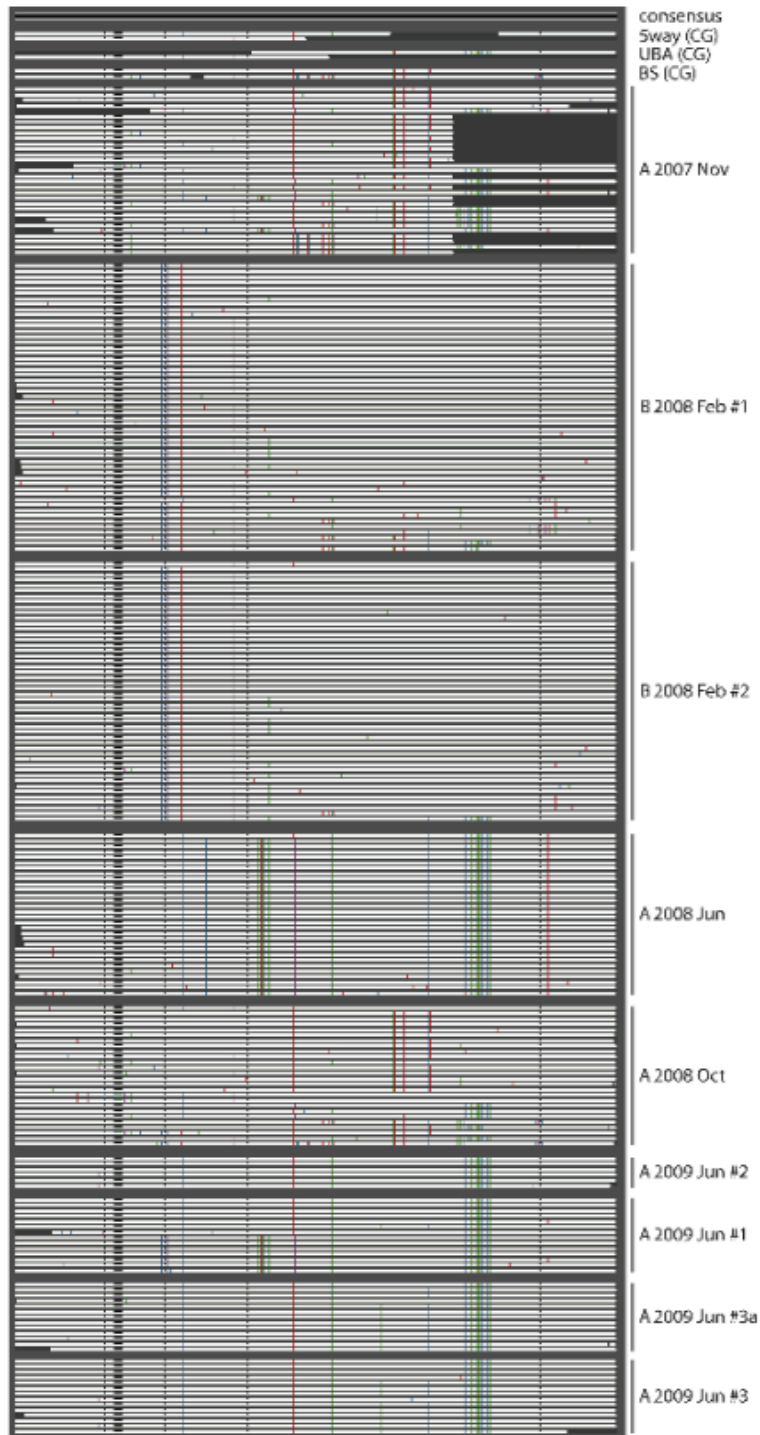


Figure 4.4. Hierarchical clustering of potential *Leptospirillum* group II spacer target across time and space. Each box indicates the relative percentages of sequences with exact spacer match and intact PAM. Scale increases from black (0%) to red (100%).



Figure 4.5. Reconstruction of *Leptospirillum* group II CRISPR loci from 5way. Loci are shown vertically from trailer to leader end, with spacers represented as wide rectangles. White rectangles represent spacers shared between at least two loci while colored rectangles represent spacers unique to a specific locus. Stripped lines in the loci show spacer loss. In the column left of each reconstructed locus, the presence of a square indicates a match to Scaffold 1127. Black squares represent perfect matches while gray squares indicate imperfect matches. The locus on the left was reconstructed with Sanger sequences while the one on the right was reconstructed with 454 sequences (Table 4.1).



Figure 4.6. Representative morphologies of virus like particles purified from the AMD biofilm. Overviews of the virus-like particle (VLP) community show a variety of morphologies (A, B, and C). A number of round particles were identified, including some with internal structure (D-F) and some with a thick outer layer (G, H). Icosahedral VLPs were also identified, which had a thin outer layer (I, J). Bullet-shaped VLPs with morphology similar to the Rhabdoviridae were also observed in the purified virus fraction (K-N), as well as spindle-shaped VLPs similar to the Fuselloviridae (O, P).

