UNIVERSITY OF CALIFORNIA
RIVERSIDE

Examining the Role of Horizontal Gene Transfer on the Evolution of CRISPR-Cas

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Evolution, Ecology, and Organismal Biology

by

Derek Miles O'Meara

September 2018

Dissertation Committee:
Dr. Leonard Nunney, Chairperson
Dr. Joel Sachs
Dr. Mark Springer

The Dissertation of Derek Miles O'Meara is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

Acknowledgements

I would like to acknowledge Leonard Nunney for his invaluable aid throughout my PhD, the UCR Biology department for their generous offering of both the Newell and Loomer awards, and my family for their continued support.

ABSTRACT OF THE DISSERTATION

Examining the Role of Horizontal Gene Transfer on the Evolution of CRISPR-Cas

by

Derek Miles O'Meara

Doctor of Philosophy, Graduate Program in Evolution, Ecology, and Organismal Biology
University of California, Riverside, September 2018
Dr. Leonard Nunney, Chairperson

CRISPR-Cas is a widespread bacterial genomic defense system characterized by its unique ability to "remember" nucleic acid sequences from invasive pathogens and, through targeted destruction of these sequnces, provide future protection against them. CRISPR-Cas is found to have variable presence/absence throughout much of the bacterial Kingdom, even at the species level. This variability is presumably caused by a mix of gain events, by which the system is passed horizontally from one bacterial genome to another, and loss events, by which CRISPR-Cas is removed from the bacterial genome driven either by a loss of its selective advantage or by active selection against the system. To further understand the evolution of CRISPR-Cas, this research was broken into three distinct chapters.

In the first chapter, a single bacterial species (*Pseudomonas psychrotolerans*) found to show variability in CRISPR-Cas presence was analyzed for evidence of both horizontal transfer and loss of the system by comparison of the bacterial phylogeny to the CRISPR-Cas phylogeny and through a search for recombination sites surrounding the CRISPR-Cas loci. Evidence indicated that there were multiple, independent losses of the CRISPR-Cas system from these strains of bacteria, potentially due to human driven changes in their environments. Further,

homologous recombination was found to be responsible for multiple independent horizontal transfers of CRISPR-Cas between the related genomes.

The second chapter followed the next logical step in zooming out to the level of the *Pseudomonas* genus in search of recombination of CRISPR-Cas at the intraspecies and interspecies levels. It was found that while intraspecies recombination of CRISPR-Cas was prevalent (as seen in the first chapter), interspecies horizontal transfer appeared to be a rare, founder-like process.

Branching away from these phylogenetic approaches, the third chapter focuses on identifying whether the CRISPR-Cas system imposes a cost on its bacterial genome by acting as a barrier towards the entry of potentially beneficial DNA. Through a phylogenetically constrained pairwise analysis of CRISPR-Cas present and CRISPR-Cas absent strains of the same species from throughout the bacterial kingdom, it was found that the barrier hypothesis was supported: strains with CRISPR-Cas had significantly fewer plasmids.

Table of Contents

List of Tables

List of Figures

General Introduction

The CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) CAS (CRISPR ASsociated proteins) system, first characterized as a prokaryotic immune system in 2005 (Mojica et al., 2005), is known for providing bacteria and archaea with viral protection in the form of adaptive immunity (Makarova et al., 2006). It acts by storing nucleic acid sequences taken from any nucleic acids entering the cell (e.g. bacteriophage) and uses this stored sequence to recognize and degrade the foreign material should it reappear in the future (Barrangou et al., 2007).

The CRISPR-CAS system is composed of three parts, the CRISPR cassette, the leader sequence, and the CAS genes (Figure A). The CRISPR cassette consists of an array of a variable number of spacers, each roughly 30 bp in length and flanked by similarly sized repeats. Repeats found at a given CRISPR locus are typically identical to one another while the spacers are unique and match with DNA of exogenous sources such as of bacteriophages and parasitic, high copy number plasmids (Mojica et al., 2005; Godde and Bickerton, 2006). New spacers are added sequentially at the end of the CRISPR cassette closest to the leader sequence (Rezzonico et al., 2011). The leader sequence is an AT rich sequence of one hundred to several hundred base pairs in length found adjacent to the CRISPR cassette and is believed to be involved with the insertion of new spacers (Karginov and Hannon, 2010; Díez-Villaseñor et al., 2013). Finally, the CAS represent an assortment of genes (between four to nine, depending on the specific CRISPR-CAS system) involved with both the inoculation of new spacers and the degradation of exogenous sources matching an existing spacer.

While much is now known about the mechanism of how the CRISPR-CAS system targets and destroys / silences exogenous genetic material, there is still much to uncover regarding the evolutionary history of the system. CRISPR cassettes have only been found in some 40% of

sequenced bacterial genomes (Grissa et al., 2007), and estimates attempting to control for sampling biases in the genomes sequenced suggest the number may be as low as 10% in bacteria (Burstein et al., 2016). Further, CRISPR-CAS's distribution does not follow a clear phylogenetic pattern of vertical transmission (Haft et al., 2005). This begs the question, "Why do some bacteria have CRISPR-CAS while others do not?"

To answer this question, we must be aware of the processes by which the variation in CRISPR-CAS presence is caused: gain and loss. Loss events represent the system being removed from the bacteria either due to selection against the system or loss of selection preserving the system, and gain events represent the transferring of functional CRISPR-CAS units between bacterial populations through the process of horizontal gene transfer. However, there are no detailed studies of these patterns within bacterial phylogenies. Within this dissertation, this problem has been approached at two levels. First, the impact of CRISPR-CAS HGT, vertical transmission, and loss at the *Pseudomonas psychrotolerans* species level (chapter 1) was identified. Second, the analysis was expanded by identifying CRISPR HGT at the genus level within *Pseudomonas* (chapter 2). Though the dissertations begins by analyzing trends of CRISPR dispersal to identify evolutionary patterns, it ends with a pairwise analysis linking CRISPR-CAS presence with a reduction in both plasmid DNA and prophage DNA taken in by the bacteria to better characterize the potential costs of maintaining a CRISPR-CAS system (chapter 3).

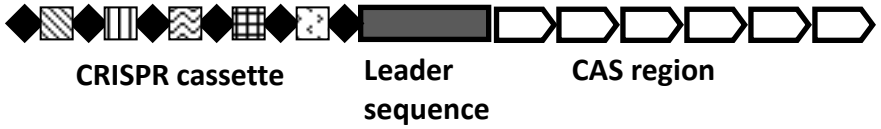

Figure A: The layout of a typical CRISPR locus consisting of the cassette, the leader sequence, and the CAS region. The CRISPR cassette is shown here with diamonds representing the identical repeats while the patterned boxes represent unique spacers that are used to target foreign DNA. (Regions not to scale).

Chapter 1

Gain, loss and recombination of CRISPR-CAS within *Pseudomonas psychrotolerans*

Abstract

Pseudomonas psychrotolerans, a bacterium found naturally in rice fields, has variable distribution of CRISPR-CAS components (the cassettes, leader sequences, and CAS genes) across the 16 available genomes. Eleven (69%) of the strains carried a type 1F-associated leader sequence and cassette (locus 1) that defined 3 groups based on the sharing of spacers; however, only 1 strain had the CAS genes required for CRISPR-CAS activity. This strain had two additional 1F CRISPR loci that were present in a degraded form (lacking the leader sequences) in 4 other strains. Five strains had no CRISPR-CAS components, one of which appeared to have resulted from the loss of a CRISPR locus. Relative to a bacterial phylogeny based on sequence close to CRISPR locus 1 as a reference, the CRISPR leader strand phylogeny was discordant due to three short (<1kb) homologous recombination events (with different breakpoints) between phylogenetically separated strains involving the CRISPR leader and/or cassettes. Our results suggest that, at least in this species, homologous recombination among existing CRISPR loci is common relative to the origination of new loci. Our analysis also revealed five recent losses of the CAS proteins, at least three of which involved homologous recombination. We propose that these recent losses may have resulted from the widespread use of heavy metal fungicides and/or the accumulation of other contaminant metals in rice fields, leading to selection for the loss of CRISPR-CAS function to enable uptake of resistance genes.

Introduction

Study of the CRISPR-CAS system has revealed multiple horizontal gene transfer (HGT) events across diverse clades of bacteria (Godde and Bickerton, 2006; Haft et al., 2005; Horvath et

al., 2008). Little has been done, however, in characterizing the relative frequency of CRISPR-CAS HGT and its effect on the distribution of the CRISPR-CAS system relative to vertical transmission and loss.

Evaluating the relative importance of CRISPR-CAS HGT in determining its prevalence among strains of the same bacterial species is an important first step for understanding how HGT and selection on the bacterial host dictate the distribution of CRISPR-CAS. Selection could promote CRISPR-CAS retention in a bacterial strain due to the benefits of immunity, or it could promote its loss due to the costs of CRISPR-CAS. Such costs could be direct costs of maintenance, or the indirect costs of limiting the influx of novel, beneficial genetic material. For example, there is evidence that the acquisition of antibiotic resistance may be inhibited by CRISPR-CAS (Palmer and Gilmore, 2010; Jiang et al., 2013).

In a scenario in which intraspecies HGT of CRISPR-CAS is high (which in the extreme implies that it acts like an infectious element), the effects of selection would be largely overwhelmed by HGT, although the absence of CRISPR-CAS from some strains could indicate selection acting against the CRISPR-CAS. On the other hand, if there is a slow rate of intraspecies HGT then the fate of CRISPR-CAS becomes closely linked with the fitness of its host, with presence indicating that selection is playing a role in its retention, while absence suggests either that some conditions disfavor bacteria with CRISPR-CAS or that selection favoring bacteria with CRISPR-CAS is weak.

The rate and nature of CRISPR-CAS HGT depends upon the mechanisms involved. The identification of CRISPR-CAS on mega-plasmids (Iacobino et al., 2013) offers a potential explanation for how this system may be moving horizontally; however, it is unknown if this or other modes of HGT predominate. By adding the framework of a bacterial phylogeny to identify

HGT events, we hope to identify the specific evolutionary events leading to the distribution of CRISPR-CAS at the level of a single species.

The study required the choice of a bacterial species polymorphic for presence/absence of CRISPR-CAS. In addition, the bacterial strains needed to show variability in spacer composition among those strains with CRISPR cassettes. Spacers define the target DNA of the CRISPR-CAS, and the presence of spacer variability ensures that the CRISPR-CAS has been functioning independently across strains (Zhang and Ye, 2017). *Pseudomonas aeruginosa* strains are known to vary in the presence of CRISPR-CAS (van Belkum et al., 2015); however, we wanted to choose a species where the strains available were not biased by clinical sampling. We found that another member of the genus, *P. psychrotolerans*, was also variable for the presence of CRISPR-CAS. *P. psychrotolerans* is a gram negative bacterium with strains primarily sequenced from rice fields in India (Midha et al., 2016).

In analyzing the distribution of CRISPR-CAS in *P. psychrotolerans*, five questions were asked:

1. What is the distribution of CRISPR-CAS among strains?
2. Can vertical transmission of the CRISPR-CAS system combined with occasional loss account for the patterns observed or is it necessary to invoke HGT?
3. Given HGT, does the CRISPR-CAS system move as a unit, or, for example, do different components of the system move separately?
4. Does HGT move CRISPR-CAS associated DNA into novel locations in the genome, perhaps together with larger insertions, or are HGT events generally associated with homologous recombination into a pre-existing CRISPR-CAS locus?
5. Is there evidence for the loss of all or part of the CRISPR-CAS region? Are particular components of the system more frequently lost?

By answering these questions, we can further understand the relative impact of horizontal transfer and vertical transmission on the evolution of the CRISPR-CAS system and begin to develop hypotheses regarding the relative benefits and costs of carrying CRISPR-CAS.

Methods

Bacterial sequence used

The genomes used in the analysis were all of the *P. psychrotolerans* sequenced genomes available on NCBI as of January 1st 2018 (NCBI Resource Coordinators, 2017) that showed less than 99.5% aligned identity (as recorded by NCBI) to any other sequenced genome already chosen, ignoring gaps in the alignment.

Identification of CRISPR cassettes and CAS genes

CRISPR cassettes in *P. psychrotolerans* were identified through the use of the CRISPRFinder program (Grissa et al., 2007). This program identifies CRISPR cassettes by searching for repeats (at least 3) of 20 to 40 bp in length spaced apart by 20 to 40 bp within a bacterial genome. The possibility that degraded cassettes were present at these same locations in other strains was examined by aligning all strains at loci found positive for a CRISPR cassette. Cassettes found in this way had their putative repeat sequences compared to the consensus repeat sequence at that locus using the program FASTA to determine the probability (as a value for expected number of hits) that the similarities between two aligned sequences could have arisen by chance (Pearson and Lipman, 1988).

The CRISPRFinder web tool was also used to search for identifiable CAS genes within 20 kb 5' and 3' of all identified CRISPR cassettes using BLAST. Once a CAS gene was identified, both BLASTn (nucleotide) and BLASTp (protein) were used against the genomes of other *P. psychrotolerans* strains to detect CAS genes that may be located away from the CRISPR-

cassette. Further efforts were made to identify possibly unidentified CAS genes by searching within 20 kb 5' and 3' of known CRISPR cassettes for at least 4 adjacent hypothetical proteins with the same orientation, since 4 is the fewest number of genes as of yet found in any functional CAS (Makarova et al., 2015). CAS type was identified by matching identified CAS genes with known CAS types, each of which is defined by its unique set of CAS genes (Makarova et al., 2015).

Bacterial phylogeny:

A *P. psychrotolerans* phylogeny was created using the 30 kb immediately 5' of the most prevalent CRISPR cassette locus (defined as locus 1), a region that could also be identified in strains lacking locus 1. All genes within this region were identified and concatenated using the NCBI annotations. The concatenated DNA sequences were aligned with CLUSTALW (Larkin et al., 2007) and the maximum likelihood phylogeny and bootstrapping were performed using RAxML-HPC2 on XSEDE tool through CIPRES (Miller et al., 2012). *Pseudomonas citronellolis* strain P3B5 was used as an outgroup. This phylogeny was used to examine the distribution of CRISPR-CAS and, using only strains showing any presence of locus 1, as a reference in detecting HGT within the adjacent CRISPR region.

To detect any lack of concordance between the tree topologies of individual genes within the 30bk region and the tree topology of the concatenated genes, the maximum likelihood tree for each gene was compared to one constrained to have the same topology as the concatenated gene tree. PAUP 4.0* (Swofford, 2003) was used to calculate maximum likelihood scores for both the constrained and unconstrained tree for each individual gene. A one-tailed Kishino Hasegawa test (Kishino and Hasegawa, 1989; Hasegawa and Kishino, 1989) was then used to determine if the maximum likelihood scores for the constrained and unconstrained tree were significantly different from one another. We used a one-tailed analysis because the constrained tree cannot

have a higher maximum likelihood score than the unconstrained tree. The Holm-Bonferroni method was used to control for multiple comparisons and adjust significance values from the original α=.05 (Holm, 1979). Significant results indicate that the gene in question shares a different evolutionary history from the others, due to homologous recombination following HGT into one or more strains, and these genes were removed from the concatenated phylogeny.

CRISPR leader phylogeny concordance with the bacterial phylogeny

A phylogeny was created using the CRISPR leader of the most abundant type of CRISPR-CAS (which in *P. psychrotolerans* is type 1F) across all loci. The leader is immediately adjacent to the 3' end of the CRISPR cassettes. A leader sequence for this type 1F CRISPR-CAS system was identified by Alkhnbashi et al. (2016) in *P. aeruginosa*, and we used this sequence as a scaffold to identify the length of the leader starting at the 3' end of the CRISPR cassettes. CLUSTALW was used to align the sequences and both maximum likelihood and bootstrap analyses were created through the RAxML-HPC2 on XSEDE tool through CIPRES.

The bacterial phylogeny and the leader phylogeny were tested for concordance using PAUP* 4.0. Likelihood scores were compared for the unconstrained bacterial tree and the bacterial tree constrained by the leader phylogenies topology using the Kishino Hasegawa analysis (see above). Significant results indicated that, due to HGT, one or more of the leader sequences had different evolutionary history from the bacterial phylogeny.

Shared spacer analysis

Spacers are derived from foreign DNA and added sequentially (Rezzonico et al., 2011). Spacer composition was used to identify groups of related CRISPR cassettes, since matching spacers are considered to share a single origin given the highly stochastic mechanism by which they are created (Pourcel et al., 2005; Kupczok and Bollback, 2013). Related cassettes were identified as those that had at least a single matching spacer between them.

To find matching spacers, a database of the *P. psychrotolerans* spacers was created and each spacer sequence was compared to all others using BLAST. Using BLAST cutoff of $e < 10^{-10}$ identified matching pairs with up to a single bp differences in the 29 bp long spacers, allowing for the possibility of a single point mutation since the original incorporation of each spacer. The same results were found when using the larger e value cutoff of 10^{-8} which allowed up to 3 mismatching bp.

Introgression site identification

Successful HGT results either in the insertion of a novel sequence creating a new locus or homologous recombination into an existing locus. To identify probable recombination sites within 500 bp 5' and 3' of the CRISPR leader/cassette locus of interest, we used Nunney's Introgression test (Nunney et al., 2012). This test determines if a given region of the genome shares a different evolutionary history from its surrounding regions. All strains with CRISPR were aligned at the locus of interest and all fixed sites were removed, leaving behind only SNP sites. This reduced sequence was used for pairwise comparisons, where each site is classified as either "different" (D) or "same" (S) between the two strains. We were interested in identifying regions more similar than expected, based on the surrounding genome, indicating HGT between the ancestors of the pair of strains.

For our analysis, we were only interested in those possible recombination regions that spanned at least part of the CRISPR leader and/or cassette region. We defined possible 5' recombination boundaries by detecting stretches of consecutive S sites that were longer than we would expect to see (with a probability of <0.05) given the ratio of S to D sites spanning from the 5' edge of the analysis up to that point. Given a proportion of S sites in the range 0.37 – 0.65, for $0.37<S<0.47$, four consecutive S sites are needed; for $0.47<S<0.55$, five consecutive S sites are needed; for $0.55<S<0.61$, six consecutive S sites are needed; and from $0.61<S<0.65$, seven

consecutive S sites are needed. Once a 5' potential boundary is identified, the same is done from the 3' direction to find a potential 3' end for the recombination boundary. We next repeat this same process from between the two potential boundaries, starting from the 3' end, but instead using D sites as the reference to determine if the 5' boundary need be extended. Given a proportion of D sites in the range 0 – 0.14, if $0<D<0.04$, two D sites within five adjacent sites is enough to define a boundary; if $0.04<D<0.1$, two D sites within four is enough; and if $0.1<D<0.14$, two D sites within three is enough. This is repeated from the 5' end to the 3' end to determine if the 3' boundary should be extended.

The region between these two final boundaries is considered the potential recombination region. The ratios of D and S sites on either side of a potential breakpoint (stretching until the 500 bp boundary or until the next potential breakpoint) were tested for equality using the introgression test (Nunney et al., 2012). Rejection of the null hypothesis provided is evidence of a recombination breakpoint at the tested site. For our analysis, we were only interested in those possible recombination regions spanned over the CRISPR leader and/or cassette region.

If more than two strains are being tested for a region being shared through recombination, rather than D sites, "polymorphic" (P) sites will be used to describe those bases that are polymorphic with at least one of the strains having a base also found in at least one of the reference strains. Other combinations are ambiguous and not scored. The same method applies for detecting and testing the recombination boundaries, only with P sites replacing D sites.

Generally, the cassette region cannot be used in the introgression analysis given the potentially rapid turnover of spacers and their adjacent repeats. However, the most ancestral repeat within a cassette (which is farthest from the leader sequence) often contains an accumulation of mutations relative to the repeat consensus based on the more recent repeats (Horvath et al., 2008), suggesting that it is conserved (i.e. it predates its adjacent spacer). The

ancestral repeats were considered sufficiently conserved to be used in the introgression analysis if, across the strains, they shared unique sequence features not found in the other repeats.

Results

P. psychrotolerans CRISPR loci

Sixteen fully sequenced unique *P. psychrotolerans* genomes were identified, and 3 different CRISPR cassette loci were located within them (labeled 1-3, see Figure 1.1; Table 1.1). CRISPR cassettes at locus 1 were the most prevalent and were found, together with a ~130 bp leader sequence at their 3' end, in 11 of the 16 strains. Five of these strains (SB11, SB18, NS201, NS274, and NS383), also carried both locus 2 and locus 3; however, in all but SB11, these cassettes were highly degraded, with their repeat consensus showing only about 70% similarity with the SB11 (based on a FASTA analysis) and their leader sequences were absent. SB11 had an intact leader sequence at both locus 2 and locus 3 (Figure 1.1). The degraded CRISPR cassettes at locus 2 and locus 3 were not identified by the CRISPRFinder program as they had too many mutations between the repeats, nor was the locus 1 CRISPR cassette of PRS08 since it contained too few spacers. The remaining 5 strains had no identifiable CRISPR cassettes, and *P. citronellolis* P3B5, a close relative used as the outgroup for the bacterial phylogeny, had no evidence of CRISPR-CAS anywhere within its genome.

The CRISPR-CAS system needs functional CAS genes along with the CRISPR cassette and leader sequence to actively protect the bacterial genome and to gain new immunities (Vorontsova et al., 2015). However, SB11 was the only strain to have any CAS genes. This CAS cluster was located between the locus 2 and locus 3 CRISPR leaders (Figure 1.1) and included the 6 CAS genes (Cas1, Cas2-Cas3, csy1, csy2, csy3, and cas6f) typical of the type 1F CAS system (Makarova et al., 2015).

In the 10 strains with locus 1 cassettes but no CAS genes, two different genomic architectures were found at the locus 2/CAS/locus 3 region. As noted above, strains NS274, NS383, SB18, and NS201 were all found to have two small degraded locus 2 and 3 cassettes (and no leaders) on opposite strands immediately adjacent to each other at this region, with the entire CAS region absent (Figure 1.1). On the other hand, strains PRS08, SB5, RSA46, SB14, NS376, and SB8 were all missing the entirety of the locus 2 and 3 CRISPR cassettes, leaders, and CAS genes (Figure 1.1).

Bacterial phylogeny

A phylogeny of the 16 *P. psychrotolerans* strains was constructed using the coding regions within 30kb 5' of the CRISPR locus 1 (Figure 1.1). There are 25 annotated genes within this 30kb region, and each was tested for concordance with the phylogeny of the concatenated coding region. Four were found to be discordant, likely due to HGT at these loci. The first of these 4 discordant genes was the closest to the CRISPR locus and was later identified as having multiple recombination regions extending into it (see below). The other three genes were consecutively the 12th, 13th, and 14th farthest genes from locus 1. The differing topology of these 3 genes appears to have been caused by strains NS274 and NS383, which are sisters in the consensus, being more distantly related to each other over this region. The four discordant genes were removed from the analysis. The resulting bacterial phylogeny (Figure 1.2) was topologically identical to the original phylogeny created from all 25 genes.

Of the five strains lacking any CRISPR-CAS loci, one (NS2) was found to be closely related to strains carrying the locus 1 CRISPR (Figure 1.2), with the genomic data showing a complete deletion of the locus and its replacement with approximately 4kb of unique sequence (Figure 1.1). The remaining 4 strains lacking locus 1 formed their own clade basal to the rest

(Figure 1.2) and, given that the tree was rooted with *P. citronellolis* P3B5 that also lacked any sign of CRISPR-CAS loci, there is no evidence that these strains ever contained locus 1.

Leader phylogeny concordance with the bacterial phylogeny

There were six different leader sequences identified, one each at locus 2 and locus 3 in strain SB11, and four across the 11 strains with locus 1. The phylogeny of the leader sequence had strong bootstrap support for all nodes (Figure 1.3), with the leaders from locus 2 and 3 distantly related to those from locus 1: locus 1 shared 47% sequence identity with locus 2 and 55% sequence identity with locus 3; while locus 2 and 3 shared 60% identity. For comparison, the lowest sequence identity found between 2 strains of locus 1 was 91%.

Concordance between the bacterial phylogeny (with strains lacking leaders omitted) and the locus 1 leader phylogeny was tested using the Kishino Hasegawa analysis. The null hypothesis of concordance was rejected ($p=0.0002$), indicating the occurrence of HGT in the leader of locus 1. This result is primarily due to the identical leaders in the phylogenetically separated strains of SB11 and NS201/SB18, and, similarly, in PRS08 and NS383/NS274 (Figure 1.2).

Shared spacers

The number of spacers found at the three CRISPR loci ranged from 1-5 at locus 1, and, in SB11, 7 at locus 2 and 9 at locus 3 (Figure 1.3). We confirmed that those spacers found within the degraded CRISPR cassettes at locus 2 and 3 did not match any other spacers. Notably, they did not match those from the complete cassettes at locus 2 and 3 in SB11. However, due to the heavy degradation of their repeats and the lack of any leader sequence near them, we did not include them in the spacer analysis. There was no sharing of spacers across loci (Figure 1.3).

There were 16 unique spacers found at CRISPR locus 1. Of these 16, 13 were shared between at least two strains, and only 3 were unique to a single strain (SB11). Based on shared

spacers at locus 1 and the unique spacers at locus 2 and locus 3, five independent groups were apparent (A-E, see Figure 1.3). Members of group A all shared at least 1 spacer, as did the members of group C, while the 5 members of group B shared at least 4 spacers. Moreover, all matching spacers were found in the same order, consistent with a shared origin. However, none of these groups form a monophyletic clade when mapped onto the bacterial phylogeny (see shape symbols in Figure 1.2). Further, group B does not form a monophyletic clade within the leader phylogeny (Figure 1.3). These patterns indicate that there were one or more recombination events involving the CRISPR locus 1 cassette region and that the leader and cassette did not always move in tandem.

Introgression analysis of CRISPR locus 1

To detect HGT between strains that involved CRISPR locus 1 leaders/cassettes, we tested a region that stretched 500bp on either side of the leader sequence using Nunney's introgression test (Nunney et al., 2012). The majority of the cassette was excluded from the analysis due to its dynamic turnover, but the locus 1 ancestral repeat (furthest from the leader) was included. This repeat showed strong evidence of shared ancestry, marked by the presence of a 4bp deletion across all strains which was not found in any other repeat. Within the ancestral repeat, there were two SNP sites not shared between all strains that were used in the analysis (see cassette region, Figure 1.4).

Three phylogenetic inconsistencies were examined, one within each of the three spacer groupings (A-C) at locus 1. Each group contained at least one strain that was phylogenetically distant from the others (Figure 1.2): in group A, SB11 was phylogenetically distant from SB18 and NS201; in group B, SB5 and RSA46 are phylogenetically distant from SB14, SB8, and NS376; and in group C, PRS08 is phylogenetically distant from NS383 and NS274. In all 3 cases, the introgression analysis revealed a region of significantly greater sequence similarity of

the otherwise distantly related strains that overlapped the CRISPR cassette and/or the leader sequence (rectangles in Figure 1.4) when compared to both their 5' and 3' adjacent regions (Table 1.2). This pattern is strong evidence of three independent recombination events resulting from HGT, noting that none of the three recombinant sequences found within the tests shared any recombination breakpoints with the others (Figure 1.4). Furthermore, the sequence 5' and 3' of the recombined region showed the same pattern of similarity (Table 1.2), indicating that they shared the same evolutionary history, further supporting the view that there was a single recombination event between them. It can also be seen from Figure 1.4 that there was a small region of about 50bp at the 3' end of the leader sequence in the group B comparison suggestive of a recombination. Testing of the potential breakpoints surrounding the sequence of 7 "S" sites showed that, while the 3' breakpoint was statistically significant ($P<0.01$), the 5' breakpoint was not. Notably, the sites between this region and the larger upstream recombination showed no indication of heightened similarity, indicating that the 50bp region was not an extension of the upstream recombination.

The recombined regions varied in their relationship to the CRISPR locus (Figure 1.4). The group A exchange was around 462 bp and involved the whole of the CRISPR locus, as did the small group C exchange of about 223 bp. However, in the group B exchange of about 218 bp, the 3' end breakpoint was between the cassette and the leader sequence (Figure 1.4), demonstrating that the CRISPR cassette and the leader sequence do not always move together.

The direction of the exchanges was investigated by comparing the bacterial phylogeny to that of the recombined region. For group A, the leader phylogeny indicated that the exchanged leader sequence was basal, assuming the much more distantly related group D and E leader sequences are out-groups (Figure 1.3). This result is consistent with the position of SB11 in the bacterial phylogeny (Figure 1.2), indicating the introgression of sequence from SB11 into the

lineage of SB18 and NS201. The direction of the introgression of the group C sequence is ambiguous, given the two trees, and the group B exchange between SB5/RSA46 and SB14/NS376/SB8 did not involve the leader sequence.

Detecting loss and recombination at the CAS region

Only a single strain of *P. psychrotolerans* was found with CAS genes, raising the question of whether or not this loss represented a single or multiple loss events. This problem can be approached in two ways. First, the presence of unique vs. shared spacers can be used to determine how recently a cassette has been acquiring new spacers, as expected of a functional CRISPR-CAS system (Zhang and Ye, 2017), and second, the genomic architecture around the excised genes can be used to test the adequacy of a single loss hypothesis.

The strain SB11 is the only strain found with CAS genes, and, consistent with a functioning CRISPR-CAS, has a set of 3 unique spacers (Figure 1.3). It also shares an ancestral spacer with the two other group A strains that must have been present at the time of the HGT that homogenized their locus 1 cassette and leader, believed to be from SB11 to the ancestor of the other two strains (Figure 1.4); however, since that transfer occurred, the recipient (the ancestor of NS201/SB18) accumulated 4 new spacers, indicating continued CRISPR-CAS activity. Furthermore, the completely different spacer makeup between group A and the two other spacer groups B and C (Figure 1.3) is indicative that CRISPR-CAS was functioning in both group B and C for at least some time following the CRISPR cassettes' divergence from one another. These data indicate at least 3 independent losses of the CAS genes.

As noted above, in those bacterial strains with CRISPR cassettes and leaders present at locus 1 but no CAS genes, two different genomic architectures were found. Architecture 1, which included strains with cassettes from group A and C, retained the (now degraded) cassettes of locus 2 and locus 3, while architecture 2, which included strains with cassettes from group B and

C, involved the complete deletion of these loci (Figure 1.1). The involvement of group C in both architectures indicates a fourth event in the CAS region. The most parsimonious explanation for this pattern seems to be that (a) the CAS region and the leaders of locus 2 and 3 (architecture 1) were lost from the ancestor of the group A strains SB18/NS201 but that this deletion was later transferred via HGT to the ancestor of group C strains NS274/NS383 (or vice versa), and that (b) the CAS region and all of loci 2 and 3 (architecture 2) was first lost from the ancestor of either group B clade SB14/SB8/NS376 or clade SB5/RSA46, second was transferred via HGT to the other clade, and third was later transferred via HGT to the group C strain PRS08. This hypothesis requires five events that result in the loss of the CAS genes.

To test this hypothesis, we searched for evidence of HGT between the ancestors of NS274/NS383 and of SB18/NS201 (architecture 1), between the two clades of the group B strains (architecture 2), and between PRS08 and one or more of the group B strains (architecture 2). In the case of architecture 1, we first looked for potential recombination breakpoints 500 bp 5' of locus 2 and 500 bp 3' of locus 3. The 5' region was extended an extra 127 bp to the 3' end of the locus 1 leader sequence as a potential recombination boundary was found to extend past the first 500 bp. The 634 bp region of the degraded CRISPR cassettes of locus 2 and locus 3 was not included as it was absent from most strains; however, the sequence is almost identical between the four strains with architecture 1 (other than a missing 60 bp region from the NS274/NS383 taxa and 5 polymorphic sites between NS274/NS383 and SB18/NS201). A recombination event linking the four architecture 1 strains was found to be 950 bp in length with boundaries roughly 300 bp 5' of locus 2 ($p<0.001$) in the ~1kb region shown in Figure 1.1 and within 22 bp of the 3' end of locus 3 ($p<0.001$) (Table 1.3; Figure 1.5). This finding supports the hypothesis that the phylogenetically distant NS274/NS383 and NS201/SB18 share the same genomic architecture at the site of their missing CAS genes through recombination rather than through independent

events. Furthermore, the presence of heavily degraded CRISPR cassettes at loci 2 and 3 provides evidence that the loss of CAS in this piece of DNA may be older than the loss from any of these 4 strains (as measured by the intact state of the locus 1 cassette), suggesting that the inserted region may have originated from elsewhere and was inserted into one of these lineages and then transferred to the other.

Architecture 2, in which the entirety of the locus 2 and locus 3 regions is missing, was found amongst all 5 Group B strains and the Group C PRS08 strain. First, we looked for a recombination region between the 5 Group B strains (ignoring the group C strain, PRS08). A recombinant region roughly 170 bp long spanning from 114 bp 5' of locus 2 ($p<0.001$) and 50 bp 3'f of locus 3 ($p<0.001$) was identified between these 6 strains (Architecture 2B column, Table 1.3; Figure 1.5). We next looked for recombination that could explain the similarities between the Group C strain (PRS08) and the Group B strains by comparing PRS08 to SB5/RSA46, the group B strain(s) with which it shared the most similarities at this region. A recombinant region approximately 540 bp long spanning from what appears to be the same 5' site as the previous event, 114 bp 5' of locus 2 ($p<0.001$), but spanning to a novel breakpoint 420 bp 3' of locus 3 ($p<0.01$), was identified between these strains (Architecture 2C column, Table 1.3; Figure 1.5). These results support the hypothesis that these phylogenetically unrelated strains share a common genomic architecture surrounding the missing CAS genes due to recombination between strains rather than independent loss events. Thus, in summary, there appears to have been an initial deletion, followed by a recombination event that occurred between strains of group B leading to a shared architecture characterized by a complete absence of locus 2, locus3, and the CAS genes. This event was then followed by a separate transfer from the SB5/RSA46 ancestor into PRS08 (with a longer 3' end).

Finding the precise location of the 3' breakpoint for genomic architecture 1 surrounding the missing CAS is limited by a lack of polymorphic sites within the first 23 bp of the region 3' of locus 3. However, this highly conserved region provides a scaffold by which exogenous DNA might bind and begin the recombination process.

Our results support the hypothesis that there have been five independent losses of the CAS region from *P. psychrotolerans*, with evidence for at least 2 independent losses either by deletion or recombination being found within the first genomic architecture analyzed and 3 for the second genomic architecture, with at least 3 and probably 4 of these events involving loss through recombination.

Discussion

While many other studies have characterized CRISPR-CAS presence/absence within a given species (Palmer and Gilmore, 2010; van Belkum et al., 2015; Delaney et al., 2012; Touchon et al., 2011) or among clades of the bacterial kingdom (Burstein et al., 2016), this is the first study to document both the distribution of CRISPR-CAS within a species and its links to specific HGT events. Analysis of 16 fully sequenced genomes of *P.* psychrotolerans revealed a single CRISPR-CAS system with the locus 1 CRISPR present in 11 of them plus all or part of two other CRISPR loci found in 5 of these 11 strains. There was clear evidence of locus 1 CRISPR leaders and cassettes moving horizontally among strains with multiple HGT events necessary to explain the discordance between the bacterial and CRISPR locus 1 leader phylogenies and the distribution of spacers in the locus 1 CRISPR cassettes, plus similarly clear evidence of both deletion and HGT being involved in the loss of CAS genes from 10 of the 11 strains that possess the locus 1 CRISPR.

While 3 CRISPR loci were found within *P. psychrotolerans*, only one strain contained the three complete loci plus the required CAS genes. Four other strains contained the locus 1 CRISPR plus degraded versions of the locus 2 and 3 cassettes, and six others contained only locus 1, but these 10 strains lacked any members of the expected cluster of CAS genes.

Three recombination events, involving 350 – 1000 bp, were identified involving the locus 1 CRISPR cassette and/or leader. All three of these recombinant events had unique 5' and 3' breakpoints (Figure 1.4). Two of these events included both the cassette and leader of locus 1; however, the third involved only the cassette. This third event demonstrates that CRISPR cassettes may recombine with and replace an existing CRISPR cassette, so that it cannot be assumed that CRISPR cassettes and their adjacent leaders have evolved in tandem.

A phylogeny based on the CRISPR leader sequence showed that loci 1, 2, and 3 are distantly related to each other relative to the relationship among strains at locus 1, showing that their divergence was much earlier than the divergence seen within locus 1 (Figure 1.3). This pattern, combined with evidence of HGT involving the locus 1 CRISPR (Figure 1.4), indicates that HGT resulting in homologous recombination is the dominant form of genetic change, while HGT resulting in the formation of novel CRISPR loci is rare.

The finding that the gain of new CRISPR loci is rare raises the reverse question of how commonly CRISPR loci are lost. Degraded cassettes of loci 2 and 3 were found in two strains, probably resulting from a recombination event that replaced the pre-existing cassettes (and deleted both of the leader sequences and the CAS loci); however, this does not provide information on the frequency of such events. At locus 1, where the data are more extensive, the phylogenetic data only show evidence of a single strain (or its recent ancestor) losing its locus 1 CRISPR region. The NS2 strain, which was found to be unambiguously within the clade of the 11 strains carrying CRISPR locus 1, had no CRISPR presence whatsoever (Figure 1.2).