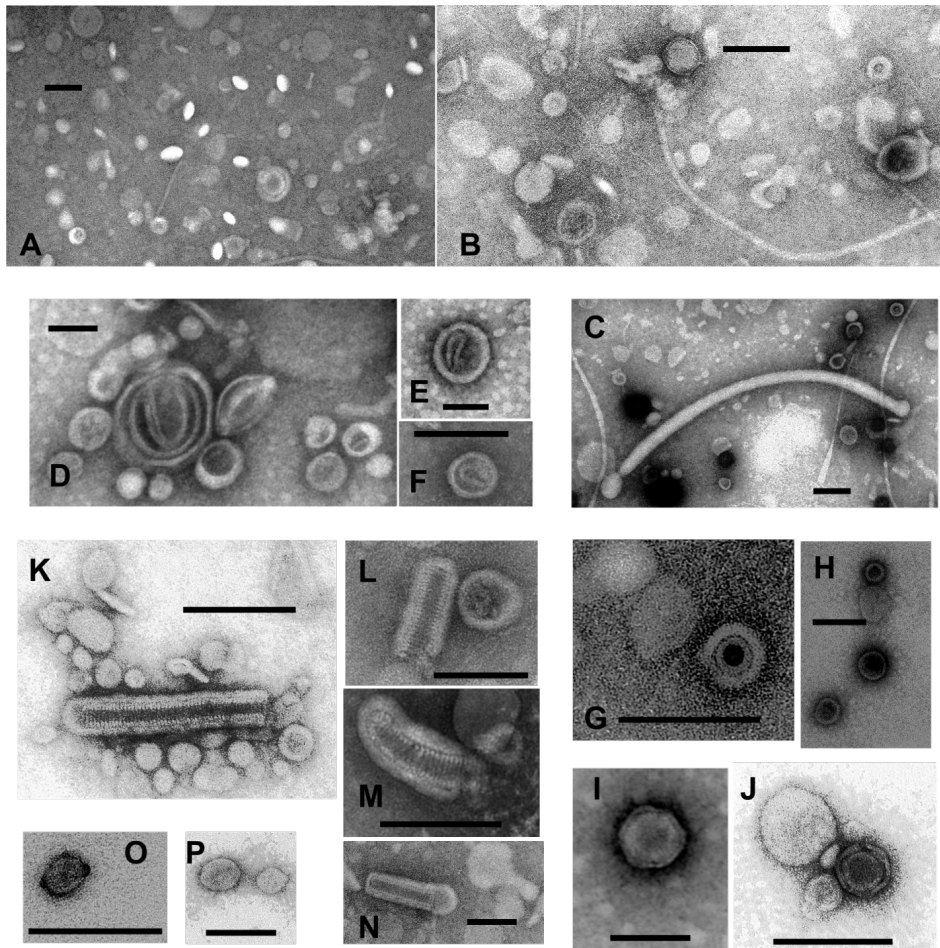
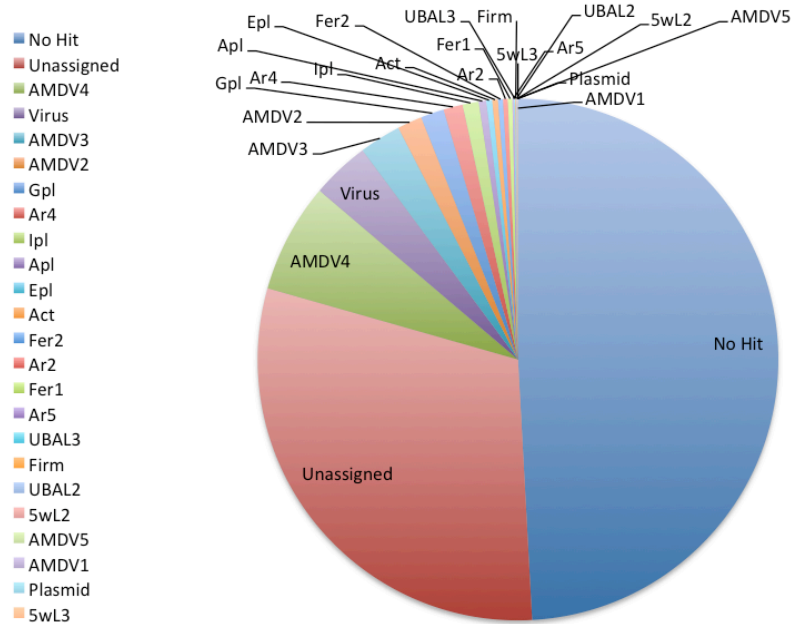


Figure 4.7. Similarity matches between total virus concentrate reads to AMD microbial genomes, viruses, and plasmids. A. Pie represents portions of reads with translated protein matches to proteins from AMD microbial genomes, viruses, and plasmids. B. Pie represents portions of reads with nucleotide matches to archaeal viruses.

A.



B.

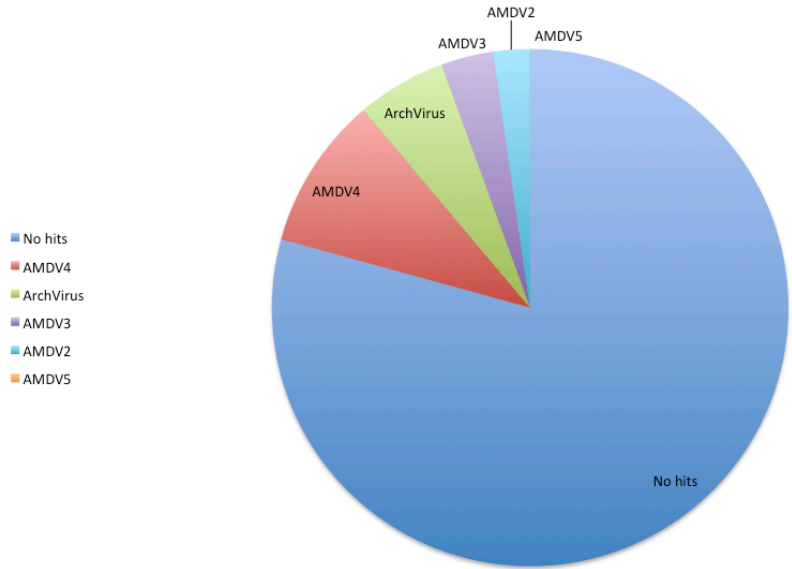


Figure 4.8. Reconstructed E-plasma #1 loci and spacer matches to AMDV2. In each sampling, the CRISPR spacers (boxes) are aligned vertically according to their ordering in the metagenomic reads, with CRISPR repeats removed for compactness. Boxes filled with the same color at the same vertical position represent identical spacers. Black-filled boxes show flanking genetic material and white-filled boxes denote cell-specific spacers found only once in the dataset. When separated spacers can be linked via paired reads, the intervening region is shown as a grey bar. When spacers match AMDV2 reads, filled squares indicate perfect matches while open circles indicate imperfect matches. A. Spacers matches to AMDV2 derived from June 2005, and July 2005 (Sanger). B. Spacers matches to AMDV2 derived from November 2007 (454, virus concentrate).