CRISPR cassettes in genomes lacking CAS genes are referred to as orphaned CRISPR cassettes and have been previously found to be common in *Enterococci faecalis* (Palmer and Gilmore, 2010), *Listeria monocytogenes* (Bikard and Marraffini, 2013), and *Aggregatibacter actinomycetemcomitans* (Jorth and Whiteley, 2012). In the present study, orphaned cassettes were the rule. Of the 11 *P. psychrotolerans* strains with at least one CRISPR cassette, only one strain (SB11) had CAS genes present. Analysis of the spacer composition at locus 1 identified three spacer groups that had non-overlapping spacer sequences (Figure 1.3) indicating continued functionality following their divergence, consistent with a minimum of three CAS loss events. Analysis of the genomic architecture in the region where the CAS genes were found in SB11 (between loci 2 and 3; see Figure 1.1) suggested at least two more loss events, and that at least 3 of the losses had resulted from an HGT of sequence lacking the CAS genes that resulted in homologous recombination and hence a deletion of the genes in the recipient (Table 1.3; Figure 1.5).

It is not known what type of HGT leads to the recombination events observed in this species. One possibility is transformation, allowing bacteria to incorporate via homologous recombination DNA from dying neighbors. This possibility is certainly consistent with the finding that changes in CRISPR-CAS were primarily due to small (<1kb) regions of homologous recombination. Another possibility is that CRISPR-CAS related sequence is sometimes carried on plasmids (Iacobino et al., 2013). A BLAST search of known type 1F CAS genes revealed an entire type 1F CRISPR-CAS system within a plasmid of *Vibrio alginolyticus* (Genbank reference: NZ_CP013486.1). Though not from within the same species, this provides support for plasmid transfer as a potential mechanism by which CRISPR-CAS gain and loss within *P. psychrotolerans* could have been facilitated.

The recent widespread independent losses (at least 5) of CRISPR-CAS functionality from *P. psychrotolerans* by CAS loss strongly points towards the possibility of a change in the selective pressure that had previously favored CRISPR-CAS. The relative abundance of the locus 1 cassette (69% of strains) supports the view that there was historically some benefit to retaining CRISPR-CAS, but only 6% (1 strain) now retains CRISPR-CAS function. One hypothesis to account for CRISPR-CAS no longer being beneficial for the bacteria relates to their ecology. The majority of these bacterial strains come from rice fields in India, where the use of fungicides over the past 60 years has sharply increased due to increased worldwide production needs (Prasanna et al., 2013). These fungicides often contain copper and other compounds harmful to bacteria, driving selection within the bacterial populations for resistance to the fungicides. It has been hypothesized that CRISPR-CAS reduces the ability of some bacteria to intake new plasmids (Palmer and Gilmore, 2010; Jiang et al., 2013). Given that resistance genes for traits such as copper tolerance are often found on plasmids (Cooksey, 1990), it is possible that the multiple losses of the CAS genes from *P. psychrotolerans* are due to the adaptive advantage for these bacteria to acquire new beneficial genes through HGT.

We see some evidence for this in PRS08, the only fully assembled strain of *P. psychrotolerans*, with one recognized plasmid (NCBI reference sequence: NZ_CP018759.1) carrying about 160 genes (the vast majority of which are uncharacterized). The benefit derived from this plasmid is unknown, but a BLAST search against the plasmid with the CRISPR spacers found across *P. psychrotolerans* revealed a match (found at both site 14932 and 109174 within the plasmid) with the only spacer found in PRS08 (also found in the other two Group B strains). Though not necessarily evidence for selection against the CRISPR-CAS system, it does indicate that the strain historically targeted this plasmid and that the loss of CRISPR-CAS functionality was necessary for this plasmid to be taken in.

Future studies may identify related ways by which human interaction has had unintended consequences of the evolution of CRISPR-CAS within wild bacterial populations.

Table 1.1. CRISPR loci and CAS genes found in *P. psychrotolerans*

Strain	Locus 1	Locus 2	Locus 3	CAS?
SB11	Yes	Yes	Yes	yes
NS274	Yes	Yes [a]	Yes [a]	No
SB18	Yes	Yes [a]	Yes [a]	No
NS201	Yes	Yes [a]	Yes [a]	No
NS383	Yes	Yes [a]	Yes [a]	No
SB8	Yes	No	No	No
SB14	Yes	No	No	No
NS376	Yes	No	No	No
SB5	Yes	No	No	No
RSA46	Yes	No	No	No
PRS08	Yes [a]	No	No	No
NS2	No	No	No	No
NS337	No	No	No	No
SDS18	No	No	No	No
DSM15758	No	No	No	No
L19	no	No	No	No

a CRISPR cassette was either too degraded or short to be identified via CRISPRFinder but rather was identified through alignment at this locus.

Table 1.2: Introgression test results CRISPR locus 1 comparing group A (SB18/NS201 vs SB11), group B (SB14/NS376/SB18 vs SB5/RSA46) and group C (NS274/NS383 vs PRS08)

Comparison	5' boundary	3' boundary	Approx. rec. region length	5' site ratio	Rec. site ratio	3' site ratio	5' vs rec. region	3' vs rec. region	5' vs 3'
Group A	3290652 - 3290655	3291108 - 3291121	462 bp	23S:27D	60S:0D	20S:12D	0.000***	0.000***	0.600
Group B	3290498 - 3290506	3290681 - 3290759	218 bp	16S:14D	26S :0D	41S:45D	0.000***	0.000***	0.758
Group C	3290650 - 3290652	3290868 - 3290880	223 bp	31S:18D	20S:2D	36S:35D	0.027*	0.001**	0.240

Note: Positions within the genome relate strain PRS08

Significance levels: * $p<0.05$, ** $p<0.01$, *** $p<0.001$

Table 1.3: Introgression test results for the region surrounding CRISPR locus 2 and 3 for genomic architecture 1 (NS274/NS383 vs SB18/NS201), genomic architecture 2B (NS376/SB8 vs SB14 vs SB5/RSA46) and genomic architecture 2C (SB5/RSA46 vs PRS08)

Comparison	5' boundary	3' boundary	Approx. rec. region length	5' site ratio	Rec. site ratio	3' site ratio	5' vs rec. region	3' vs rec. region	5' vs 3'
Architecture 1	3291201 - 3291205	3291503 - 3291526	950 bp	30S:17D	69S:0D	18S:14D	0.001***	0.000***	0.657
Architecture 2B	3291378 - 3291390	3291554 - 3291557	170 bp	11S:22P	49S:2P	7S:22P	0.001***	0.000***	0.758
Architecture 2C	3291378 - 3291390	3291920 - 3291931	540 bp	16S:17D	67S:4D	6S:4D	0.000***	0.004**	0.240

Note: Positions within the genome relate strain PRS08

Significance levels: ** p<0.01, *** p<0.001

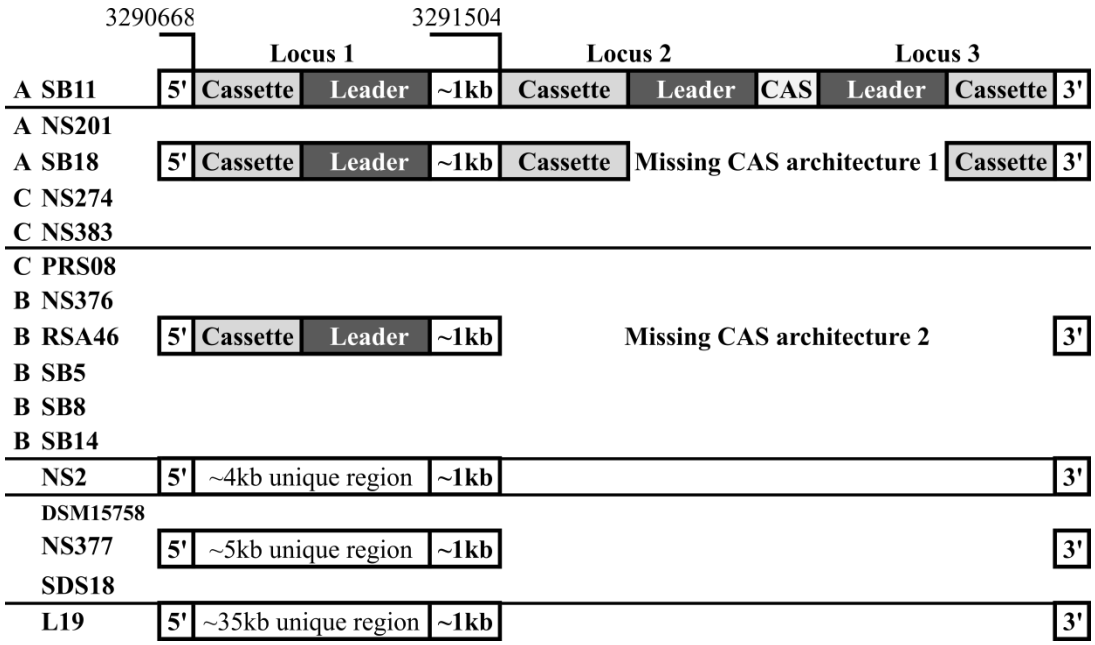

Figure 1.1: Graphical representation (not to scale) of the three CRISPR cassette loci within P. *psychrotolerans.* Letters (A, B, and C) to the left of strain names indicate what spacer group that strain's locus 1 CRISPR cassette belongs to. Arrows indicate the strand on which the CRISPR cassette is located (arrows point away from the ancestral ends of the cassette). Only the PRS08 genome was fully assembled and is used as the genomic reference. The unique regions of differing sizes found at locus 1 within those strains lacking a CRISPR cassette align at their 3' end but not their 5' end.

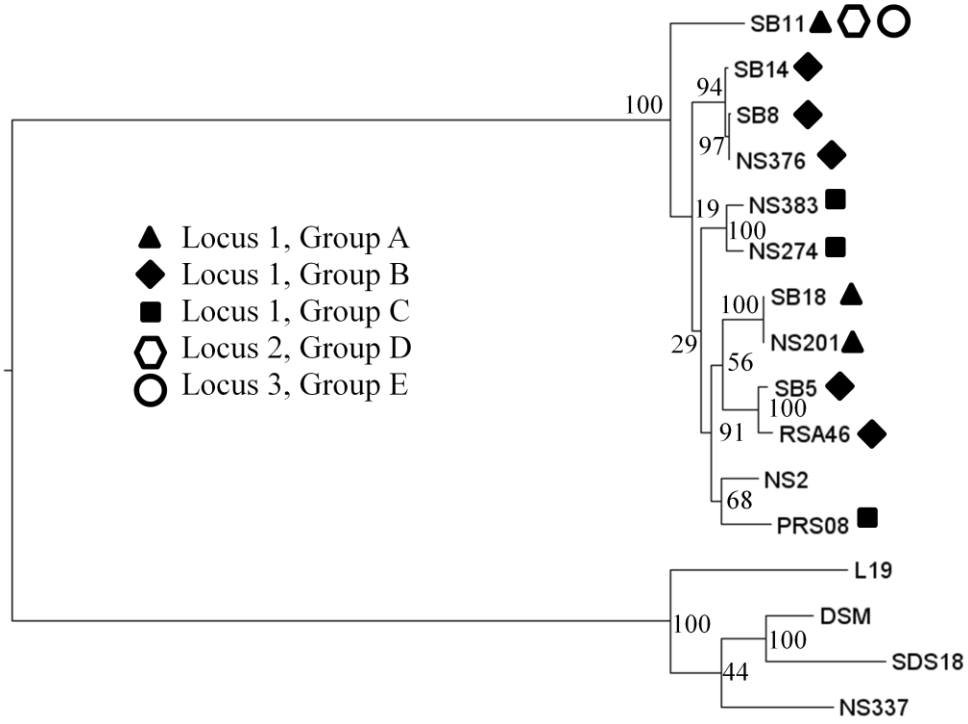

Figure 1.2: *Pseudomonas psychrotolerans* phylogeny of the bacterial genome immediately 5' of the locus 1 CRISPR. It is based on 21 of the 25 coding genes found within the 30 kb region 5' of the CRISPR cassette (the remaining 4 showed a discordant phylogeny due to recombination). This phylogeny is rooted by *Pseudomonas citronellolis*. Shapes shown next to strain names indicate presence of CRISPR cassettes, the locus from which they came, and the spacer group they fall into (see Figure 4). Strains with no shape next to their name had no CRISPR presence.

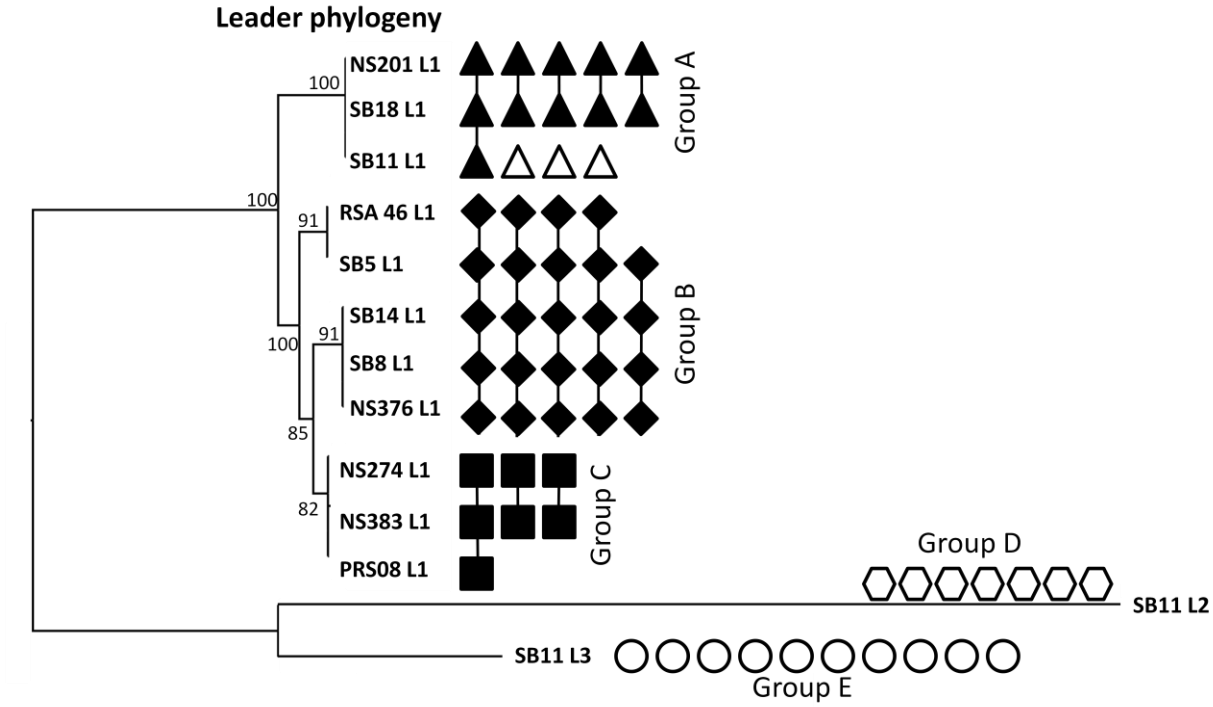

Figure 1.3: Leader phylogeny and spacer distribution from CRISPR loci 1, 2, and 3 of *P. psychrotolerans.* This tree is unrooted. The phylogeny labels indicate strain and leader locus (L1, L2, L3). The shapes adjacent to the leader phylogeny represent the spacers found within the corresponding CRISPR cassette, with the differing shapes representing the different groupings (Group A – E). Groupings are formed when two or more cassettes share at least a single spacer (as indicated by black lines connecting the identical spacers). Black spacers have a matching spacer found within another strain while white spacers have no matches. Spacers are shown in sequential order with the oldest on the left.

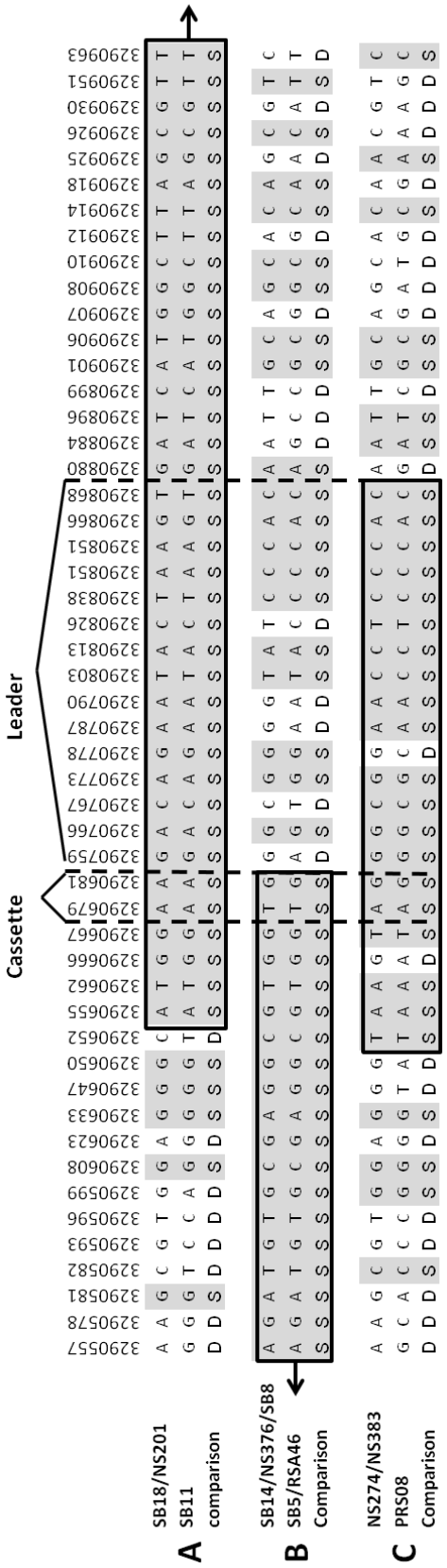

Figure 1.4: Recombination events at CRISPR locus 1, showing all polymorphic sites (SNPs) near the CRISPR cassette and leader at locus 1 in *P. psychrotolerans*. SNP sites were based on the 11 strains containing locus 1 and the numbers above the sites indicate the genomic position of each site based on strain PRS08. SNP sites within the CRISPR cassette's most ancestral repeat and leader strand are indicated by the dashed lines. Comparisons come from the spacer groupings created in earlier analysis as noted the letters A, B, and C next to each comparison. Under each paired sequence, it is indicated whether the two strains are different (D) at a given site or the same (S), with S sites highlighted in grey. Black boxes drawn around the bases indicate significant introgression regions. Arrows indicate that the introgression boundary for comparison A is found 3' of the region shown while the introgression boundary for comparison B is found 5' of the shown region.