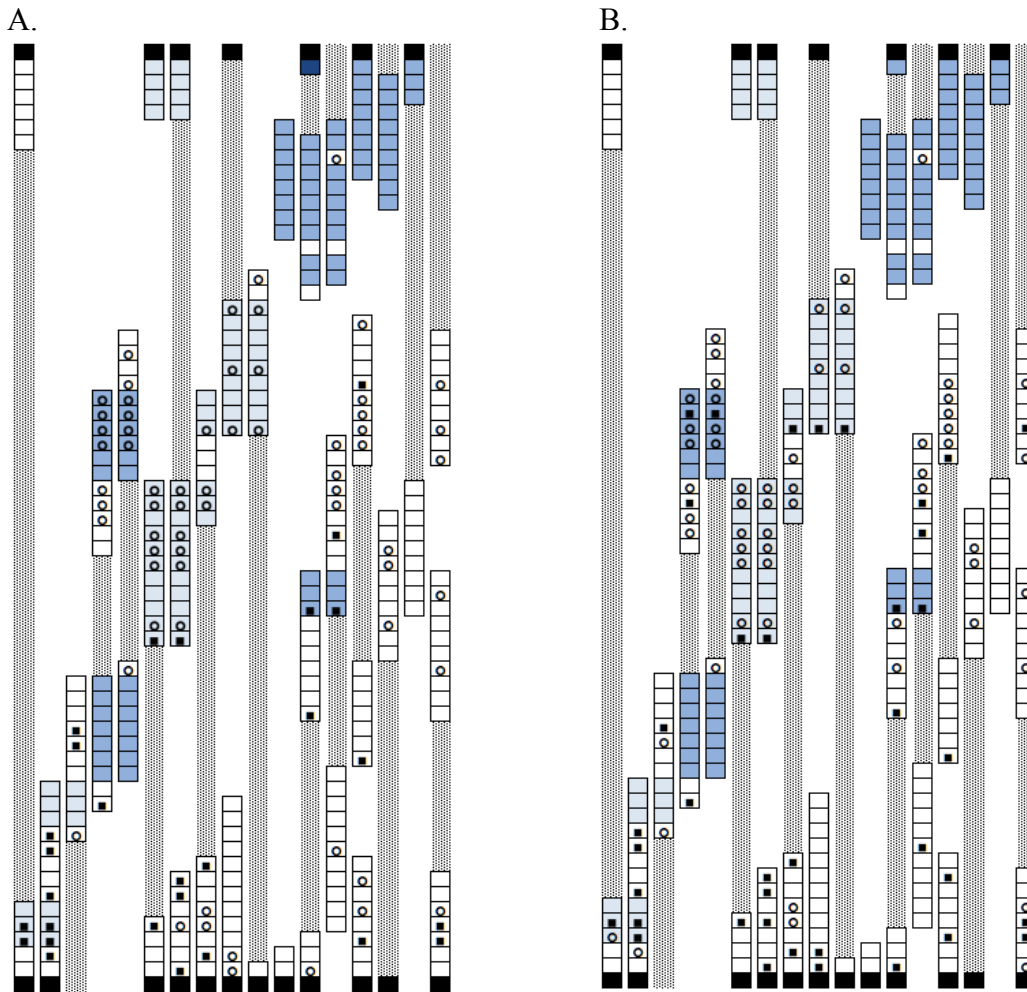
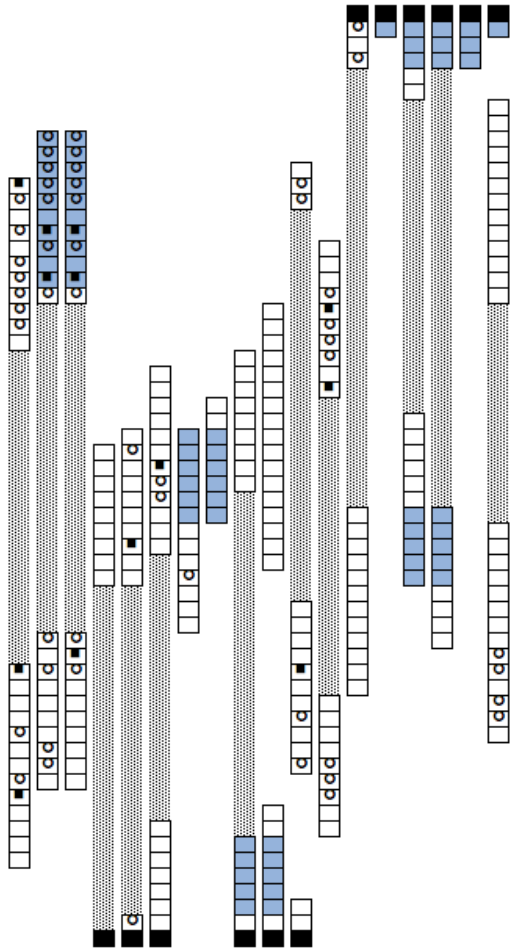


Figure 4.9. Reconstructed E-plasma loci #2 and spacer matches to AMDV2. Loci and spacer matches represented in the same manner as Figure 4.8. A. Spacers matches to AMDV2 derived from March 2002, June 2005, and July 2005 (Sanger). B. Spacers matches to AMDV2 derived from November 2007 (454, virus concentrate).

A.



B.

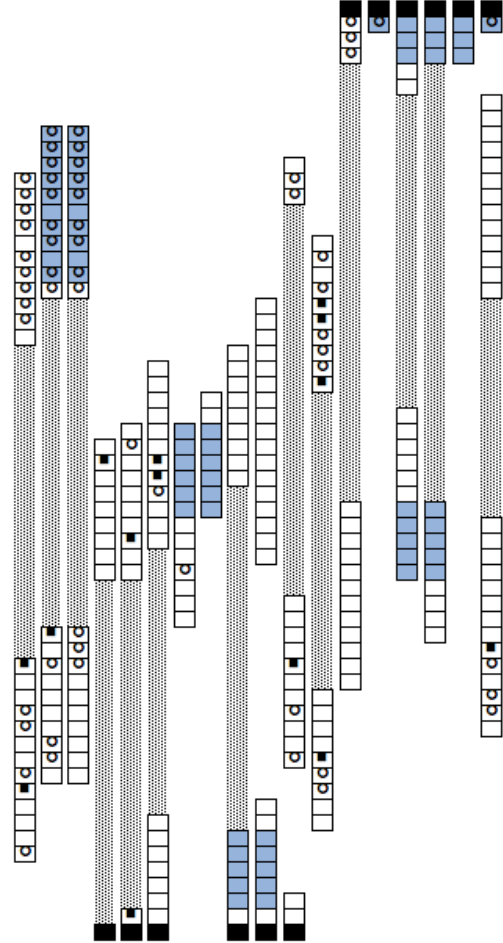


Figure 4.10. Reconstructed E-plasma loci #1 (A.) and #2 (B.) and spacer matches to multiple archaeal viruses. In each sampling, the CRISPR spacers (boxes) are aligned vertically according to their ordering in the metagenomic reads, with CRISPR repeats removed for compactness. Boxes outlined in thick black lines at the same vertical position represent identical spacers. Black-filled boxes show flanking genetic material. When separated spacers can be linked via paired reads, the intervening region is shown as a grey bar. When spacers match archaeal viruses, filled squares indicate perfect matches while open circles indicate imperfect matches.