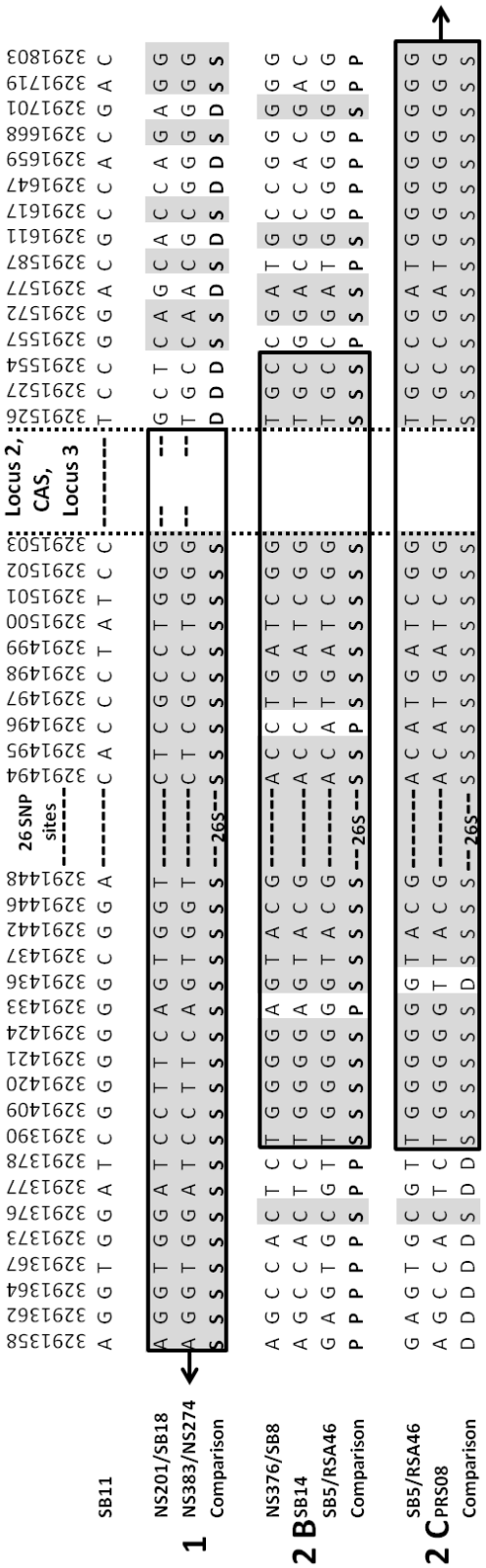

Figure 1.5: Recombination events surrounding CRISPR locus 2 and locus 3, showing all polymorphic sites (SNPs) near the missing CAS region across all strains of *P. psychrotolerans* containing locus 1. The numbers above the sites indicate the genomic position of each site based on strain PRS08. The entirety of the region of locus 2 and locus 3, with the CAS genes between them, is condensed between the dotted lines with the dashes representing the condensed DNA. SB11 contained the entire locus 2 – CAS - locus 3 region while the architecture 1 strains contained only locus 2 and 3 degraded cassettes. The remaining strains completely lacked this region. Both deletion architectures (labeled 1 and 2) are shown here, with architecture 2 broken into a comparison of strains from spacer Group B while ignoring Group C (2B) and a comparison between a Group B pair and Group C (2C). All three comparisons indicate whether the strains within that comparison were the same (S) for a given base, with S sites highlighted in grey. In comparison 1 and 2C, it is indicated whether the two strains are different (D) at a given site, while comparison 2B shows whether the site was polymorphic (P) for the strains in the comparison. Black boxes drawn around the bases indicate significant introgression regions. Arrows indicate that the introgression boundary for comparison 1 is 5' of the region shown and the introgression boundary for comparison 2C is 3' of the region shown. 26 consecutive (S) sites shared across all three comparisons have been condensed for ease of viewing.

Chapter 2

Detection of interspecies HGT of CRISPR-Cas in the bacterial genus, *Pseudomonas*

Abstract

The prokaryotic defense system, CRISPR-Cas, is well known to have undergone horizontal gene transfer (HGT) events between distantly related species of bacteria, while a recent study found evidence of frequent HGT among strains of *Pseudomonas psychrotolerans* that resulted in frequent homologous recombination of components of a CRISPR locus and in recombination-related deletion of the Cas genes. The hypothesis tested is that HGT of CRISPR-Cas above the species level is limited due to homologous recombination becoming ineffective and because a CRISPR cassette transferred from one species to another offers little immediate selective benefit for the recipient as the cassette is unlikely to contain adaptive resistance for phage relevant to its new host species. To determine whether these barriers limit the interspecific HGT of CRISPR-Cas, we examined strains within a clade of the *Pseudomonas* genus. Strong evidence was found for the occurrence of HGT between bacterial species within this clade, though these HGT events were characterized as rare, founder events that created novel genomic loci. The finding that these HGT events were rare even among different loci within the same species of bacteria supports the hypothesis of intraspecies HGT of CRISPR-Cas being limited by the mechanism of homologous recombination between dissimilar sequences. We also report here that the majority of Cas genes found within our dataset were immediately flanked by two CRISPR cassettes. While at first this pattern appeared highly conserved, we identified fundamental differences in the orientation of the 5' and the 3' CRISPR cassette, indicating that this general flanking architecture was formed through convergent evolution.

Introduction

CRISPR-Cas is a prokaryotic system characterized by adaptive genome defense. The system is able to store short copies (25-50bp) of DNA from phage that have attacked the cell and quickly degrade this DNA should it come back in contact with the cell (Makarova et al, 2006; Barrangou et al., 2007). These short sequences (known as spacers) are found clustered together, flanked by similarly sized repeats in a region known as the CRISPR-Cassette. Adjacent to the 3' end of the CRISPR cassette is a 100-500 bp long non-coding region known as the leader sequence which is involved with the creation of new spacers (Diez-Villasenor et al., 2013). The CRISPR cassette creates crRNA (CRISPR RNA), which target memorized DNA for degradation through the help of the CRISPR associated (Cas) genes (Karginov et al., 2010).

The CRISPR-Cas systems are well spread throughout the prokaryotic taxa (Grissa et al., 2007) and it is generally assumed that horizontal gene transfer (HGT) has played a major part in the evolution of the system (Kupczok, et al., 2015; Koonin et al, 2017). Interspecific HGT of CRISPR-Cas has been well established by the identification of closely related systems in distantly related bacteria (Haft et al., 2005; Godde and Bickerton, 2006). Additional supportive evidence comes from the finding of CRISPR-Cas on horizontally transmitted mega-plasmids (Godde and Bickerton, 2006), and the detection of significantly different GC content in some Cas genes relative to the rest of the bacterial genome (Horvath et al., 2008).

At the other taxonomic extreme, it has also been shown that in *Pseudomonas psychrotolerans* there have been numerous occurrences of intraspecific HGT of CRISPR leaders and cassettes (Chapter 1), indicating that intraspecific HGT can be quite frequent; however, there was a clear bias. Successful HGT events typically resulted in the homologous recombination of relatively short regions (approximately 200 – 500 bp long) into existing CRISPR loci rather than

in the creation of novel ones. HGT followed by recombination was also implicated in several independent deletions of the complete set of Cas genes.

Less is known regarding the frequency and barriers controlling the spread of CRISPR-Cas through HGT among closely related species. Homologous recombination into a pre-existing CRISPR-Cas locus in the recipient chromosome requires there to be a region of high sequence similarity between the donor DNA (imported via transformation, transduction, or conjugation) and the recipient. This effect is expected to reduce the likelihood of successful CRISPR-Cas HGT via recombination as relatedness decreases (Rocha et al., 2005; Didelot and Maiden, 2010). As a result, the interspecific dynamics of CRISPR-Cas evolution is predicted to be much less dependent upon the introgression of sequence into pre-existing loci, and more influenced by the creation of new loci at non-homologous sites.

Regardless of the mechanism, a second possible limit to the interspecific spread of CRISPR-Cas lies in the potential lack of an immediate benefit from the newly acquired CRISPR-Cas system. Though some bacteriophages have a broad host range, there is a positive correlation in bacteria between their relatedness and susceptibility towards the same bacteriophage (Koskella and Meaden, 2013). Spacers acquired within the CRISPR-Cas of one species may therefore not be relevant to a different species of bacteria. Though new spacers will be created following HGT, the time it takes to acquire relevant spacers may reduce the newly acquired CRISPR-Cas' immediate selective benefit. If the recombination results in the replacement of an existing CRISPR cassette with a cassette from a different species of bacteria, there may even be a loss of fitness in the host due to removal of the previous cassettes.

The goal of this study was to investigate the nature and frequency of successful inter-species HGT of CRISPR-Cas by studying a group of species within the genus, *Pseudomonas*. Specifically, the study was designed to test the hypothesis that successful interspecific HGT of all

or part of CRISPR-Cas is uncommon and generally results in the formation of new loci. The results indeed showed a clear bias towards interspecies HGT creating novel loci and support our hypothesis regarding the barriers imposed on the evolution of the system through horizontal exchange.

Methods:

Creating a Pseudomonas phylogeny

A clade of the *Pseudomonas* genus was defined based around *P. psychrotolerans*, a species previously found to exhibit HGT of type 1F CRISPR cassettes and leaders between its strains (Chapter 1) and *P. aeruginosa*, one of the most frequently studied species within the genus and known to contain multiple CRISPR-Cas type 1F loci (Cady et al., 2015). We based the choice on a *Pseudomonas* phylogeny that was divided into numerous groups (Gomila et al. 2015), and selected a monophyletic clade made up of groups 15-17. This clade encompassed both of these species and had good bootstrap support isolating it from the rest of the *Pseudomonas* tree. A preliminary nucleotide BLAST (Altschul et al., 1990) of the Cas1 gene found in *P. psychrotolerans* only identified a single pseudomonas species (*P. chlororaphis*, from subgroup 7) from the Gomila et al. (2015) phylogeny that was not included within these three groups. This species was used as an outgroup.

Genomic data was collected from NCBI (NCBI Research Coordinators, 2017) for strains of bacteria from the species found in this clade. We ignored strains with greater than 99.5% gapped identity with any other strain used in our study (as reported by NCBI) to avoid comparing identical strains of bacteria. One species used in this study, *P. aeruginosa*, had over 2000 sequenced genomes available on NCBI as of Jan 1, 2017. To avoid overrepresentation of this species, we only selected those strains that were fully assembled as of this date (though all fully

sequenced genomes regardless of level of assembly were used for other species). A set of 21 ribosomal genes previously identified as ideal for the creation of prokaryotic phylogenies (Lang et al., 2013). These genes were aligned using CLUSTALW (Larkin et al., 2007) and a maximum likelihood phylogenetic analysis was created using RAxML-HPC2 on XSEDE through CIPRES (Miller et al., 2012).

Detecting CRISPR cassettes, leader sequences and Cas proteins

To compare CRISPR-Cas between the chosen strains, CRISPR cassettes and their repeat consensuses were first identified through the use of the CRISPR finder web-tool (Grissa et al., 2007). The repeat consensus found at each locus was aligned with all others using CLUSTALW. Scores for percentage similarity were then created for each pair of repeats with the goal of clustering repeats into separate groups of close relatives.

We searched for CRISPR-associated Cas genes within 20kb of the cassettes using the CRISPR finder web-tool. Cas types were identified by comparing their array of genes to the known gene composition of each Cas type (Makarova et al., 2011). For CRISPR cassettes/leaders with no nearby Cas genes (typically referred to as “orphaned" cassettes), the CRISPR-Cas type was determined based on what repeat group they matched with.

Identification of the Cas type was used to delineate the leader sequences of the associated CRISPR. The leader occurs adjacent to the 3’ end of the cassette, and the characteristic sequences associated with different CRISPR-Cas types were previously categorized by Alkhnbashi et al. (2016).

To identify if similar CRISPR systems found within a bacterial species were found at the same locus, sequence data for the first 5 genes 5’ and 3’ of each CRISPR cassette were recorded. If a given cassette was found adjacent to a Cas region then the first 5 genes on the other side of the Cas regions were used instead. Cassettes from different genomes found with the same

adjacent genes both 5' and 3' were considered to be found at the same CRISPR locus, while a pair of cassettes flanking a Cas region were considered to be 2 separate loci.

Comparing phylogenies

Only one type of CRISPR-Cas was found in >3 species (type 1F), and to determine if these CRISPR sequences were consistent with the bacterial phylogeny without invoking HGT, a phylogeny was created using the leader sequences, that together with the adjacent CRISPR cassettes, make up the type 1F CRISPR. Our analysis was restricted to type IF because phylogenetic analyses are generally only informative when comparing >3 taxa.

The leader sequences were aligned using CLUSTALW and a maximum likelihood phylogeny was created using CIPRES. Similarly, a Cas phylogeny was created using the same method based on the Cas1 gene, a gene that has been previously described as the most highly conserved of the Cas genes (Makarova et al., 2015).

HGT involving the CRISPR and/or Cas loci could be detected in two ways. First, HGT was indicated if the sequences of a single species occupied more than one position within the phylogeny, separated from each other by a different species. Second, HGT was indicated if the phylogeny was different from the topology of the bacterial phylogeny (containing only the same species). This comparison was made after any intraspecific variability was removed in the most conservative fashion possible, by removing all but one of the single species branches leaving only the species branch that most supported the bacterial phylogenies' topology. The comparison involved imposing the bacterial topology on the CRISPR or Cas phylogeny (using PAUP; Swofford, 2003) and comparing the likelihood of this tree to that of the original unconstrained tree. These values were compared for significant differences using a one-tailed Kishino Hasegawa analysis (Kishino and Hasegawa, 1989; Hasegawa and Kishino, 1989). A one-tailed test was used as the constrained tree cannot have a higher maximum likelihood score than the

unconstrained tree. A significant difference would indicate HGT. These tests were performed separately using both the leader sequences and the Cas1 gene sequences.

Similar testing was done looking at each individual CRISPR locus identified within the study to identify intraspecies HGT ocuring between CRISPR at the same genomic position as was seen in *P. psychrotolerans* (Chapter 1). For these tests, an outgroup was chosen for both the bacterial and leader phylogenies based on the closest related CRISPR locus (from within our analysis). The bacterial outgroup (*P. chlororaphis*) was not used as an outgroup here nor in the interspecies CRISPR tests as it cannot be accurately determined whether CRISPR-Cas from this species were basal to those within our data set due to the potential of interspecies HGT.

For both the interspecies and intraspecies tests for HGT (treated individually from each other), the Holm-Bonforroni correction was made to account for multiple tests (Holm, 1979).

Spacer analysis

To identify matching spacers between different CRISPR cassettes, a personal blast database was created through the BLAST+ package (Camacho et al., 2008) consisting of all of the spacers found within our *Pseudomonas* clade. This database was then run against itself to identify matches between multiple spacers. A BLAST cutoff of $e < 10^{-10}$ was used, which identified spacers with 1 or no mismatched bases between them.

Results

Pseudomonas subgroup phylogeny

Of the 24 species found in the three groups (15-17) of the *Pseudomonas* phylogeny of Gomila et al. (2015), 15 had at least one fully sequenced genome available through NCBI. In total, these 15 species included 75 unique strains, ranging from 1 to 21 strains per species (Table 2.1), noting that, although *P. aeruginosa* had over 2000 sequenced strains, we only included the

21 that were fully assembled before our cut-off of January 1st, 2017. *Pseudomonas chlororaphis* was used as an outgroup to root the phylogeny.

The three groups identified by Gomila et al (2015) were largely preserved in the resulting phylogeny (Figure 1); however, there were two notable changes. First, the two group 17 species (*P. psychrotolerans* and *P. oleovorans*), originally basal to the P. aeruginosa clade (group 15) were shifted to within the P. aeruginosa clade. Second, *P. composti,* which was originally placed as *P. aeruginosa*'s closest relative, was shifted outside of group 15. The tree of Gomila et al. (2015) used only partial sequences for 4 genes to create the phylogeny and had low resolution regarding the exact placement of these three species (bootstrap values <50%). The new placements shown in Figure 1 were very well supported (bootstrap values = 100%).

Identifying CRISPR-Cas

Three CRISPR-Cas types were identified within our 75 strains: type 1C, type 1E, and type 1F (Table 2.1; Figure 2.1). These CRISPR-Cas systems are highly divergent (Makarova et al., 2011; Makarova et al., 2015), as seen by low similarities among their Cas1 genes (between 28-41% similarity) and between their repeat consensuses (20-46% similarity) while similarities between CRISPR cassette repeats of the same Cas type ranged from 76-100%. These genetic differences between types are much greater than those found among the *Pseudomonas* species being analyzed (with the lowest gene similarities across the 10 genes being in the range of 80% similarity). For these reasons, the three CRISPR-Cas systems were analyzed independently.

Out of the 51 CRISPR cassettes found within our dataset, 15 were "orphaned", i.e. found to have no identifiable Cas genes in their genome (Table 2.1) and no more than a single orphaned CRISPR (cassette plus leader) was found in any strain. *P. psychrotolerans* has previously been shown to exhibit a high ratio of orphaned CRISPR cassettes (Chapter 1) and accounted for 10 of these 15 strains. One species only had orphaned CRISPR (*P. citronellolis* in 1/1 strain), while in

P. aeruginosa, 4/12 of the strains with CRISPR lacked Cas genes. All of the orphaned CRISPR cassettes could be typed because their repeats shared >80% identity with type 1F (and less than 35% identity with repeats from type 1C or type 1E systems).

We define a CRISPR locus as a unique region within the bacterial genome consisting of at least a CRISPR cassette and a leader strand. Any adjacent Cas genes were also included as part of this CRISPR locus. If two CRISPR cassettes were found flanking a set of Cas genes, this was considered to be a single locus though the 5' and 3' cassettes were further identified as "a" and "b" respectively (e.g. *P. aeruginosa* has two cassettes, 3a and 3b, flanking a set of Cas genes). Each species showing CRISPR presence was found to contain between 1 and 5 CRISPR loci, and importantly no CRISPR loci were conserved between species (based on their adjacent 5' and 3' genes).

We note that a previous analysis based on alignment data detected several highly degraded CRISPR cassettes lacking leader sequences within *P. psychrotolerans* (Chapter 1). Degraded cassettes of this type, which are not detected through CRISPRFinder due to the high number of mutations among their repeats, were not considered in this analysis since their evolutionary history is difficult or impossible to establish.

Evidence for intraspecific HGT of CRISPR-Cas

Intraspecific HGT of CRISPR-Cas between strains of a single bacterial species at a single genomic position would result in discordance between the phylogeny for that bacterial species and the distribution of CRISPR at that locus. To test this, all CRISPR loci found within at least 3 strains of a given species were tested to determine if they shared a similar phylogenetic topology as the bacteria they were found within. Of the 6 loci tested, *P. aeruginosa* locus 3a ($p<0.05$), *P. psychrotolerans* locus 1 ($p<0.001$), and *P. pseudocalcaligenes* ($p<0.001$) were found to have significantly different phylogenetic relationships than their bacterial phylogenies (Table 2.2).

This result is indicative of horizontal gene transfer between related strains of bacteria within pre-established CRISPR loci for half of the tested cases.

Evidence for interspecific HGT of CRISPR-Cas

Interspecific HGT of CRISPR-Cas within the chosen clade of *Pseudomonas* would result in a lack of concordance between the bacterial phylogeny and the distribution of the CRISPR-Cas. Two of the CRISPR-Cas types, 1C and 1E, were found in only 1 strain in just 1 and 3 species respectively (Table 2.1); however, they provide strong evidence of interspecific HGT. In the case of the type 1C, it was found in one strain of *P. aeruginosa* and not in the other 20 strains of that species that were analyzed, nor was it found in any of the strains of the other species of *Pseudomonas* examined. By far the most parsimonious explanation for this distribution is that the presence of this type 1C CRISPR-Cas in the single strain is due to HGT from some unknown species. Similarly, the type 1E CRISPR-Cas was found only in one of the 21 *P. aeruginosa* strains, which in itself suggests HGT; however, it was also found in two additional species within the clade, and none of the three are closely related (Figure 2.1). In both of these other species, type 1E was only found in one of several strains analyzed (Table 2.1). Furthermore, HGT (as opposed to vertical transmission) was also supported by the finding that all three type 1E CRISPR loci were located at different loci in the different species, as determined by mismatching genes 5' and 3' of the CRISPR cassettes. These examples of HGT of type 1C and type 1E suggest a movement of the complete CRISPR-Cas as a unit.

A rather different pattern of abundance was seen for type 1F. It was found in a substantial fraction of the strains of five species (Table 1.1). This allowed for several independent tests of interspecific HGT.

First, we looked at the distribution of leaders within and between species in the leader phylogeny (Figure 2.2). We found that all 23 leaders from strains of *P. aeruginosa* clustered

together into a single monophyletic group, indicating that there was no evidence of HGT between some *P. aeruginosa* strains and any of the other species. In contrast, leaders from *P. psychrotolerans* and *P. pseudoalcaligenes* were found to be paraphyletic. Both species had leaders that grouped by sequence into two distinct groups, and in both species these different groups separated in the phylogeny by the other species, plus *P. thermotolerans,* and *P.citronellolis* (Figure 2.2), providing strong evidence for the occurrence of HGT of CRISPR leaders between species of *Pseudomonas*.

Second, we looked at the distribution of type 1F across the bacterial phylogeny. As in the case of type 1E, its distribution was disjunct (Figure 2.1), suggesting either substantial HGT or alternatively, if it is assumed that the common ancestor of the clade had a type 1F CRISPR-Cas locus, then a complex pattern of deletion would be needed to give the observed result. However, the possibility of vertical transmission explaining this result is further undermined by noting that there were no shared CRISPR loci between bacterial species.

Third, the hypothesis of vertical transmission combined with deletion predicts concordance between the leader and bacterial phylogenies. Our pruned leader tree (see starred CRISPR loci in Figure 2.2) was found to be significantly different from the same tree when constrained by the bacterial topology ($p<0.023$), indicating that HGT has reshaped the leader tree, even after accounting for the clear examples of HGT that created heterogeneity in leader sequences in *P. psychrotolerans* and *P. pseudoalcaligenes*.

Within a given species and Cas type, the Cas genes were always found at the same location, e.g. in *P. aeruginosa* strains, the Cas genes were at the same location in all 10 strains with type 1F Cas, while the Cas genes were located differently in both the strain with type 1E Cas and the strain with the type 1C Cas. No Cas were found at the same location within the genome between species.

To identify whether the Cas genes have moved horizontally between bacterial species, we searched for evidence of a lack of concordance between the Cas-1 gene phylogeny from those 4 species with type 1 CRISPR-Cas (1 species, *P. citronellolis*, had a type 1F cassette but no Cas) and those same species' bacterial phylogeny. Because only a single type 1F Cas locus was found within each of the 4 species with type 1F CRISPR Cas presence, we did not have to prune the tree (other than eliminating redundant strains representing the same locus). We found these trees to share significantly different histories ($p<0.039$), indicating that the Cas genes have moved horizontally between *Pseudomonas* species. The biggest difference between the two tree topologies is the placement of *P. psychrotolerans* as a close sister to *P. aeruginosa* in the bacterial phylogeny, but *P. thermotolerans* is its closest relative within the CAS phylogeny.

The Cas1 phylogeny showed that the Cas regions formed monophyletic groupings based on the bacterial species they came from while the leaders did not (Figure 2.3). This raised the question of whether the leader sequences and the Cas genes adjacent to them are moving in tandem. A pattern seen in all but the single type 1F Cas region in *P. thermotolerans* was one in which the Cas genes were flanked on either side by a CRISPR leader and cassette (Figure 2.4), so that each set of Cas genes had two CRISPR cassettes in which to store spacers. To determine if this combination of two CRISPR loci (leader plus cassette) with one set of Cas genes between them had been moving among *Pseudomonas* species as a single unit (or possibly was shared due to homology), we checked to see if the element had the same structure at each of the Cas loci. It did not: all four of the CRISPR-Cas structures found in the different species with type 1F Cas had unique orientations (Figure 2.4). All positive strains of *P. aeruginosa* had one cassette 5' of the Cas on the same strand and a second cassette 3' of the Cas on the alternate strand, while *P. psychrotolerans* had a 5' cassette on the opposite strand and a 3' cassette on the same strand, *P. pseudoalcaligenes* had both 5' and 3' cassettes on the same strand as the Cas, and *P.*

thermotolerans had only a single 5' cassette on the opposite strand (Figure 2.4). The variation in the orientation of the two CRISPR loci relative to the Cas genes suggests independent origins and provides strong evidence for convergent evolution for this pattern of two CRISPR cassettes flanking a Cas region.

The type 1E system also shared this flanking pattern in all three of the species it was found in. Like with type 1F, there was more than a single fundamental structure for this flanking pattern. Both *P. aeruginosa* and *P. pseudoalcaligenes* type 1E systems had 5' and 3' CRISPR cassettes on the same strand as the Cas genes while *P. mendocina* had 5' and 3' cassettes on the opposite strand as the Cas genes (Figure 2.4). Again, this variation in the orientation of the flanking CRISPR loci relative to the Cas is indicative of convergent evolution.

Spacer analysis

Spacers are 25-50 bp long sequences within the CRISPR cassette that match pieces of foreign DNA and are able to target these elements for deletion through the Cas gene products (Godde and Bickerton, 2006; Barrangou et al., 2007). As a test for very recent CRISPR-Cas HGT between species, we looked for shared spacers found between any CRISPR cassettes within the *Pseudomonas* group. We identified 747 total spacers within our dataset (including those from all CRISPR-Cas types), and 29 of these spacers were found in more than one cassette (Table 2.3). All 29 of these shared spacers were found in the type 1F system (specifically 15 shared spacers out of 261 unique spacers in *P. aeruginosa*, 13 shared out of 33 unique in *P. psychrotolerans*, and 1 shared out of 127 unique spacers in *P. pseudoalcaligenes*). In all cases, shared spacers were found in the same species and at the same cassette locus. No evidence of recent HGT across species or loci was found. A previous study identified the 13 shared spacers from *P. psychrotolerans* to be caused in part by intraspecific recombination across different strains at the same locus (Chapter 1).

Discussion

It has been demonstrated from previous studies that HGT of CRISPR-Cas occurs within species (Chapter 1; Kupczok et al., 2015), between distantly related species (Haft et al., 2005; Godde and Bickerton, 2006; Horvath et al., 2008), and potentially between closely related species (Yang et al., 2015). However, little has been done to address potential barriers limiting the effect of HGT on CRISPR-Cas evolution. Homologous recombination has been shown to be a driver of intraspecies CRISPR horizontal gene transfer between identical loci (Chapter 1), but, since recombination requires regions of sequence similarity between the donor and the recipient, we hypothesized that homologous recombination involving CRISPR-Cas and its surrounding sequence would be rare between different bacterial species. This predicts that homologous CRISPR-Cas loci would be at different genomic locations in different species. Further, we proposed that, if a species gains a CRISPR cassette by interspecific HGT, it is unlikely to have an immediate selective benefit. This is because the CRISPR cassettes created in one bacterial species would be unlikely to be advantageous for another since the two species would typically be targeted by a different spectrum of phage. As a result, the probable advantage of a newly acquired CRISPR cassette would decline with declining relatedness (at least until new relevant spacers were created). Both of these concepts lead us to expect that interspecific HGT would represent rare, founder-like events relative to the rate of intraspecific HGT.

To test this hypothesis, we examined a well-defined clade of Pseudomonas to identify whether HGT of CRISPR-Cas both at the intraspecies level and at the interspecies level was taking place. In all three of the bacterial species containing multiple CRISPR-Cas positive strains (*P. aeruginosa*, *P. psychrotolerans*, and *P. pseudoalcaligenes*), we found evidence of CRISPR HGT occurring across related strains of single bacterial species based on significant discordance

between that species bacterial phylogeny and the CRISPR phylogeny. This result confirms that CRISPR-Cas is recombining between bacterial strains of the same species at shared, pre-established CRISPR loci.

We also found strong evidence for interspecies HGT based on the disjunct distribution of three CRISPR-Cas types through our *Pseudomonas* phylogeny, the lack of any conserved genomic positions showing CRISPR-Cas presence between species, the formation of paraphyletic clades (based on species) in the type 1F CRISPR leader phylogeny, and significant differences between both the type 1F leader sequence tree and the type 1F Cas1 gene tree in comparison to the bacterial phylogeny. From all of this evidence, we concluded that CRISPR-Cas has certainly been moving horizontally between species within the clade and is unlikely to be shared through homology between any two species within our dataset. However, to identify the nature of these HGT events, we looked more closely at the CRISPR leader phylogeny.

Within our data set, HGT of CRISPR across species was identified as always creating novel sites, rather than resulting in recombination into existing sites, as evidenced by the leaders at every CRISPR locus clustering into separate monophyletic groups. If HGT had occurred between two different, already established CRISPR loci, we would expect to see some divergence in the sequence relationships at the recipient locus (specifically, the recipient strain's CRISPR matching more closely with the donor locus than with the locus it is currently found). Even within a single species, we saw no evidence of genetic exchange between CRISPR loci at different genomic positions beyond that of the presumed founding HGT event (such as a likely duplication event spawning *P. aeruginosa* locus 2 from locus 3b), since they also clustered in separate groups. This is as expected under the hypothesis that there is a barrier to CRISPR-Cas HGT caused by limited recombination targets.

Further evidence for cross species HGT being rare was found in the spacer analysis, in which no CRISPR cassettes from different species (or even different loci within a single species) were shared. This finding also is as expected following the hypothesis that CRISPR cassettes formed in one bacterial species are likely not relevant in a different bacterial species, though if this were the only barrier we would expect to still see intraspecific transfer of CRISPR between different loci. However, on its own, the lack of shared spacers between loci is rather weak evidence as CRISPR spacers can experience very rapid rates of turnover (Rho et al., 2012).

Together, our finding that there was HGT occurring between like loci at the intraspecies level but no evidence of HGT occurring between species beyond that of rare founder events supports our hypothesis regarding barriers restricting CRISPR-Cas HGT. Specifically, we find support for the claim that as relatedness between bacterial species decreases, CRISPR-Cas becomes much less likely to spread between them due to limited opportunities for homologous recombination based on a lack of recombination between different established CRISPR loci of the same bacterial species. The barrier caused by CRISPR-Cas being exchanged between distantly related species not initially having spacers relevant to the recipient may still play a part, though we would not have expected to see any limits to CRISPR-Cas HGT at the intraspecies level were this the only barrier. Previous findings have shown that major bacterial clades taken from a bacterial sample are completely absent of CRISPR-Cas while other bacterial clades taken from the same environmental sample had an abundance of the system, leading the authors to predict that certain bacterial clades may share characteristics that make CRISPR-Cas inefficient (Burstein et al., 2016). We believe it possible that there may in fact be scenarios in which bacteria from these clades could benefit from HGT, though our proposed barriers greatly reduce the opportunities of acquiring such a system from the distantly related bacteria in close proximity found to have the system present.

We also noted a pattern previously seen in some *Escherichia coli* (Diez Villasenor, et al., 2010) in which two CRISPR cassettes and leader sequences would flank a set of Cas genes. While at first glance this pattern appeared to be conserved, closer inspection revealed that each *Pseudomonas* species with type 1F Cas genes present had a unique pattern of 5' and 3' CRISPR/leader orientation. Similarly, there were two different patterns of Cas-flanking 5' and 3' CRISPR cassette orientation found within the three species with type 1C CRISPR-Cas. This result indicates that the general "flanking" architecture had numerous independent origins. We believe this pattern may be explained by the selfish operon hypothesis (Lawrence and Rother, 1996), which predicts that genes frequently undergoing HGT are more likely to be found near other genes necessary for their functionality (specifically regarding mobile selfish elements). Genes that become separated from the rest will lose their functionality and find themselves at the mercy of genetic drift. From this, we predict that CRISPR cassettes and their adjacent Cas genes benefit from moving in tandem. Further, any CRISPR cassette moving in isolation would benefit most from recombining near an existing Cas gene, which can explain how so many of the Cas genes are found with cassettes on both sides when they only need a single cassette to operate.

Table 2.1. Distribution of CRISPR cassettes for each Cas type analyzed in this study. Orphaned cassettes lack associated Cas genes. Only one strain was found to contain more than one Cas type.

Species	Total number of strains	Total cassettes*	Orphaned cassettes*	Type 1C*	Type 1E*	Type 1F*	Unique 1F loci
aeruginosa	21	26 (12)	4 (4)	1 (1)	2 (1)	23 (10)	3
psychrotolerans	16	13 (11)	10 (10)	0	0	13 (11)	2
pseudoalcaligenes [a]	4	8 (3)	0	0	2 (1)	6 (3)	1
citronellolis	1	1 (1)	1 (1)	0	0	1 (1)	1
thermotolerans	1	1 (1)	0	0	0	1 (1)	1
mendocina	7	2 (1)	0	0	2 (1)	0	0
nitroreducans	4	0	0	0	0	0	0
oleovorans	3	0	0	0	0	0	0
flexibilis	4	0	0	0	0	0	0
alcaligenes	4	0	0	0	0	0	0
composti	2	0	0	0	0	0	0
alcaliphila	3	0	0	0	0	0	0
toyotomiensis	2	0	0	0	0	0	0
resinovorans	2	0	0	0	0	0	0
knackmussii	1	0	0	0	0	0	0
TOTAL	**75**	**51 (29)**	**15 (15)**	**1 (1)**	**6(3)**	**44 (26)**	**11**

* values in parenthesis represent the number of strains these cassettes were found in for each species
[a] a single strain of *P. pseudoalcaligenes* was found to contain both a type 1E and type 1F CRISPR-Cas system

Table 2.2: Test for interspecies and intraspecies HGT for various CRISPR and bacterial phylogenies using the Kishino Hasegawa analysis. Outgroups were chosen for intraspecies CRISPR based on the closest CRISPR relative from CRISPR leader sequence phylogeny.

	Comparison	outgroup strain (CRISPR Locus)	ingroup strains	p value
Interspecies	Type 1F Leader	na	5	0.023*
	Type 1F Cas	na	4	0.039*
Intraspecies	*aeruginosa* L2	psychrotolerans SB11 (L2a)	7	0.491
	aeruginosa L3a	psychrotolerans SB11 (L2a)	8	0.011*
	aeruginosa L3b	psychrotolerans SB11 (L2a)	7	0.500
	psychrotolerans L1	pseudoalcaligenes CECT (L1b)	11	0.000***
	pseudoalcaligenes L1a	thermotolerans J53 (L1)	3	0.000***
	pseudoalcaligenes L1b	psychrotolerans SB11 (L1)	3	0.205

*significant at Holm-Bonferroni corrected p<0.05
***significant at Holm-Bonferroni corrected p<0.001

Table 2.3: Spacer data looking at shared spacers between cassettes from the same locus (same species), different loci (same species) and any locus within a different species.

			Spacers shared between:		
type	Total spacers	unique spacers	Strains (same locus)	Strains (different locus)	Species
1C	39	39	0	0	0
1E	221	221	0	0	0
1F	487	428	29	0	0
unknown	4	4	0	0	0

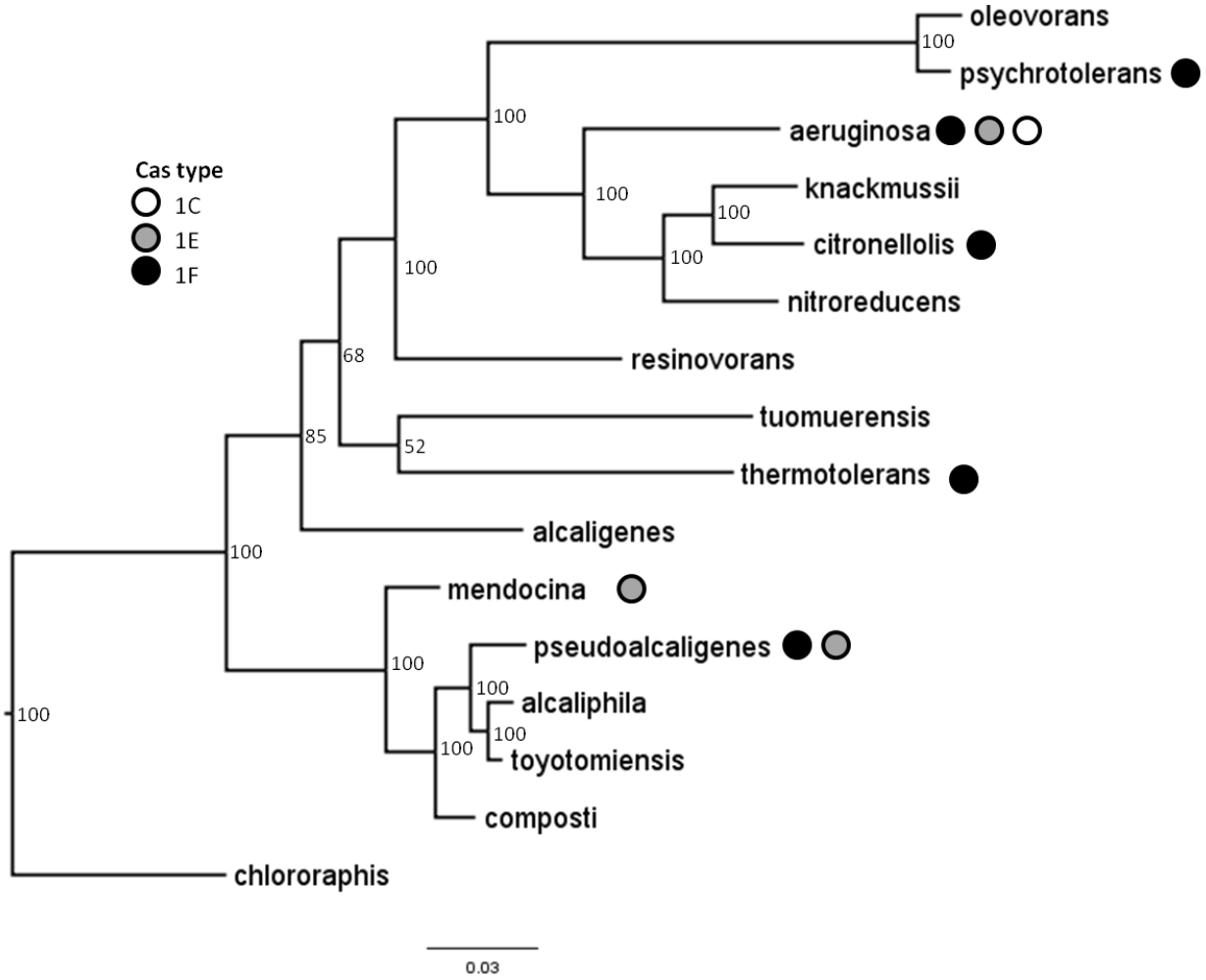

Figure 2.1: Maximum likelihood phylogeny from 21 ribosomal genes for *Pseudomonas* subgroup and corresponding CRISPR-Cas presence. Each species has a representative strain mapped above and all CRISPR-Cas systems found within each species are noted to the right of the species by the colored circles. Strains with no circles next to them had no identifiable CRISPR cassettes. The tree is rooted by *P. chlororaphis*.