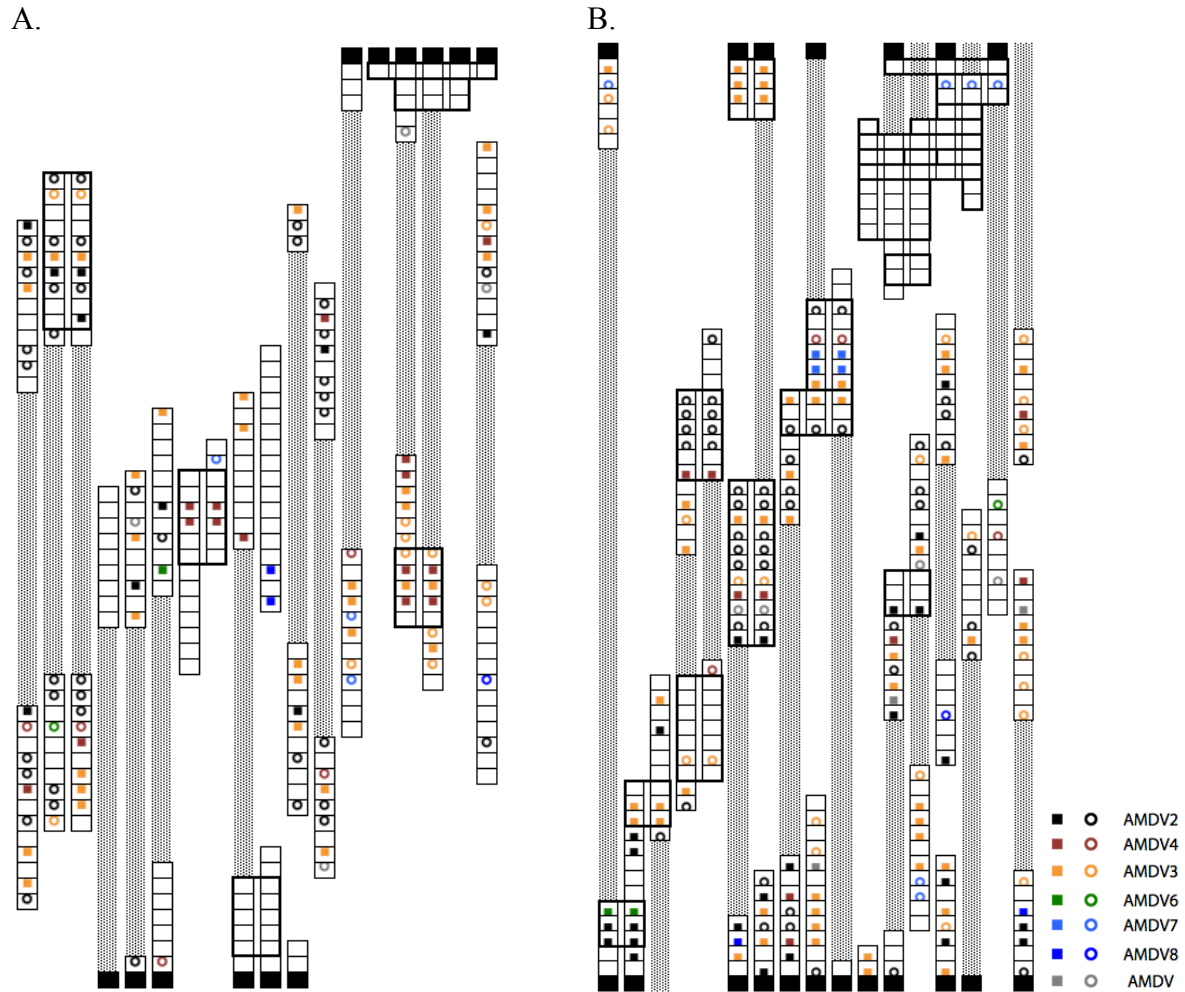
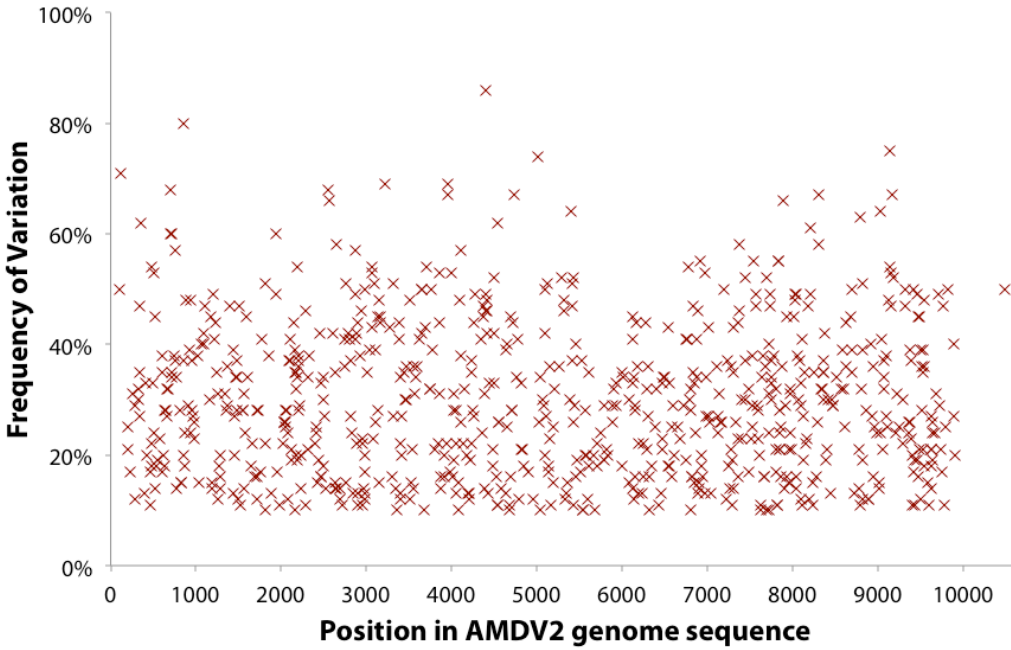


Figure 4.11. Frequency of single nucleotide polymorphisms (SNPs) across the AMDV2 genome. Graphs plot the location of polymorphisms across the AMDV2 genome location against the relative abundance of each variant position in the AMDV2 population derived from A) November 2007 and B) August 2007.

A.



B.

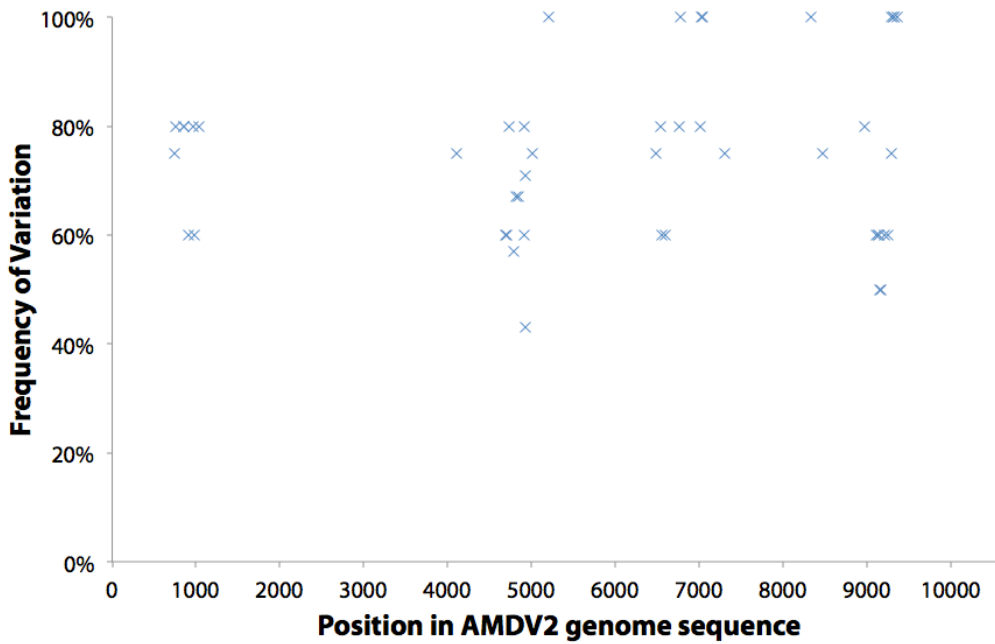


Table 4.1. Sampling and sequencing information for biofilms used in this study. Table lists the datasets from which sequences were used to assemble the loci or identify spacers from the *Leptospirillum* group II and group III, E-plasma, and G-plasma CRISP as well as those used to analyze bacteriophage AMDV1, archaeal virus AMDV2, virus concentrate (“Conc”), and viruses from AMD datasets (“Viruses”). Abbreviations: “CG” = community genomics, “Amp” = amplicons, “MDA” = MDA-treated DNA.