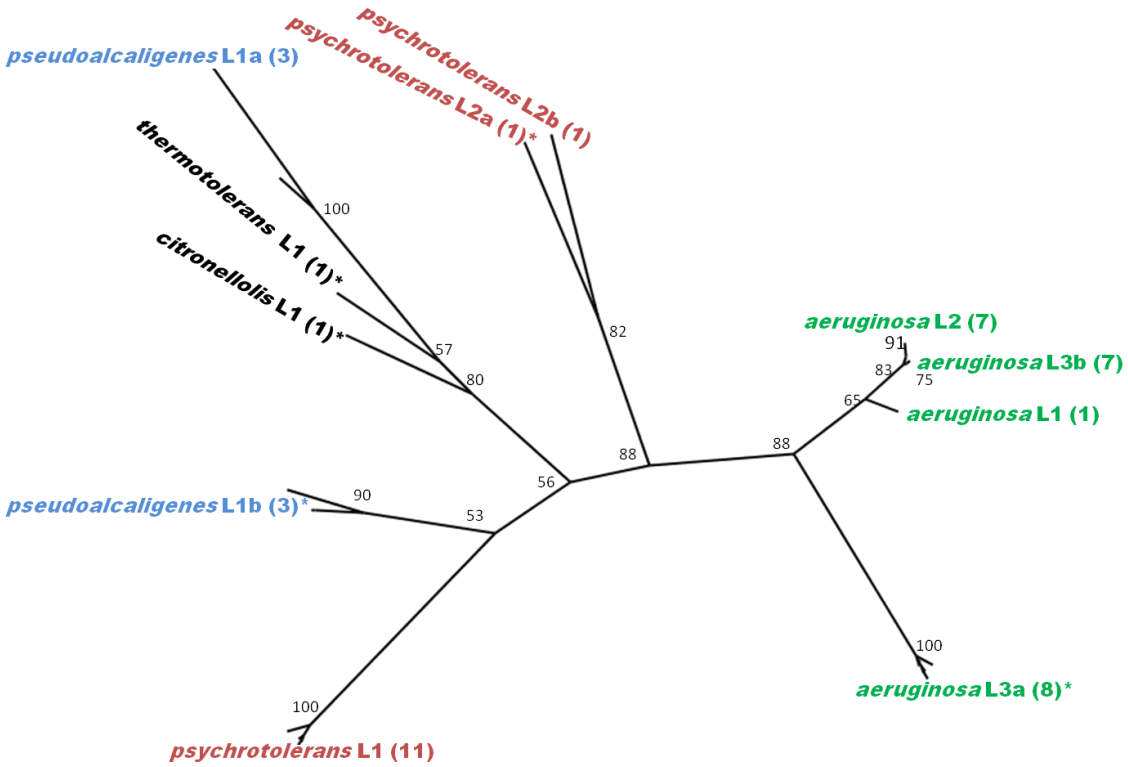

Figure 2.2: Maximum likelihood unrooted type 1F leader phylogeny. Branch ends show the name of the species the leader was found in, the locus within that species ("L" followed by the locus number), and the number of strains represented by that branch (in parentheses). The three species with more than a single strain containing a CRISPR leader and cassette (*aeruginosa, psychrotolerans,* and *pseudoalcaligenes*) have been colored to emphasize their patterns of distribution. Stared branches indicate the loci that were used to test for concordance with the bacterial phylogeny.

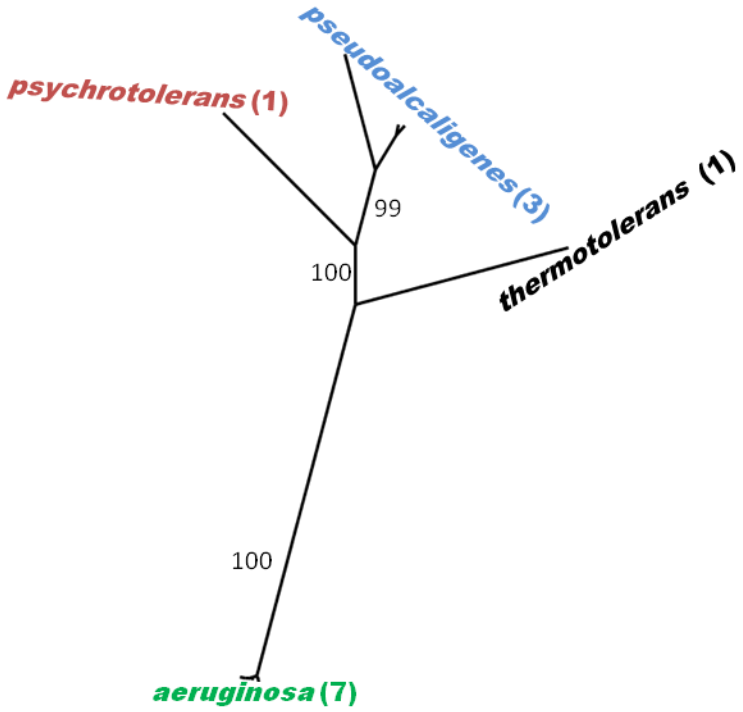

Figure 2.3: Maximum likelihood unrooted type 1F Cas1 phylogeny. Branch ends show the name of the species the Cas1 gene was found in and the number of strains represented by that branch. Cas1 found within a single species was always at the same locus within that species. Branches are not drawn to scale. Note that *P. citronellolis* had no Cas genes and is not represented here.

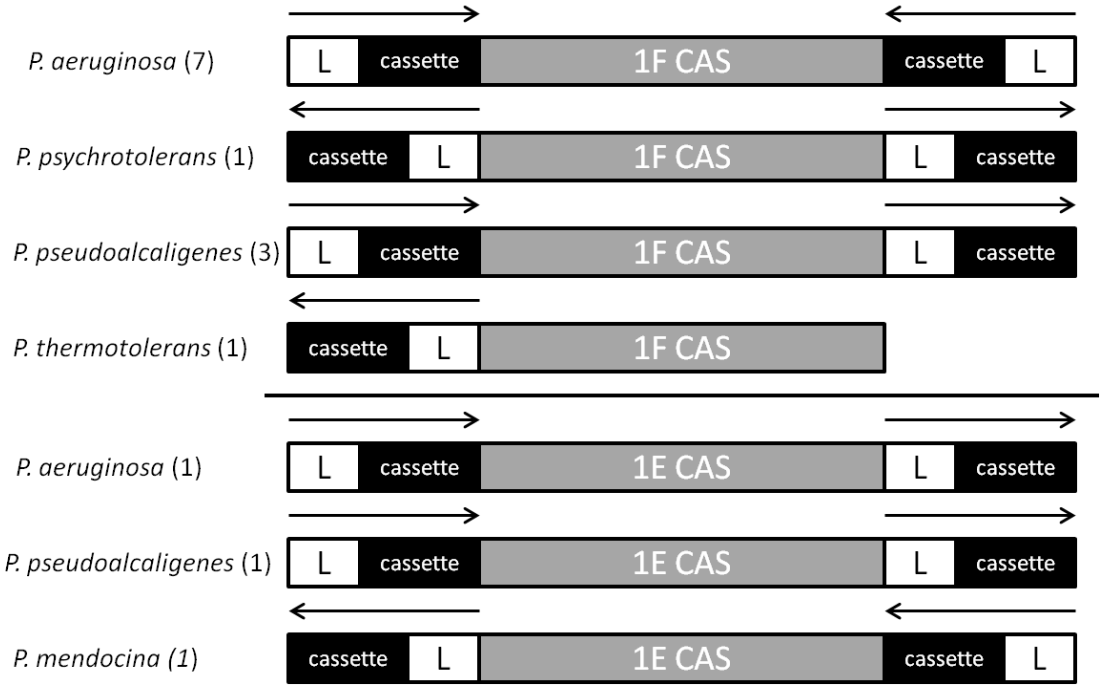

Figure 2.4: Diagram showing the topology of the type 1E and type 1F CRISPR-Cas region and nearby CRISPR cassettes and leaders for this *Pseudomonas* subgroup. Cas regions are oriented 5' to 3' for each species as a reference point. Numbers next to the species name indicate how many strains within that species are represented. Arrows above the CRISPR cassettes and leader sequences point towards the ancestral end of the cassette.

Chapter 3

A phylogenetic Test on the Role of CRISPR-Cas in Limiting Plasmid Acquisition and Prophage Integration in Bacteria

Abstract

CRISPR-Cas is a prokaryotic defense system capable of protecting the cell from damaging foreign genetic elements. However, some such elements can be beneficial and bacteria with CRISPR-Cas may incur a cost of reduced intake of mutualistic plasmids and prophage. To test the hypothesis that CRISPR-Cas limits the horizontal transfer of potentially beneficial genetic material, we compared the distribution of both plasmids and prophage in CRISPR-Cas positive and negative strains across a set of 37 bacterial families, each characterized by only one CRISPR-Cas positive and one CRISPR-Cas negative strain from the same species. This design controlled for phylogenetic bias and environmental covariates in the distribution of CRISPR-Cas. We report a significant negative association between CRISPR-Cas presence and plasmid count, with those strains of bacteria with CRISPR-Cas containing on average less than half the plasmid count of their paired CRISPR-Cas positive strain (0.93 vs. 1.93). CRISPR-Cas positive strains had 31% fewer intact prophage, but the effect was highly variable and not significant. These results support the hypothesis that CRISPR-Cas can influence the rate of plasmid-mediated HGT and, given the abundant evidence of beneficial genes carried by plasmids, provides a clear example of a cost associated with the CRISPR-Cas system.

Introduction

The prokaryotic defense system, CRISPR-Cas, has become well known for its ability to provide immunity for its bacterial host against foreign genetic elements (Barrangou et al., 2007; Marraffini and Sontheimer, 2008). The system is composed of a CRISPR Cassette region, made

up of a series of 25-50 bp long DNA sequences known as “spacers”, each flanked by similarly sized “repeat” sequences, and the Cas (CRISPR associated) genes (Barrangou et al., 2007). The spacers are derived from exogenous sources of genetic material and are able to target these sources (through sequence matching) for degradation carried out by the Cas genes (Haurwitz et al., 2010). The primary benefit of the CRISPR-Cas system comes from its ability to protect bacteria from pathogenic elements such as bacteriophage. However, it has been proposed that the immunity provided by the CRISPR-Cas system may come at a cost to the host cell (Jiang et al., 2013; Vale et al., 2015). These proposed costs include metabolic costs (Weinberger and Gilmore, 2012), risks of self-targeting autoimmunity (Stern et al., 2010), and the cost of reducing the ability of the bacteria to uptake genes from its environment (Marraffini and Sontheimer, 2008; Jiang et al., 2013; Palmer and Gilmore, 2010).

We focus here on the last of these, the proposed cost of reduced acquisition of novel DNA. CRISPR-Cas systems can prevent DNA uptake via plasmid conjugation (Marraffini and Sontheimer, 2008), transformation with naked DNA (Bikard et al., 2012), and viral transduction (Edgar and Qimron, 2010). Previous experimental studies supporting the negative effect of preventing DNA uptake have found CRISPR-Cas to act as a barrier towards the uptake of drug resistant plasmids in both *Staphylococci* (Marraffini and Sontheimer, 2008) and *Enterococci* (Palmer and Gilmore, 2010), although this was not seen in *E. coli* (Touchon et al., 2012). Similarly, a handful of studies show CRISPR-Cas’s ability to protect the cell from temperate (lysogenic) phage as evident by indentifying spacers matching prophage sequence (Briner et al., 2015), experimental evidence of CRISPR-Cas reducing the chance of lytic phage integration (Edgar and Qimron et al., 2010), and by identifying a negative correlation between prophage counts and CRISPR-Cas presence between strains of a *Streptococcus pyogenes* (Nozawa et al., 2011). The success of lysogenic phage depends primarily upon vertical transmission rather than

infectious horizontal transfer, and consequently selection favors lysogenic phage that bring with them genes that increase host fitness (Obeng et al., 2016). For example, it has been proposed that elevated levels of antibiotic resistance within *Staphylococcus aureus* are spread through transduction of lysogenic phage (Haaber et al., 2017). Bacteria protected by CRISPR-Cas systems may again have a cost imposed on them by limiting the intake of new, beneficial genes through transduction.

The generality of the costs of limiting DNA acquisition has yet to be established. To examine this issue, a previous study focused on identifying whether those bacteria with CRISPR-Cas had an effect on the intake of new genes through the analysis of the available prokaryote genomes (1237 from bacteria with 43% CRISPR-Cas positive; 114 from archaea with 84% CRISPR-Cas positive) (Gophna et al., 2015). The authors found contradictory associations between CRISPR-Cas presence/absence or activity (assumed to be measured by the number of spacers) and measures of horizontal gene transfer (HGT). For example, in the genomes of bacteria lacking CRISPR-Cas, they found significantly fewer prophages and novel genes measured by atypical dinucleotide ratios (contrary to expectation), but more novel genes measured by singletons (as expected). However, despite using a large sample of genomes, these genomes were chosen without any control for potentially important phylogenetic or ecological difference among the CRISPR groupings (presence/absence or high/low spacer count). For example, growth temperature was shown to be an important confounding factor, with high temperature species having the most spacers (Gophna et al., 2015), and it was inevitable that some taxa were heavily over sampled relative to others, creating the potential for serious bias (Felsenstein, 1985).

To eliminate these potential sources of bias we analyzed data from across the eubacterial kingdom comparing genomes with and without CRISPR-Cas randomly chosen from within the

same bacterial species, using only one species to represent each family. Using this approach, we looked for a negative correlation between CRISPR-Cas presence and the acquisition of new genetic material across a wide array of bacteria. The two primary hypotheses being tested were whether CRISPR-Cas presence reduced the number of plasmids and whether it reduced the level of prophage integration. We also examined whether there was any general effect of CRISPR-Cas in limiting genome size, after plasmids and prophage were excluded. A secondary pair of questions addresses whether CRISPR-Cas presence influenced the size of plasmids or of prophage that are taken up and retained by bacteria. By identifying whether CRISPR-Cas acts as a functional barrier to accumulation of these diverse types of DNA via HGT, we may better understand the costs and benefits involved in maintaining CRISPR-Cas.

Methods:

Identifying CRISPR-Cas positive and CRISPR-Cas negative strains for pairwise analysis

Initial screening was carried out using CRISPRdb (Grissa, et al., 2007), a database representing close to 7000 bacterial strains (regardless of CRISPR presence/absence) representing over 2100 bacterial species (as of 4/01/2018). Bacterial genomes entered in this database are searched for CRISPR Cassettes using the program, CRISPRFinder (Grissa et al., 2007). Using this information, bacterial species were considered for inclusion in the analysis if they had at least one completely sequenced strain having no identifiable CRISPR Cassette and at least one containing a confirmed CRISPR Cassette. This intraspecific pairing was designed to eliminate the range of ecological and physiological biases that can arise when comparisons of CRISPR-Cas presence and absence are confounded with species differences.

Each bacterial species satisfying this criterion was included in the next step which controlled for phylogenetic bias. The first species alphabetically (by genus, then species name) from each family was identified and one CRISPR-Cas positive and one CRISPR-Cas negative

strain was selected randomly for analysis from those available on 4/01/2018 in the NCBI database (NCBI Research Coordinators, 2017) using the date (day/month) of deposition (those deposited earliest in the year were chosen, regardless of the specific year). To this end, strains were searched for CRISPR-Cas via the CRISPR-Finder program (Grissa et al., 2007) starting with those having a submission date closest to January 1st. If multiple strains were submitted on the same day, the first strain in alpha-numeric ordering was used. For each species, the first strain having a confirmed CRISPR Cassette with at least 10 spacers plus having at least 4 nearby Cas genes within its genomic DNA was chosen as the CRISPR-Cas positive strain. Similarly, the first strain of the same species with neither a CRISPR Cassette nor Cas genes was labeled as the CRISPR-Cas negative strain. Those strains showing confirmed CRISPR Cassettes, but with fewer than 10 spacers or with fewer than 4 adjacent Cas genes, were omitted from the analysis as it is uncertain whether these strains had functional CRISPR-Cas.

Strains not fully assembled were ignored as a full assembly is required to accurately identify plasmid presence. To increase the power of our plasmid analysis, if the selected species did not have any plasmids present in either its CRISPR-Cas absent or CRISPR-Cas present strain, we moved on to the next bacterial species alphabetically within the family until a species with plasmid presence in at least one of the two strains was found. If no species from the family was found to have plasmids in its selected strains, then we reverted to using the first species alphabetically within the family with both a CRISPR-Cas absent and present strain. If no species fulfilled these criteria, the family was not represented.

Tests for differences between CRISPR-Cas positive and CRISPR-Cas negative bacteria

Bacterial strains chosen for this analysis had their genome sequences and their plasmid counts downloaded from NCBI. Plasmid counts reference the number of unique plasmids found within the bacterial strain, not copy number for individual plasmids as this information is

generally not available. To determine if CRISPR-Cas presence reduced plasmid count, a one-tailed Wilcoxon signed-rank test was used through R (R Core Team, 2018). Though there are known representation issues with plasmid counts from public databases under-representing wild bacterial populations (Jørgensen et al., 2014), we see no reason why this would increase our risk of type 1 error.

The total prophage count within a bacterial strain was determined using the program PHASTER (Arndt et al., 2016). PHASTER works by searching the bacterial genome for putative stretches of genes that are highly similar to known prophage genes and identifies "intact" prophage as those above a certain threshold based on the percentage of known prophage length found in the system. The number of intact prophage was recorded for each pair of bacterial strains, and the numbers in CRISPR-Cas positive and negative strains was compared, again using a one-tailed Wilcoxon signed-rank test.

"Other" DNA was found for each strain by subtracting the total amount of total prophage DNA (including questionable and incomplete prophage as recorded by PHASTER) from the total chromosomal DNA (as recorded by NCBI). Data were log transformed to control for skew and a one-tailed paired t-test was performed through R to identify if the size of CRISPR-Cas positive genomes were smaller than genomes lacking CRISPR-Cas.

To control for issues of multiple testing, the Holm-Bonferroni corrected alpha values were used to identify significant values between the three analyses (Holm, 1979).

A second and independent analysis investigating plasmid size in CRISPR-Cas positive and negative strains was done by averaging the plasmid sizes for each strain of bacteria for species in which both strains had at least a single plasmid. These averages were then log transformed to control for skew and treated as pairwise data in a two-tailed paired t-test. The same procedure was repeated for prophage size.

Finding plasmid-targeting CRISPR spacers

All CRISPR spacers found within our analysis were tested for matches with known plasmid sequences. Spacers were analyzed with CRISPRTarget (Biswas et al., 2013) using the Refseq plasmid and phage databases to search for matches. Plasmid sequences found to match any spacers from our dataset were analyzed for CRISPR cassettes using CRISPRFinder to ensure that these matches were targeting the plasmid and not simply a related CRISPR cassette found within the plasmid.

Results:

Identification of CRISPR-Cas variable species

A total of 132 species from 45 different families with variable CRISPR-Cas presence/absence were identified from CRISPRdb. Analysis of these CRISPR-Cas variable species identified 37 families represented by at least one species with a fully sequenced CRISPR-Cas positive strain (as determined by having at least 10 spacers and at least 4 nearby Cas genes) and a fully sequenced CRISPR-Cas negative strain. These 37 families spanned over 6 bacterial phyla with 4 coming from *Actinobacteria*, 2 from *Bacteroidetes*, 1 from *Chloroflexi*, 1 from *Eubacteria*, 9 from *Firmicutes*, and 20 from *Proteobacteria* (Table 3.1). Out of the 37 species, 9 had no plasmids in either strain, 14 had no prophage in either strain, and of these 5 had neither.

Plasmid count and the presence/absence of CRISPR-Cas

For the analysis of the effect of CRISPR-Cas on plasmid number, we identified 28 species (representing 28 families) with at least one plasmid in one or both of the paired strains from the 37 total families (Table 3.1). In these 28 species, strains with CRISPR-Cas carried on average significantly fewer plasmids, 0.93 versus 1.93 in CRISPR-Cas negative strains, a 52% drop ($p=0.007$, significant at $p<0.05$ after the 3-test correction; see Table 2) with the CRISPR-Cas

negative strain having more plasmids in 15 species (53.6%), fewer in 8 (28.6%), and in the remaining 5 species (17.9%) both strains had the same number of plasmids (Table 3.1). This result supports the hypothesis that the CRISPR-Cas system reduces the likelihood of a bacterium acquiring plasmids.

The same result was obtained when total plasmid DNA was compared between paired CRISPR-Cas positive and negative strains ($z=-2.46$; uncorrected $p=0.007$). This measurement potentially conflates the effects of CRISPR-Cas on plasmid acquisition and on the size of said plasmids (Xu et al., 2015).

To test the hypothesis that size might influence the vulnerability of a plasmid to CRISPR-Cas, we examined the average size of plasmids in the two types of strain. Only 10 species had both CRISPR-Cas present and absent strains with at least one plasmid (Table 3.1) meaning that the test had little power. The geometric mean for these plasmids was 150 kb for CRISPR-Cas absent strains and 101 kb for CRISPR-Cas present strains; however, this 33% decrease in size (based on the geometric mean) in those strains with CRISPR-Cas was not significant (Table 3.2) with only half (5) of the 10 species showing a smaller average plasmid size in the CRISPR-Cas negative strains.

Prophage count and the presence/absence of CRISPR-Cas

Out of the 23 species with intact prophage found in at least one of the two strains, 11 (48%) species had higher intact prophage counts in the CRISPR-Cas absent strains while 8 (35%) had more intact prophage in the CRISPR-Cas present strains and the remaining 4 (17%) had the same number between the two strains (Table 3.1). The average intact prophage count was 1.78 for the CRISPR-Cas absent genomes and 1.22 for CRISPR-Cas present genomes (ignoring species where both strains had no intact prophage). This 31% decrease in CRISPR-Cas negative strains is in the expected direction but was not significant based on a Wilcox Signed-Rank test

(p=0.097; Table 3.2). When the total DNA of complete prophage was compared between paired CRISPR-Cas positive and negative strains the result was the same (z=-1.12; uncorrected p=0.131).

As in the case of plasmids, we tested the hypothesis that size might influence the vulnerability of a prophage to CRISPR-Cas. Nine species had prophage present in both strains (Table 1) and the average prophage size was not significantly different between the groups, with geometric means of 80 kb for the CRISPR-Cas absent strains and 78 kb for the CRISPR-Cas present strains (Table 3.2). Like the plasmid length analysis, the sample size was small and the test had limited power.

Test for difference in "other" genomic DNA based on CRISPR-Cas presence

Across the 37 families, only 18 of the 37 species (48.6%) were found to have more "other" DNA in the genomes of CRISPR-Cas negative strains (Table 3.1) contrary to the hypothesis that such strains would accumulate more genomic DNA. The geometric mean for the CRISPR-Cas absent genomes was 3,290 kb while the geometric mean for the CRISPR-Cas present genomes was 2,864 kb, but this small 13.0% decrease was not significant (Table 3.2).

Search for plasmid-targeting spacers

Of the 1515 spacers within our dataset, only 22 (1.5%) matched with sequences found outside of bacterial chromosomes, namely from plasmids and phage. However, even with this small sample of known spacer targets, 12 of the 28 bacterial species represented with at least a single plasmid found in either the CRISPR-Cas positive or CRISPR-Cas negative strain had at least a single spacer matching a known plasmid sequence. Further, 11 of these 12 matched with plasmids found within the same species of bacteria as that spacer (although, consistent with the function of CRISPR-Cas, they were never matched a plasmid within the same strain). Since plasmid sometimes carry CRISPR cassettes (Godde and Bickerton, 2006), it was confirmed that

none of the matching sequences identified in the plasmids were part of CRISPR cassettes within the plasmid.

Discussion:

To understand the distribution of CRISPR-Cas among bacteria, it is important to determine if CRISPR-Cas serves as a barrier against the acquisition of new, potentially beneficial genes. To this end, we looked at DNA-content differences between paired strains from the same species of bacteria with and without CRISPR-Cas, one pair from each of 37 bacterial families. This pairing of strains from a single species to represent each bacterial family removed the potential bias that can drive the results of large-scale association studies (Felsenstein, 1985). In this study, we focused primarily on identifying whether CRISPR-Cas presence reduced the number of plasmids or of intact prophage. We found that CRISPR-Cas presence significantly reduced plasmid count from 1.93 to 0.93 ($P<0.007$). Note that this result is significant after correcting for the triple testing of plasmid number, prophage number, and amount of other genomic DNA ($p <0.05$, Table 3.2). It also remains significant if we ignore the implicit 1-tailed nature of the hypothesis and apply a 2-tailed test. There was no significant effect of CRISPR-Cas presence on the number of intact prophage, although it was in the predicted direction with an average of 1.22 intact prophage present in the CRISPR-Cas positive strains relative to 1.78 present in the negative strains, nor was there evidence of the accumulation of other genomic DNA in strains lacking CRISPR-Cas, although again the means were in the expected direction (3.3Mb vs. 2.9Mb; Table 3.2).

The significant reduction in plasmid number in CRISPR-Cas positive strains is in line with previous work on single species where the presence of CRISPR-Cas has been found to limit plasmid conferred drug resistance (Marraffini and Sontheimer, 2008; Palmer and Gilmore, 2010),

although the negative association between CRISPR-Cas and plasmid count is not always seen e.g. in *Escherichia coli* (Touchon, 2011). This apparently stochastic nature of the difference was also seen in the present study, where although 54% of CRISPR-Cas positive strains had fewer plasmids, 29% had more. However, it was perhaps notable that this 29% of reversals occurred primarily in species where the average plasmid count across the 2 strains was small (mean = 1.6) compared to 3.3 in cases where the expected pattern was observed. The prevalence of many plasmid-targeting spacers in our dataset is further evidence that the CRISPR-Cas system is responsible for limiting the intake of new plasmids.

The effect of CRISPR-Cas presence on the number of intact prophage found within the genome was not significant. While the mean number of prophage was 46% greater in CRISPR-Cas negative strains, the effect was far from uniform, so that the median number in both groups was 1 prophage/genome. However, as in the case of plasmids, all extreme differences (>2 prophage) showed more prophage in the CRISPR-Cas negative strains. While this result was not significant, it is consistent with most previous work that identified a link in individual species of bacteria between CRISPR-Cas presence and lower prophage content (Edgar and Qimron, 2010; Briner et al., 2015; Nozawa et al., 2011; Hatoum-Aslan and Marraffini, 2014). The opposite relationship was found by Gophna et al. (2015), who found significantly more prophage in the genomes of CRISPR-Cas positive strains in their large-scale study of 1399 genomes. However, this result may have been influenced by phylogenetic and/or ecological biases as the study lacked any control for the oversampling of particular species, families, or environments among the genomes analyzed.

The lack of a significant association with prophage integration could also be due to some CRISPR-Cas systems targeting RNA. In these systems, prophage are not targeted until the prophage is transcribed (Goldberg and Marraffini; 2015), so an intact prophage found within

some CRISPR-Cas positive strains may in fact be kept in check by an RNA-targeting CRISPR-Cas.

We also examined the average size of plasmids and prophage in the CRISPR-Cas positive and negative strains to determine if the size of invading DNA elements might influence the effectiveness of CRISPR-Cas. For this test we could only consider species where both the positive and negative strains contained at least one of the element type being considered. Unfortunately, this reduced the sample size to only 10 species for the plasmid comparison and 8 species for the prophage comparison. It was found that, on average, the geometric mean for plasmids was 33% smaller in the CRISPR-Cas positive strains (150kb vs. 101kb), but this difference was far from significant (Table 3.2). It remains to be seen if there is a real effect of larger plasmids being more likely to be detected and targeted by CRISPR-Cas. Geometric means for prophages were 2.5% larger in CRISPR-Cas positive strains (80 kb vs. 78 kb), a small non-significant difference.

It is possible that the presence of CRISPR-Cas could reduce the overall size of the genome by limiting the incorporation of non-viral novel genes into the genome. Our analysis of "other DNA" (excluding plasmids and all prophage DNA) revealed means in the expected direction (2.9Mb vs. 3.3Mb), but the difference was not significant. Although Gophna et al. (2015) came to the same conclusion through an analysis of singletons (defined as genes not shared with closely related genomes), their results are difficult to interpret for two reasons. First, they found significantly more singletons in CRISPR-Cas negative strains, but, in apparent contrast, within CRISPR-Cas positive strains singletons increased with the number of spacers. Since they interpreted the number of spacers as a measure of CRISPR-Cas activity, they concluded that there was no net effect of CRISPR-Cas. However, their assumption that the number of spacers is a measure of CRISPR-Cas activity should be treated with caution since

spacer count conflates two different effects, the rate of spacer acquisition and the rate of spacer loss. Second, as noted above, the study suffers from a lack of phylogenetic or ecological control that could seriously bias their results. For example, strains with more spacers could be predominantly from species at higher risk of genetic exchange from exogenous sources.

Our analysis sacrificed large sample size for rigorous phylogenetic and ecological controls, one consequence of which was the inability to analyze data from Archaea (only 2 archaea species in CRISPRdb showed intraspecific variability in CRISPR cassette presence). Phylogenetic control was achieved by using only one species per family and ecological control was achieved by comparing a pair of strains from the same species. Comparing CRISPR-Cas positive and negative strains of the same species minimizes the effect of ecological covariates that may be linked to HGT, such as growth temperature (Gophna et al., 2015) and other factors such as genome size. Further, in defining CRISPR-Cas positive strains, we used a strict cutoff determining functionality: at least 10 spacers and the presence of Cas genes. These criteria were designed to reduce the chance of including strains that may have lost CRISPR-Cas activity but still show clear evidence of a cassette (so-called orphaned cassettes; Palmer and Gilmore, 2010).

In summary, our findings support the hypothesis that CRISPR-Cas acts as a barrier against the acquisition of new plasmids into the bacterial cell; however, the effect, though quite extreme, is stochastic in nature so that the barrier is far from rigid. As beneficial genes for antibiotic resistance (Svara and Rankin, 2016) and heavy metal resistance (Cooksey, 1990) are often found in plasmids, scenarios may arise where conjugation is necessary for survival. Under these conditions, bacterial populations that had previously been protected by CRISPR-Cas systems would be under heavy selection for mutations that limited the functionality of the CRISPR-Cas. The loss of CRISPR-Cas through selection for the uptake of a mobile genetic element has been demonstrated experimentally(Jiang et al., 2013; Bikard et al., 2012). Results

such as these can aid in explaining why so many bacteria lack the otherwise beneficial CRISPR-Cas system (Burstein et al., 2016).

Table 1. Plasmid counts (pl), intact prophage count (pro) and "other" DNA length for CRISPR-CAS negative (CR-) and CRISPR-CAS positive (CR+) strains of bacteria.

Phylum	Family	Genus and species	CR- pl	CR+ pl	CR- pro	CR+ pro	CR- "other"	CR+ "other"
Actinobacteria	*Actinomycetaceae*	*Trueperella pyogenes*	0	0	0	1	2253422	2224794
	Bifidobacteriaceae	*Bifidobacterium longum*	0	2	0	0	2725058	2348892
	Corynebacteriaceae	*Corynebacterium glutamicum*	1	1	0	0	3300239	3184739
	Propionibacteriaceae	*Acidipropionibacterium acidipropionici*	0	1	0	0	3614155	3639226
Bacteroidetes	*Bacteroidaceae*	*Bacteroides fragilis*	2	0	0	1	5185511	4865615
	Flavobacteriaceae	*Elizabethkingia meningoseptica*	0	0	0	0	3961259	3862225
Chloroflexi	*Dehalococcoidaceae*	*Dehalococcoides mccartyi*	0	0	1	0	1386335	1307154
Eubacteria	*Listeriaceae*	*Listeria monocytogenes*	0	0	0	0	2745517	2853591
Firmicutes	*Aerococcaceae*	*Aerococcus urinae*	0	0	0	0	2038874	1947562
	Bacillaceae	*Bacillus pumilus*	1	1	1	1	3664486	242673
	Clostridiaceae	*Clostridium baratii*	1	1	1	1	3044150	3108366
	Enterococcaceae	*Enterococcus faecalis*	4	0	3	0	2843973	2725025
	Lactobacillaceae	*Lactobacillus amylovorus*	2	0	1	1	2037001	1886167
	Leuconostocaceae	*Leuconostoc gelidum*	3	0	3	0	1869141	1872599
	Paenivacillaceae	*Paenibacillus polymyxa*	2	0	0	2	5578290	5693536
	Staphylococcaceae	*Staphylococcus equorum*	5	0	0	1	2704339	2776093
	Streptococcaceae	*Streptococcus salivarius*	0	1	1	0	2144344	2183393
Proteobacteria	*Acetobacteraceae*	*Gluconobacter oxydans*	5	0	0	0	2718639	3562517
	Acidithiobacillaceae	*Acidthiobacillus ferrooxidans*	0	0	0	0	2854438	2976897
	Aeromonadaceae	*Aeromonas veronii*	0	0	3	0	4800509	4274686
	Alteromonadaceae	*Alteromonas mediterranea*	1	0	0	0	4391537	4528668
	Burkholderiaceae	*Ralstonia solanacearum*	1	3	2	0	3667342	3478832
	Campylobacteraceae	*Campylobacter coli*	2	1	0	1	1674643	1670474
	Enterobacteriaceae	*Cedecea neteri*	0	1	2	3	4792443	5091102
	Helicobacteraceae	*Helicobacter cetorum*	1	1	0	0	1785366	1914546
	Legionellaceae	*Legionella pneumophila*	0	1	0	0	3378443	3493776
	Moraxellaceae	*Acinetobacter baumannii*	3	1	3	2	5924711	424711
	Morganellaceae	*Xenorhabdus bovienii*	2	0	8	7	4114701	3725398
	Myxococcaceae	*Anaeromyxobacter dehalogenans*	0	0	0	0	5013479	5019329
	Pasteurellaceae	*Aggregatibacter actinomycetemcomitans*	0	0	0	1	2091003	2318408
	Pseudomonadaceae	*Pseudomonas chlororaphis*	0	1	3	2	6986473	6493909
	Rhodobacteraceae	*Rhodobacter sphaeroides*	5	2	3	2	4142174	4338239
	Rhodospirillaceae	*Azospirillum brasilense*	5	5	0	0	3026793	2986426
	Shewanellaceae	*Shewanella baltica*	4	1	1	1	5043376	5058386
	Vibrionaceae	*Vibrio harveyi*	2	0	0	0	5996729	5933621
	Xanthomonadaceae	*Xanthomonas citri*	0	2	0	1	3865165	5109645
	Yersiniaceae	*Serratia fonticola*	2	0	5	0	3940621	4030506

Table 2: Paired tests comparing the genomes of CRISPR-Cas present (CC+) and absent (CC-) strains.

Variable	Sample size	CC- mean[3]	CC+ mean[3]	Test value	significance (H - B corrected)
Plasmid count[1]	28	1.93	0.93	z=-2.45	p=0.007* (0.021)
Intact prophage count[1]	23	1.78	1.22	z=-1.27	p=0.097 (0.194)
"Other" DNA	37	3,290 kb	2,864 kb	t=-1.36	p=0.091 (0.192)
Plasmid size[2]	10	150 kb	101 kb	t=0.94	p=0.371
Prophage size[2]	9	80 kb	78 kb	t=0.18	p=0.865

[1] Comparisons of plasmid and prophage count excluded 0,0 pairs.

[2] Comparisons of plasmid and prophage size required both strains to have plasmids or prophage, respectively.

[3] geometric mean used for "other" DNA and both plasmid and prophage sizes

*Significant at Holm-Bonferonni corrected $p<0.05$

General Conclusion

The main focus of this dissertation has been to better understand why the genomic defense system, CRISPR-Cas, has such a variable rate of presence/absence throughout much of the bacterial kingdom. From the start, it was recognized that the variable presence /absence of the CRISPR-Cas system when comparing related bacterial taxa must be due to some combination of horizontal gain events and loss events. Here, I summarize the findings of my dissertation and how these two distinct effects have been elucidated.