Location	Date	#	Type	Seq	Lepto II	Lepto III	E-plasma	G-plasma	AMVD1	AMDV2	Conc	Viruses
5way	Mar 2002	1	CG	Sanger	X	X						
	Mar 2002	1	Amp	454	X	X						
A (UBA)	Jun 2005	1	CG	Sanger	X	X	X	X		X		
	Jul 2005	1	Amp	454	X							
	Nov 2005	1	CG	Sanger		X	X	X		X		
	Nov 2007	1	MDA	454						X	X	
	Nov 2007	1	Amp	Sanger					X			
	Jun 2008	1	Amp	Sanger					X			
	Oct 2008	1	Amp	Sanger					X			
	Jun 2009	4	Amp	Sanger					X			
AB 20m	Jun 2006	4	CG	Illumina								X
AB muck	May 2007	1	CG	Illumina								X
	Aug 2007	1	CG	Illumina								X
B 2m	Feb 2008	1	Amp	Sanger					X			
B 8m	Feb 2008	1	Amp	Sanger					X			
C 10m	Aug 2006	1	CG	Illumina								X
	Nov 2006	1	CG	Illumina								X
	Aug 2007	1	CG	Illumina								X
C 75m	Jun 2006	1	CG	454	X							
	Aug 2006	1	CG	454	X							
	Nov 2006	1	CG	454	X							
	May 2007	1	CG	454	X							
	Aug 2007	1	CG	454	X					X		
	Nov 2007	1	CG	Illumina								X
	Jun 2008	3	CG	Illumina								X
	Oct 2008	1	CG	Illumina								X
	Sept 2010	1	CG	Illumina								X
	Nov 2010	1	CG	Illumina								X
	Dec 2010	1	CG	Illumina								X

References

- Allen, E.E., and Banfield, J.F. (2005) Community genomics in microbial ecology and evolution. *Nat Rev Micro* **3**: 489-498.
- Allen, E.E., Tyson, G.W., Whitaker, R.J., Detter, J.C., Richardson, P.M., and Banfield, J.F. (2007) Genome dynamics in a natural archaeal population. *Proceedings of the National Academy of Sciences* **104**: 1883-1888.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389-3402.
- Altshuler, D., Pollara, V.J., Cowles, C.R., Van Etten, W.J., Baldwin, J., Linton, L., and Lander, E.S. (2000) An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407**: 513-516.
- Andersson, A.F., and Banfield, J.F. (2008) Virus Population Dynamics and Acquired Virus Resistance in Natural Microbial Communities. *Science* **320**: 1047-1050.
- Arrigo, K.R. (2005) Marine microorganisms and global nutrient cycles. *Nature* **437**: 349-355.
- Baker, B.J., and Banfield, J.F. (2003) Microbial communities in acid mine drainage. *FEMS Microbiology Ecology* **44**: 139-152.
- Baker, B.J., Comolli, L.R., Dick, G.J., Hauser, L.J., Hyatt, D., Dill, B.D. et al. (2010) Enigmatic, ultrasmall, uncultivated Archaea. *Proceedings of the National Academy of Sciences* **107**: 8806-8811.
- Banfield, J.F., and Young, M. (2009) Variety--the Splice of Life--in Microbial Communities. *Science* **326**: 1198-1199.
- Banfield, J.F., VerBerkmoes, N.C., Hettich, R.L., and Thelen, M.P. (2005) Proteogenomic Approaches for the Molecular Characterization of Natural Microbial Communities. *OMICS: A Journal of Integrative Biology* **9**: 301-333.
- Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S. et al. (2007) CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* **315**: 1709-1712.
- Barrett, R.D.H., M'Gonigle, L.K., and Otto, S.P. (2006) The distribution of beneficial mutant effects under strong selection. *Genetics* **174**: 2071-2079.
- Beaumont, H.J.E., Gallie, J., Kost, C., Ferguson, G.C., and Rainey, P.B. (2009) Experimental evolution of bet hedging. *Nature* **462**: 90-93.
- Bohannon, B.J.M., and Lenski, R.E. (2000) Linking genetic change to community evolution: insights from studies of bacteria and bacteriophage. *Ecology Letters* **3**: 362-377.
- Bolotin, A., Quinquis, B., Sorokin, A., and Ehrlich, S.D. (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* **151**: 2551-2561.
- Bolotin, A., Quinquis, B., Renault, P., Sorokin, A., Ehrlich, S.D., Kulakauskas, S. et al. (2004) Complete sequence and comparative genome analysis of the dairy bacterium *Streptococcus thermophilus*. *Nat Biotech* **22**: 1554-1558.
- Bond, P.L., Smriga, S.P., and Banfield, J.F. (2000) Phylogeny of microorganisms populating a thick, subaerial, predominantly lithotrophic biofilm at an extreme acid mine drainage site. *Applied and Environmental Microbiology* **66**: 3842-3849.
- Bondy-Denomy, J., Pawluk, A., Maxwell, K.L., and Davidson, A.R. (2013) Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature* **493**: 429-432.

Bosecker, K. (1997) Bioleaching: metal solubilization by microorganisms. *FEMS Microbiology Reviews* **20**: 591-604.

Brockhurst, M.A., Morgan, A.D., Fenton, A., and Buckling, A. (2007) Experimental coevolution with bacteria and phage: The *Pseudomonas fluorescens*, λ 2 model system. *Infection, Genetics and Evolution* **7**: 547-552.

Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W.L. et al. (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res* **18**: 763-770.

Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L. et al. (2008) Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science* **321**: 960-964.

Brussow, H., Canchaya, C., and Hardt, W.-D. (2004) Phages and the Evolution of Bacterial Pathogens: from Genomic Rearrangements to Lysogenic Conversion. *Microbiol Mol Biol Rev* **68**: 560-602.

Buckling, A., and Rainey, P.B. (2002) Antagonistic coevolution between a bacterium and a bacteriophage. *Proceedings of the Royal Society of London Series B: Biological Sciences* **269**: 931-936.

Cady, K.C., and O'Toole, G.A. (2011) Non-identity targeting of *Yersinia*-subtype CRISPR-phage interaction requires the Csy and Cas3 proteins. *Journal of Bacteriology*.

Cady, K.C., White, A.S., Hammond, J.H., Abendroth, M.D., Karthikeyan, R.S.G., Lalitha, P. et al. (2011) Prevalence, conservation and functional analysis of *Yersinia* and *Escherichia* CRISPR regions in clinical *Pseudomonas aeruginosa* isolates. *Microbiology* **157**: 430-437.

Canchaya, C., Fournous, G., Chibani-Chennoufi, S., Dillmann, M.-L., and Brussow, H. (2003) Phage as agents of lateral gene transfer. *Current Opinion in Microbiology* **6**: 417-424.