Horizontal gain of the CRISPR-Cas system had previously been shown to have occurred between distantly related bacterial taxa (Haft et al., 2005; Godde and Bickerton, 2006; Horvath et al., 2010), though beyond this little had been done to further characterize the process. In looking at a single bacterial species (*Pseudomonas psychrotolerans*) showing CRISPR-Cas variability (Chapter 1), we found evidence for multiple horizontal events by which CRISPR-Cas moved between related bacterial strains. Further, we can be certain that these events were caused by homologous recombination based on the short lengths of the recombined sites and the differing recombination boundaries surrounding the recombined CRISPR locus. This recombination occurred between an already established CRISPR locus, meaning that the recombined CRISPR cassette and leader sequences were likely replacing those already found within the region. The adjacent components were even found to have differing evolutionary histories in one case, as evident by a recombination boundary containing the CRISPR cassette but not the adjacent leader sequence.

This general pattern of intraspecies recombination between existing CRISPR loci was again seen in several other species of the *Pseudomonas* genus (Chapter 2). However, the mechanism of homologous recombination is predicted to become less likely, creating a barrier to horizontal transfer, as the relatedness of the bacterial sequences decreases. As hypothesized, no

recombination between existing CRISPR loci was found at the interspecies level for the *Pseudomonas* species used in this analysis. This was demonstrated by every CRISPR locus forming its own monophyletic grouping for both the CRISPR leader sequence phylogeny and the Cas gene phylogeny. Instead, horizontal gene transfer between bacterial species followed a rare, founder like pattern where a single horizontal transfer event of CRISPR-Cas occurred at a novel locus. This newly acquired CRISPR-Cas could act as a progenitor for future intraspecies homologous recombination events.

It was also hypothesized that CRISPR cassettes transferred at the interspecies level would be less likely to provide an immediate selective benefit, assuming a decline in the overlap of the specific phages attacking the donor and recipient bacteria, though no evidence was found that could disentangle this effect from the effect of a recombination barrier.

Though our results have shown homologous recombination to be a major factor in the spread of CRISPR-Cas across bacterial genomes, it was also found to be involved with the loss of CRISPR-Cas functionality. Evidence for this was seen by the multiple, independent losses of the Cas genes within *P. psychrotolerans* (Chapter 1). There were two different genomic architectures characterized by absent Cas genes at this site, and both architectures were discordant from the bacterial phylogeny. Between the two architectures, three independent recombination events were identified that fully explain the phylogenetic discordance. The most parsimonious explanation here was that loss of the Cas genes was driven by homologous recombination splicing out the existing Cas genes for a donor sequence absent of the Cas genes. This is particularly interesting as it demonstrates a single, relatively common mutational change (when compared to rates of base pair substitutions) capable of completely removing CRISPR-Cas functionality (as CRISPR cannot function without the Cas genes; Palmer and Gilmore, 2010).

This helps to explain the extremely fast response to selection for loss of CRISPR-Cas functionality shown experimentally (Jiang et al., 2013).

Previous researchers have proposed a particularly interesting hypothesis that CRISPR-Cas imposes a cost on its bacterial host genome by acting as a barrier towards the integration of potentially beneficial genes from exogenous sources (Marraffini and Sontheimer, 2008; Palmer and Gilmore, 2010; Jiang et al., 2013). Though this hypothesis has been supported both experimentally (Jiang et al., 2013) and through identification of a negative association between CRISPR-Cas presence and the uptake of drug-resistance plasmids within *Staphylococci* (Marraffini and Sontheimer, 2008) and *Enterococci* (Palmer and Gilmore, 2010), this result was not found in *Escherichia Coli* (Touchon et al., 2012). To provide a more general overview, an exhaustive analysis of paired strains of bacteria with and without CRISPR-Cas from throughout the bacterial kingdom was performed, employing strict phylogenetic controls (Chapter 3). A significant negative correlation between CRISPR-Cas presence and plasmid count was reported here, with those strains lacking CRISPR-Cas containing an average of 1 more plasmid than their CRISPR-Cas present counterpart.

The results found in this dissertation paint the picture of a dynamic CRISPR-Cas evolutionary history of gains and losses. While recombination allows for rapid exchange of CRISPR-Cas between related genomes and less frequent founder-like exchange between more distant genomes, the same process allows for functionality to be lost when the system becomes more costly than beneficial. The switch in selection can be at least partially explained by the cost the CRISPR-Cas system imposes on the bacterial genome by reducing the intake of potentially beneficial genes. Understanding of the nature of CRISPR-Cas evolution allows for hypothesis regarding certain patterns, such as the hypothesis (though largely untested at the moment) at the end of chapter 1 regarding the multiple recent losses of CRISPR-Cas functionality in *P.*

psychrotolerans being driven by man-made changes to the bacteria's environment. We may also hypothesize that bacteria containing drug resistant accessory genes are more likely to have compromised CRISPR-Cas systems and are potentially more susceptible to phage therapies than their non-drug resistant counterparts. Through the research presented in this dissertation and future research building off of these results, we may better understand more broadly the tradeoffs involved in the "double edged sword" of genomic defense.

References

Alkhnbashi, O.S., Shah, S.A., Garrett, R.A., Saunders, S.J., Costa, F., and Backofen, R. (2016). Characterizing leader sequences of CRISPR loci. Bioinformatics *32*, i576–i585.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. Journal of Molecular Biology *215*, 403–410.

Arndt, D., Grant, J.R., Marcu, A., Sajed, T., Pon, A., Liang, Y., and Wishart, D.S. (2016). PHASTER: a better, faster version of the PHAST phage search tool. Nucleic Acids Research *44*, W16–W21.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. Science *315*, 1709–1712.

van Belkum, A., Soriaga, L.B., LaFave, M.C., Akella, S., Veyrieras, J.-B., Barbu, E.M., Shortridge, D., Blanc, B., Hannum, G., Zambardi, G., et al. (2015). Phylogenetic Distribution of CRISPR-Cas Systems in Antibiotic-Resistant *Pseudomonas aeruginosa*. MBio *6*, e01796-15.

Bikard, D., and Marraffini, L.A. (2013). Control of gene expression by CRISPR-Cas systems. F1000Prime Reports *5*.

Bikard, D., Hatoum-Aslan, A., Mucida, D., and Marraffini, L.A. (2012). CRISPR Interference Can Prevent Natural Transformation and Virulence Acquisition during In Vivo Bacterial Infection. Cell Host & Microbe *12*, 177–186.

Biswas, A., Gagnon, J.N., Brouns, S.J.J., Fineran, P.C., and Brown, C.M. (2013). CRISPRTarget: Bioinformatic prediction and analysis of crRNA targets. RNA Biology *10*, 817–827.

Briner, A.E., Lugli, G.A., Milani, C., Duranti, S., Turroni, F., Gueimonde, M., Margolles, A., van Sinderen, D., Ventura, M., and Barrangou, R. (2015). Occurrence and Diversity of CRISPR-Cas Systems in the Genus Bifidobacterium. PLOS ONE *10*, e0133661.

Burstein, D., Sun, C.L., Brown, C.T., Sharon, I., Anantharaman, K., Probst, A.J., Thomas, B.C., and Banfield, J.F. (2016). Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nature Communications *7*, 10613.

Cooksey, D.A. (1990). Plasmid-Determined Copper Resistance in Pseudomonas syringae from Impatiens. Appl. Environ. Microbiol. *56*, 13–16.

Delaney, N.F., Balenger, S., Bonneaud, C., Marx, C.J., Hill, G.E., Ferguson-Noel, N., Tsai, P., Rodrigo, A., and Edwards, S.V. (2012). Ultrafast Evolution and Loss of CRISPRs Following a Host Shift in a Novel Wildlife Pathogen, Mycoplasma gallisepticum. PLoS Genetics *8*, e1002511.

Didelot, X., and Maiden, M.C.J. (2010). Impact of recombination on bacterial evolution. Trends in Microbiology *18*, 315–322.

Diez-Villasenor, C., Almendros, C., Garcia-Martinez, J., and Mojica, F. (2010). Diversity of CRISPR loci in Escherichia coli. Microbiology *156*, 1351–1361.

Díez-Villaseñor, C., Guzmán, N.M., Almendros, C., García-Martínez, J., and Mojica, F.J.M. (2013). CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of Escherichia coli . RNA Biology *10*, 792–802.

Edgar, R., and Qimron, U. (2010). The Escherichia coli CRISPR System Protects from Lysogenization, Lysogens, and Prophage Induction. Journal of Bacteriology *192*, 6291–6294.

Felsenstein, J. (1985). Phylogenies and the Comparative Method. The American Naturalist *125*, 1–15.

Godde, J.S., and Bickerton, A. (2006). The Repetitive DNA Elements Called CRISPRs and Their Associated Genes: Evidence of Horizontal Transfer Among Prokaryotes. Journal of Molecular Evolution *62*, 718–729.

Goldberg, G.W., and Marraffini, L.A. (2015). Resistance and tolerance to foreign elements by prokaryotic immune systems — curating the genome. Nature Reviews Immunology *15*, 717–724.

Gomila, M., PeÃ±a, A., Mulet, M., Lalucat, J., and GarcÃa-ValdÃ©s, E. (2015). Phylogenomics and systematics in Pseudomonas. Frontiers in Microbiology *6*.

Gophna, U., Kristensen, D.M., Wolf, Y.I., Popa, O., Drevet, C., and Koonin, E.V. (2015). No evidence of inhibition of horizontal gene transfer by CRISPR–Cas on evolutionary timescales. The ISME Journal *9*, 2021–2027.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007a). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Research *35*, W52–W57.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007b). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics *8*, 172.

Haaber, J., Penadés, J.R., and Ingmer, H. (2017). Transfer of Antibiotic Resistance in Staphylococcus aureus. Trends in Microbiology *25*, 893–905.

Haft, D.H., Selengut, J., Mongodin, E.F., and Nelson, K.E. (2005). A guild of forty-five CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. In PLoS Comput Biol, p.

Hasegawa, M., and Kishino, H. (1989). CONFIDENCE LIMITS ON THE MAXIMUM-LIKELIHOOD ESTIMATE OF THE HOMINOID TREE FROM MITOCHONDRIAL-DNA SEQUENCES. Evolution *43*, 672–677.

Hatoum-Aslan, A., and Marraffini, L.A. (2014). Impact of CRISPR immunity on the emergence and virulence of bacterial pathogens. Current Opinion in Microbiology *17*, 82–90.

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.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. Scandinavian Journal of Statistics *6*, 65–70.

Horvath, P., Romero, D.A., Coute-Monvoisin, A.-C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C., and Barrangou, R. (2008). Diversity, Activity, and Evolution of CRISPR Loci in Streptococcus thermophilus. Journal of Bacteriology *190*, 1401–1412.

Iacobino, A., Scalfaro, C., and Franciosa, G. (2013). Structure and Genetic Content of the Megaplasmids of Neurotoxigenic Clostridium butyricum Type E Strains from Italy. PLoS ONE *8*, e71324.

Jiang, W., Maniv, I., Arain, F., Wang, Y., Levin, B.R., and Marraffini, L.A. (2013). Dealing with the Evolutionary Downside of CRISPR Immunity: Bacteria and Beneficial Plasmids. PLoS Genetics *9*, e1003844.

Jørgensen, T.S., Xu, Z., Hansen, M.A., Sørensen, S.J., and Hansen, L.H. (2014). Hundreds of Circular Novel Plasmids and DNA Elements Identified in a Rat Cecum Metamobilome. PLoS ONE *9*, e87924.

Jorth, P., and Whiteley, M. (2012). An Evolutionary Link between Natural Transformation and CRISPR Adaptive Immunity. MBio *3*, e00309-12-e00309-12.

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

Kishino, H., and Hasegawa, M. (1989). Evaluation of the Maximum Likelihood Estimate of the Evolutionary Tree Topologies from DNA Sequence Data, and the Branching Order in Hominoidea. Journal of Molecular Evolution *29*, 170–179.

Koonin, E.V., Makarova, K.S., and Zhang, F. (2017). Diversity, classification and evolution of CRISPR-Cas systems. Current Opinion in Microbiology *37*, 67–78.

Koskella, B., and Meaden, S. (2013). Understanding Bacteriophage Specificity in Natural Microbial Communities. Viruses *5*, 806–823.

Kupczok, A., and Bollback, J.P. (2013). Probabilistic models for CRISPR spacer content evolution. BMC Evolutionary Biology *13*, 54.

Kupczok, A., Landan, G., and Dagan, T. (2015). The contribution of genetic recombination to CRISPR array evolution. Genome Biology and Evolution.

Lang, J.M., Darling, A.E., and Eisen, J.A. (2013). Phylogeny of Bacterial and Archaeal Genomes Using Conserved Genes: Supertrees and Supermatrices. PLoS ONE *8*, e62510.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., et al. (2007). Clustal W and Clustal X version 2.0. Bioinformatics *23*, 2947–2948.

Lawrence, J.G., and Roth, J.R. (1996). Selfish Operons: Horizontal Transfer May Drive the Evolution of Gene Clusters. Genetics *143*, 1843–1860.

Makarova, K.S., Grishin, N.V., Shabalina, S.A., Wolf, Y.I., and Koonin, E.V. (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., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J.J., Charpentier, E., Haft, D.H., et al. (2015). An updated evolutionary classification of CRISPR–Cas systems. Nature Reviews Microbiology *13*, 722–736.

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

Midha, S., Bansal, K., Sharma, S., Kumar, N., Patil, P.P., Chaudhry, V., and Patil, P.B. (2016). Genomic Resource of Rice Seed Associated Bacteria. Frontiers in Microbiology *6*.

Miller, M.A., Pfeiffer, W., and Schwartz, T. (2012). The CIPRES science gateway: enabling high-impact science for phylogenetics researchers with limited resources. (ACM Press), p. 1.

Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J., and Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J Mol Evol *60*.

NCBI Resource Coordinators (2017). Database Resources of the National Center for Biotechnology Information. Nucleic Acids Research *45*, D12–D17.

Nozawa, T., Furukawa, N., Aikawa, C., Watanabe, T., Haobam, B., Kurokawa, K., Maruyama, F., and Nakagawa, I. (2011). CRISPR Inhibition of Prophage Acquisition in Streptococcus pyogenes. PLoS ONE *6*, e19543.

Nunney, L., Yuan, X., Bromley, R.E., and Stouthamer, R. (2012). Detecting Genetic Introgression: High Levels of Intersubspecific Recombination Found in Xylella fastidiosa in Brazil. Applied and Environmental Microbiology *78*, 4702–4714.

Obeng, N., Pratama, A.A., and Elsas, J.D. van (2016). The Significance of Mutualistic Phages for Bacterial Ecology and Evolution. Trends in Microbiology *24*, 440–449.

Palmer, K.L., and Gilmore, M.S. (2010). Multidrug-Resistant Enterococci Lack CRISPR-cas. MBio *1*, e00227-10-e00227-19.

Pearson, W.R., and Lipman, D.J. (1988). Improved tools for biological sequence comparison. Proceedings of the National Academy of Sciences of the United States of America *85*, 2444–2448.

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*.

Prasanna, M.K., Sidde, D.K., Moudgal, R., Kiran, N., Pandurange, K.T., and Vishwanath, K. (2013). Impact of Fungicides on Rice Production in India. In Fungicides - Showcases of Integrated Plant Disease Management from Around the World, M. Nita, ed. (InTech), p.

R Core Team (2018). R: A language and environment for statistical computing. (Vienna, Austria).

Rezzonico, F., Smits, T.H.M., and Duffy, B. (2011). Diversity, Evolution, and Functionality of Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) Regions in the Fire Blight Pathogen Erwinia amylovora. Applied and Environmental Microbiology *77*, 3819–3829.

Rho, M., Wu, Y.-W., Tang, H., Doak, T.G., and Ye, Y. (2012). Diverse CRISPRs Evolving in Human Microbiomes. PLoS Genetics *8*, e1002441.

Rocha, E.P.C., Cornet, E., and Michel, B. (2005). Comparative and Evolutionary Analysis of the Bacterial Homologous Recombination Systems. PLoS Genetics *1*, e15.

Stern, A., Keren, L., Wurtzel, O., Amitai, G., and Sorek, R. (2010). Self-targeting by CRISPR: gene regulation or autoimmunity? Trends in Genetics *26*, 335–340.

Svara, F., and Rankin, D.J. (2011). The evolution of plasmid-carried antibiotic resistance. BMC Evolutionary Biology *11*.

Swofford, D.L. (2003). PAUP*. Phylogenetic Analysis Using Parsimony (*and other Methods) Version 4. USA. Nat. Biotechnol. *18*, 233–234.

Touchon, M., and Rocha, E.P.C. (2010). The Small, Slow and Specialized CRISPR and Anti-CRISPR of Escherichia and Salmonella. PLoS ONE *5*, e11126.

Touchon, M., Charpentier, S., Clermont, O., Rocha, E.P.C., Denamur, E., and Branger, C. (2011). CRISPR Distribution within the Escherichia coli Species Is Not Suggestive of Immunity-Associated Diversifying Selection. Journal of Bacteriology *193*, 2460–2467.

Touchon, M., Charpentier, S., Pognard, D., Picard, B., Arlet, G., Rocha, E.P.C., Denamur, E., and Branger, C. (2012). Antibiotic resistance plasmids spread among natural isolates of Escherichia coli in spite of CRISPR elements. Microbiology *158*, 2997–3004.

Vale, P.F., Lafforgue, G., Gatchitch, F., Gardan, R., Moineau, S., and Gandon, S. (2015). Costs of CRISPR-Cas-mediated resistance in *Streptococcus thermophilus*. Proceedings of the Royal Society B: Biological Sciences *282*, 20151270.

Vorontsova, D., Datsenko, K.A., Medvedeva, S., Bondy-Denomy, J., Savitskaya, E.E., Pougach, K., Logacheva, M., Wiedenheft, B., Davidson, A.R., Severinov, K., et al. (2015). Foreign DNA acquisition by the I-F CRISPR–Cas system requires all components of the interference machinery. Nucleic Acids Research *43*, 10848–10860.

Weinberger, A.D., and Gilmore, M.S. (2012). CRISPR-Cas: To Take Up DNA or Not—That Is the Question. Cell Host & Microbe *12*, 125–126.

Xu, L., Paterson, A.D., Turpin, W., and Xu, W. (2015). Assessment and Selection of Competing Models for Zero-Inflated Microbiome Data. PLOS ONE *10*, e0129606.

Yang, C., Li, P., Su, W., Li, H., Liu, H., Yang, G., Xie, J., Yi, S., Wang, J., Cui, X., et al. (2015). Polymorphism of CRISPR shows separated natural groupings of *Shigella* subtypes and evidence of horizontal transfer of CRISPR. RNA Biology *12*, 1109–1120.

Zhang, Q., and Ye, Y. (2017). Not all predicted CRISPR–Cas systems are equal: isolated cas genes and classes of CRISPR like elements. BMC Bioinformatics *18*.