Carmel, L., and Koonin, E.V. (2009) A universal nonmonotonic relationship between gene compactness and expression levels in multicellular eukaryotes. *Genome biology and evolution* **1**: 382.

Castillo-Davis, C.I., Mekhedov, S.L., Hartl, D.L., Koonin, E.V., and Kondrashov, F.A. (2002) Selection for short introns in highly expressed genes. *Nature genetics* **31**: 415-418.

Chen, F., Lu, J.R., Binder, B.J., Liu, Y.C., and Hodson, R.E. (2001) Application of digital image analysis and flow cytometry to enumerate marine viruses stained with SYBR gold. *Applied and Environmental Microbiology* **67**: 539-545.

Chibani-Chennoufi, S., Bruttin, A., Dillmann, M.L., and Brussow, H. (2004) Phage-host interaction: an ecological perspective. *Journal of Bacteriology* **186**: 3677-3686.

Cohen, D. (1966) Optimizing reproduction in a randomly varying environment. *Journal of theoretical biology* **12**: 119-129.

Dawkins, R., Krebs, J.R., and Krebs, J. (1979) Arms races between and within species. *Proceedings of the Royal Society of London Series B Biological Sciences* **205**: 489-511.

Denef, V.J., and Banfield, J.F. (2012) In Situ Evolutionary Rate Measurements Show Ecological Success of Recently Emerged Bacterial Hybrids. *Science* **336**: 462-466.

Denef, V.J., Mueller, R.S., and Banfield, J.F. (2010a) AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. *ISME J* **4**: 599-610.

Denef, V.J., Kalnejais, L.H., Mueller, R.S., Wilmes, P., Baker, B.J., Thomas, B.C. et al. (2010b) Proteogenomic basis for ecological divergence of closely related bacteria in natural acidophilic microbial communities. *Proceedings of the National Academy of Sciences* **107**: 2383-2390.

Desai, M.M., and Fisher, D.S. (2007) Beneficial mutation, selection balance and the effect of linkage on positive selection. *Genetics* **176**: 1759-1798.

- Deveau, H., Barrangou, R., Garneau, J.E., Labonte, J., Fremaux, C., Boyaval, P. et al. (2008) Phage Response to CRISPR-Encoded Resistance in *Streptococcus thermophilus*. *J Bacteriol* **190**: 1390-1400.
- Dick, G.J., Andersson, A.F., Baker, B.J., Simmons, S.L., Thomas, B.C., Yelton, A.P., and Banfield, J.F. (2009) Community-wide analysis of microbial genome sequence signatures. *Genome Biol* **10**: R85.
- Díez-Villaseñor, C., Almendros, C., García-Martínez, J., and Mojica, F.J.M. (2010) Diversity of CRISPR loci in *Escherichia coli*. *Microbiology* **156**: 1351-1361.
- Edgar, R., and Qimron, U. (2010) The *Escherichia coli* CRISPR system protects from Φ lysogenization, lysogens, and prophage induction. *Journal of Bacteriology* **192**: 6291-6294.
- Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792-1797.
- Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460-2461.
- Edwards, K.J., Bond, P.L., Gihring, T.M., and Banfield, J.F. (2000) An Archaeal Iron-Oxidizing Extreme Acidophile Important in Acid Mine Drainage. *Science* **287**: 1796-1799.
- Edwards, R.A., and Rohwer, F. (2005) Viral metagenomics. *Nature Reviews Microbiology* **3**: 504-510.
- Eppley, J., Tyson, G., Getz, W., and Banfield, J. (2007) Strainer: software for analysis of population variation in community genomic datasets. *BMC Bioinformatics* **8**: 398.
- Ewing, B., and Green, P. (1998) Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Res* **8**: 186-194.
- Ewing, B., Hillier, L., Wendl, M.C., and Green, P. (1998) Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assessment. *Genome Res* **8**: 175-185.
- Fuhrman, J.A. (1999) Marine viruses and their biogeochemical and ecological effects. *Nature* **399**: 541-548.
- Furukawa, K. (2003) „Super bugs,“ for bioremediation. *Trends in Biotechnology* **21**: 187-190.
- Garneau, J.E., Dupuis, M.-E., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P. et al. (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* **468**: 67-71.
- Garrett, R.A., Shah, S.A., Vestergaard, G., Deng, L., Gudbergdottir, S., Kenchappa, C.S. et al. (2011) CRISPR-based immune systems of the Sulfolobales: complexity and diversity. *Biochemical Society Transactions* **39**: 51.
- Gerrish, P.J., and Lenski, R.E. (1998) The fate of competing beneficial mutations in an asexual population. *Genetica* **102**: 127-144.
- Giovannoni, S.J., and Vergin, K.L. (2012) Seasonality in Ocean Microbial Communities. *Science* **335**: 671-676.
- Goltsman, D.S.A., Deneff, V.J., Singer, S.W., VerBerkmoes, N.C., Lefsrud, M., Mueller, R.S. et al. (2009) Community genomic and proteomic analyses of chemoautotrophic iron-oxidizing „*Leptospirillum rubrum*,“ (group II) and „*Leptospirillum ferrodiazotrophum*,“ (group III) bacteria in acid mine drainage biofilms. *Applied and Environmental Microbiology* **75**: 4599-4615.
- Gómez, P., and Buckling, A. (2011) Bacteria-phage antagonistic coevolution in soil. *Science* **332**: 106-109.

Gomez-Alvarez, V., Teal, T.K., and Schmidt, T.M. (2009) Systematic artifacts in metagenomes from complex microbial communities. *The ISME journal* **3**: 1314-1317.

Gordon, D., Abajian, C., and Green, P. (1998) Consed: A Graphical Tool for Sequence Finishing. *Genome Res* **8**: 195-202.

Grissa, I., Bouchon, P., Pourcel, C., and Vergnaud, G. (2008) On-line resources for bacterial micro-evolution studies using MLVA or CRISPR typing. *Biochimie* **90**: 660-668.

Gudbergsdottir, S., Deng, L., Chen, Z., Jensen, J.V.K., Jensen, L.R., She, Q., and Garrett, R.A. (2011) Dynamic properties of the *Sulfolobus* CRISPR/Cas and CRISPR/Cmr systems when challenged with vector, Åborne viral and plasmid genes and protospacers. *Molecular Microbiology* **79**: 35-49.

Haerter, J.O., Trusina, A., and Sneppen, K. (2011) Targeted bacterial immunity buffers phage diversity. *Journal of virology* **85**: 10554-10560.

Haible, D., Kober, S., and Jeske, H. (2006) Rolling circle amplification revolutionizes diagnosis and genomics of geminiviruses. *Journal of Virological Methods* **135**: 9-16.

Handelsman, J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews* **68**: 669-685.

Haurwitz, R.E., Jinek, M., Wiedenheft, B., Zhou, K., and Doudna, J.A. (2010) Sequence- and Structure-Specific RNA Processing by a CRISPR Endonuclease. *Science* **329**: 1355-1358.

He, J., and Deem, M.W. (2010) Heterogeneous diversity of spacers within CRISPR (clustered regularly interspaced short palindromic repeats). *Physical review letters* **105**: 128102.

Heidelberg, J.F., Nelson, W.C., Schoenfeld, T., and Bhaya, D. (2009) Germ Warfare in a Microbial Mat Community: CRISPRs Provide Insights into the Co-Evolution of Host and Viral Genomes. *PLoS ONE* **4**.

Held, N.L., and Whitaker, R.J. (2009) Viral biogeography revealed by signatures in *Sulfolobus islandicus* genomes. *Environ Microbiol* **11**: 457-466.

Horvath, P., and Barrangou, R. (2010) CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**: 167-170.

Horvath, P., Romero, D.A., Coute-Monvoisin, A.-C., Richards, M., Deveau, H., Moineau, S. et al. (2008) Diversity, Activity, and Evolution of CRISPR Loci in *Streptococcus thermophilus*. *J Bacteriol* **190**: 1401-1412.

Hugenholtz, P. (2002) Exploring prokaryotic diversity in the genomic era. *Genome Biol* **3**: 1-0003.0008.

Huggins, A.R., and Sandine, W.E. (1984) Differentiation of Fast and Slow Milk-Coagulating Isolates in Strains of Lactic Streptococci. *J Dairy Sci* **67**: 1674-1679.

Huse, S.M., Huber, J.A., Morrison, H.G., Sogin, M.L., and Welch, D.M. (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol* **8**: R143.

Hyman, P., and Abedon, S.T. (2010) Chapter 7 - Bacteriophage Host Range and Bacterial Resistance. In *Advances in Applied Microbiology*. Allen, I.L., Sima, S., and Geoffrey, M.G. (eds): Academic Press, pp. 217-248.

Johnson, D.B., and Hallberg, K.B. (2005) Acid mine drainage remediation options: a review. *Science of The Total Environment* **338**: 3-14.

Karginov, F.V., and Hannon, G.J. (2010) The CRISPR System: Small RNA-Guided Defense in Bacteria and Archaea. *Molecular Cell* **37**: 7-19.

Kim, K., Chang, H., Nam, Y., Roh, S., Kim, M., Sung, Y. et al. (2008) Amplification of uncultured single-stranded DNA viruses from rice paddy soil. *Applied and Environmental Microbiology* **74**: 5975-5985.

Koonin, E.V., and Wolf, Y.I. (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research* **36**: 6688-6719.

Kunin, V., Sorek, R., and Hugenholtz, P. (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biology* **8**: R61.

Kuo, C.H., and Ochman, H. (2009) Deletional bias across the three domains of life. *Genome biology and evolution* **1**: 145.

Kuo, C.H., and Ochman, H. (2010) The extinction dynamics of bacterial pseudogenes. *PLoS genetics* **6**: e1001050.

Labrie, S.J., Samson, J.E., and Moineau, S. (2010) Bacteriophage resistance mechanisms. *Nat Rev Micro* **8**: 317-327.

Leplae, R., Lima-Mendez, G., and Toussaint, A. (2010) ACLAME: A CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Research* **38**: D57-D61.

Leroy, F., and De Vuyst, L. (2004) Lactic acid bacteria as functional starter cultures for the food fermentation industry. *Trends in Food Science & Technology* **15**: 67-78.

Levesque, C., Duplessis, M., Labonte, J., Labrie, S., Fremaux, C., Tremblay, D., and Moineau, S. (2005) Genomic Organization and Molecular Analysis of Virulent Bacteriophage 2972 Infecting an Exopolysaccharide-Producing *Streptococcus thermophilus* Strain. *Appl Environ Microbiol* **71**: 4057-4068.

Levin, B.R. (2010) Nasty Viruses, Costly Plasmids, Population Dynamics, and the Conditions for Establishing and Maintaining CRISPR-Mediated Adaptive Immunity in Bacteria. *PLoS Genet* **6**: e1001171.

Lo, I., Denef, V.J., VerBerkmoes, N.C., Shah, M.B., Goltsman, D., DiBartolo, G. et al. (2007) Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* **446**: 537-541.

Lythgoe, K.A., and Chao, L. (2003) Mechanisms of coexistence of a bacteria and a bacteriophage in a spatially homogeneous environment. *Ecology Letters* **6**: 326-334.

Makarova, K., Grishin, N., Shabalina, S., Wolf, Y., and Koonin, E. (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biology Direct* **1**: 7.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J.J., Charpentier, E., Horvath, P. et al. (2011) Evolution and classification of the CRISPR-Cas systems. *Nat Rev Micro* **9**: 467-477.

Manica, A., Zebec, Z., Teichmann, D., and Schleper, C. (2011) In vivo activity of CRISPR-mediated virus defence in a hyperthermophilic archaeon. *Molecular Microbiology* **80**: 481-491.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bembien, L.A. et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376-380.

Marraffini, L.A., and Sontheimer, E.J. (2008) CRISPR Interference Limits Horizontal Gene Transfer in *Staphylococci* by Targeting DNA. *Science* **322**: 1843-1845.

Marraffini, L.A., and Sontheimer, E.J. (2010) CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet* **11**: 181-190.

Mojica, F.J.M., Díez-Villaseñor, C.s., García-Martínez, J., and Soria, E. (2005) Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. *Journal of Molecular Evolution* **60**: 174-182.

Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J., and Almendros, C. (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* **155**: 733-740.

Morran, L.T., Schmidt, O.G., Gelarden, I.A., Parrish, R.C., and Lively, C.M. (2011) Running with the Red Queen: host-parasite coevolution selects for biparental sex. *Science* **333**: 216-218.

Nowak, M.A., and May, R. (2001) *Virus dynamics: mathematical principles of immunology and virology*: Oxford University Press, USA.

Paez-Espino, D., Morovic, W., Sun, C.L., Thomas, B.C., Ueda, K.-i., Stahl, B. et al. (2013) Strong bias in the bacterial CRISPR elements that confer immunity to phage. *Nat Commun* **4**: 1430.

Pal, C., Macia, M.D., Oliver, A., Schachar, I., and Buckling, A. (2007) Coevolution with viruses drives the evolution of bacterial mutation rates. *Nature* **450**: 1079-1081.

Palmer, K.L., and Gilmore, M.S. (2010) Multidrug-resistant enterococci lack CRISPR-cas. *MBio* **1**.

Palmer, K.L., and Whiteley, M. (2011) DMS3-42: The secret to CRISPR-dependent biofilm inhibition in *Pseudomonas aeruginosa*. *Journal of Bacteriology* **193**: 3431-3432.

Patel, A., Noble, R.T., Steele, J.A., Schwabach, M.S., Hewson, I., and Fuhrman, J.A. (2007) Virus and prokaryote enumeration from planktonic aquatic environments by epifluorescence microscopy with SYBR Green I. *Nature Protocols* **2**: 269-276.

Paterson, S., Vogwill, T., Buckling, A., Benmayor, R., Spiers, A.J., Thomson, N.R. et al. (2010) Antagonistic coevolution accelerates molecular evolution. *Nature* **464**: 275-278.

Pourcel, C., Salvignol, G., and Vergnaud, G. (2005) 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.

Prangishvili, D., Forterre, P., and Garrett, R.A. (2006) Viruses of the Archaea: a unifying view. *Nature Reviews Microbiology* **4**: 837-848.

Pride, D.T., Sun, C.L., Salzman, J., Rao, N., Loomer, P., Armitage, G.C. et al. (2011) Analysis of streptococcal CRISPRs from human saliva reveals substantial sequence diversity within and between subjects over time. *Genome Res* **21**: 126-136.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P. et al. (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**: D590-D596.

Riesenfeld, C.S., Schloss, P.D., and Handelsman, J. (2004) Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* **38**: 525-552.

Rodriguez-Brito, B., Li, L.L., Wegley, L., Furlan, M., Angly, F., Breitbart, M. et al. (2010) Viral and microbial community dynamics in four aquatic environments. *The ISME journal* **4**: 739-751.

Rousseeuw, P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics* **20**: 53-65.

Sambrook, J., and Russell, D.W. (eds) (2001) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor, NY, USA: Cold Spring Harbor Laboratory Press.

Sapranaukas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., and Siksnys, V. (2011) The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic Acids Res* **39**: 9275-9282.

Satyanarayana, T., Gowda, S., Ayllón, M.A., and Dawson, W.O. (2004) Closterovirus bipolar virion: Evidence for initiation of assembly by minor coat protein and its restriction to the genomic RNA 5' region. *Proceedings of the National Academy of Sciences* **101**: 799-804.

Schouls, L.M., Reulen, S., Duim, B., Wagenaar, J.A., Willems, R.J.L., Dingle, K.E. et al. (2003) Comparative genotyping of *Campylobacter jejuni* by amplified fragment length polymorphism,

multilocus sequence typing, and short repeat sequencing: strain diversity, host range, and recombination. *Journal of Clinical Microbiology* **41**: 15-26.

Schrenk, M.O., Edwards, K.J., Goodman, R.M., Hamers, R.J., and Banfield, J.F. (1998) Distribution of *Thiobacillus ferrooxidans* and *Leptospirillum ferrooxidans*: Implications for Generation of Acid Mine Drainage. *Science* **279**: 1519-1522.

Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B. et al. (2011) Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. *Proceedings of the National Academy of Sciences* **108**: 10098-10103.

Sherr, E., and Sherr, B. (2002) Significance of predation by protists in aquatic microbial food webs. *Antonie van Leeuwenhoek* **81**: 293-308.

Shibata, A., Goto, Y., Saito, H., Kikuchi, T., Toda, T., and Taguchi, S. (2006) Comparison of SYBR Green I and SYBR Gold stains for enumerating bacteria and viruses by epifluorescence microscopy. *Aquatic Microbial Ecology* **43**: 223-231.

Simmons, S.L., DiBartolo, G., Deneff, V.J., Goltsman, D.S.A., Thelen, M.P., and Banfield, J.F. (2008) Population Genomic Analysis of Strain Variation in *Leptospirillum* Group II Bacteria Involved in Acid Mine Drainage Formation. *PLoS Biol* **6**: 1427-1442.

Sorokin, V.A., Gelfand, M.S., and Artamonova, I.I. (2010) Evolutionary dynamics of clustered irregularly interspaced short palindromic repeat systems in the ocean metagenome. *Applied and Environmental Microbiology* **76**: 2136-2144.

Staley, J.T., and Konopka, A. (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Reviews in Microbiology* **39**: 321-346.

Stang, A., Korn, K., Wildner, O., and Überla, K. (2005) Characterization of virus isolates by particle-associated nucleic acid PCR. *Journal of Clinical Microbiology* **43**: 716-720.

Stern, A., and Sorek, R. (2011) The phage-host arms race: Shaping the evolution of microbes. *BioEssays* **33**: 43-51.

Stocker, R. (2012) Marine Microbes See a Sea of Gradients. *Science* **338**: 628-633.

Suttle, C.A. (2005) Viruses in the sea. *Nature* **437**: 356-361.

Suttle, C.A. (2007) Marine viruses - major players in the global ecosystem. *Nat Rev Micro* **5**: 801-812.

Thingstad, T., and Lignell, R. (1997) Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquatic Microbial Ecology* **13**: 19-27.

Tringe, S.G., von Mering, C., Kobayashi, A., Salamov, A.A., Chen, K., Chang, H.W. et al. (2005) Comparative Metagenomics of Microbial Communities. *Science* **308**: 554-557.

Tyson, G.W., and Banfield, J.F. (2008) Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ Microbiol* **10**: 200-207.

Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M. et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37-43.

van der Oost, J., Jore, M.M., Westra, E.R., Lundgren, M., and Brouns, S.J.J. (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends in Biochemical Sciences* **34**: 401-407.

Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A. et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66-74.

Weinbauer, M.G. (2004) Ecology of prokaryotic viruses. *FEMS Microbiology Reviews* **28**: 127-181.

Weinbauer, M.G., and Höfle, M.G. (1998) Significance of Viral Lysis and Flagellate Grazing as Factors Controlling Bacterioplankton Production in a Eutrophic Lake. *Applied and Environmental Microbiology* **64**: 431-438.

Weinbauer, M.G., and Rassoulzadegan, F. (2004) Are viruses driving microbial diversification and diversity? *Environ Microbiol* **6**: 1-11.

Wiedenheft, B., van Duijn, E., Bultema, J.B., Waghmare, S.P., Zhou, K., Barendregt, A. et al. (2011) RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. *Proceedings of the National Academy of Sciences* **108**: 10092-10097.

Wilmes, P., Simmons, S.L., Denef, V.J., and Banfield, J.F. (2008) The dynamic genetic repertoire of microbial communities. *FEMS Microbiology Reviews* **33**: 109-132.

Wilson, G.G., and Murray, N.E. (1991) Restriction and modification systems. *Annual review of genetics* **25**: 585-627.

Yelton, A.P., Thomas, B.C., Simmons, S.L., Wilmes, P., Zemla, A., Thelen, M.P. et al. (2011) A semi-quantitative, synteny-based method to improve functional predictions for hypothetical and poorly annotated bacterial and archaeal genes. *PLoS Computational Biology* **7**: e1002230.

Zdobnov, E.M., and Apweiler, R. (2001) InterProScan, an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**: 847-848.

Zegans, M.E., Wagner, J.C., Cady, K.C., Murphy, D.M., Hammond, J.H., and O'Toole, G.A. (2009) Interaction between bacteriophage DMS3 and host CRISPR region inhibits group behaviors of *Pseudomonas aeruginosa*. *Journal of Bacteriology* **191**: 210-219.

Zehr, J.P., and Ward, B.B. (2002) Nitrogen Cycling in the Ocean: New Perspectives on Processes and Paradigms. *Applied and Environmental Microbiology* **68**: 1015-1024.

Zheng, Y., Roberts, R.J., and Kasif, S. (2004) Identification of genes with fast-evolving regions in microbial genomes. *Nucleic Acids Res* **32**: 6347-6357.