A tripartite survey of CRISPR-Cas diversity: Assessing the contours and limits of natural variation in CRISPR-Cas systems by associated genes and spacer acquisition biases

Alexander McKay, BSc (Hons)

Australian National University

A thesis submitted for the Doctor of Philosophy (PhD)

The Australian National University

John Curtin School of Medical Research

Declaration.

I declare that the research presented in this work is original work that I conducted during my candidature at the Australian National University, with the following exceptions:

Dr. Gaetan Burgio (MD, PhD) performed tree based phylogenetics in Chapter 4 and power statistics in Chapter 5.

Prof Eric Stone helped to develop the statistical method in assessing spacer acquisition biases in Chapter 5.

Alexander McKay, BSc (Hons)

PhD candidate

Genome Editing and Microbial Immunity

Division of Genome Sciences and Cancer

The John Curtin School of Medical Research

The Australian National University

Word count: 47021

Acknowledgements.

I'd like to express my gratitude to the Australian National University for giving me the opportunity to undertake this program, and be part of the university community over the last decade. I'm also deeply appreciative of the National Computational Infrastructure (NCI) for allowing me the use of their resources to conduct my investigation.

It's difficult to express in words the enormity and depth of the sense of debt and gratitude for some of the people I've known at ANU over this period. When I first contacted Gaetan Burgio, my present supervisor about doing a CRISPR-related project, I first thought he may have been a fictional character. I'd never met him in any biochemistry courses, and naturally assumed he may have been an affiliated or visiting academic from a separate university. So naturally, I was very surprised when he responded to my email. Being part of the Burgio lab has altered who I am and the course of my life in ways I could never have conceived of. Perhaps I am biased but in many ways he's the paragon of what for me, it means to be a researcher. Exceptionally skilled and learned in the field, what really distinguishes him from many other investigators is how much he cares, both for his research, his lab and his family. It is a clichéd adage. But all greatness comes from the heart, and Gaetan has this in spades. I could not have done this without his belief, encouragement and dedication and that's a feeling I intend to honour for decades to come.

Next I'd like to thank the other members of my supervisory panel: Dan Andrews, Eduardo Eyras and Eric Stone. Every one of you has been amazing supportive at different moments along the journey. A special thanks goes to Dan Andrews and Eric. Dan, it meant a lot to me, and still does, to have had the chance to work as part of your group in the year preceding the start of this program. Over the last 7 years, since meeting you, you've always been a great advisor and confidant. To Eric, thanks for bearing with me and lending your assistance developing the statistical tests used for assessing the significant of spacer acquisition biases. It would've have been a real

struggle without your input and guidance.

Although not present on a full time basis in the lab anymore, I'd like to extend a special thanks to Lora Starrs, for both her excellent teaching and feedback as a mentor as well as being a great friend. My only regret is unlike the rest of the lab members (barring Gaetan) we never quite went rock climbing together.

Among the students in the lab, there are a lot of people who've made the journey a really special one. Ni Ni, Fei-ju, Myu, Emily, Paris, Ethan, Julia, Parnika, Norika. So many people have past through the lab on their own Journeys. I think the one thing about doing a PhD I really disliked the most, was that, in being bound here in Canberra, I haven't had a chance to catch up and go on more adventures with everyone. But hopefully I will very soon.

Of course this journey wouldn't been the same without the 3 people I shared it with as fellow PhD students. To Arash H. Dastjerdi, I said it in week one when you arrived and I'll say it again: You are a crazy dude. Never change. I really do cherish every off the rails conservation we've had over the past few years as well as the adventures hiking and climbing. You have in you the greatest compassion, especially when given the chance to effect it through people.

To Jovita De Silva,
You are uniquely special and important to me and a very close (hopefully life-long) friend. In truth, since you left earlier in year, it pains me every time I look at or walk by the desk where you used to sit. In the past few years, I've found your compassion and consciousness very admirable and perhaps even a bit enviable. It's that looking forward I will strive more for myself. I hope that you remain at peace and continue to find happiness and contentless in your life.

Finally to Anthony Newman,
It's rare to share the kind of bond I have with you on this sort of journey. Having been

friends not just since the start of the program but all the way back since the year I first came to Canberra. I said I'd keep this short, so put simply. You are the greatest friend I've ever had, and it means everything to me to have been able to share this experience with you.

List of abbreviations.

AAA+	ATPase Associated with diverse cellular Activities
Abi	Abortive Infection
BLAST	Basic Local Alignment Search Tool
BREX	Bacteriophage Exclusion
CARF	CRISPR-associated Rossmann Fold
CASCADE	CRISPR-associated complex for antiviral defense
CAST	CRISPR-associated Transposons
CBASS	Cyclic-oligonucleotide-based anti-phage signalling systems
cGAS-STING	Cyclic GMP-AMP Synthase
CHAT	Caspase HetF Associated with Tprs
CRISPR-Cas	Clustered Regularly Interspaced Short Palindromic Repeats -CRISPR associated
crRNA	CRISPR RNA
DDBJ	DNA Data Bank of Japan
DISARM	Defense Island System Associated with Restriction-Modification
DNA	Deoxyribose Nucleic Acid
ds	Double-Stranded
DSB	Double-Stranded Breaks
E. coli	Escherichia coli
EMBL	European Molecular Biology Laboratory
ENA	European Nucleotide Archive
GBA	Guilt By Association
GOLD	Genomes On-Line Database
HD	Histidine-Aspartate (nuclease domain motif)
HEPN	Higher Eukaryotes and Prokaryotes Nucleotide-binding
HMM	Hidden Markov Model
HNH	Histidine-Asparagine-Histidine (nuclease domain motif)
HTH	Helix-Turn-Helix (nucleic acid binding motif)
ICTV	International Committee on the Taxonomy of Viruses
IHF	Integration Host Factor
IMG/VR	Integrated Microbial Genomes/Virus Repository
IS	Insertion Sequence
JGI	Joint Genome Institute
MGE	Mobile Genetic Element
mRNA	Messenger RNA
NAD+	Nicotinamide Adenine Dinucleotide
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
OMEGA	Obligate Mobile Element Guided Activity

ORF	Open Reading Frame
PADLOC	Prokaryotic Antiviral Defence Locator
pAgos	Prokaryotic Argonaute systems
PAM	Protospacer Adjacent Motif
PDB	Protein Data Bank
Pfam	Protein Families database
PILER-CR	Parsimonious Inference of a Library of Elementary Repeats - CRISPR
PLD	Phospholipase-Like Domain
PPS	Priming Proto-Spacer
RecBCD	Recombinational repair genes B,C,D
R-M	Restriction-Modification
RNA	Ribose Nucleic Acid
RNP	Ribonucleoprotein
RT	Reverse Transcriptase
RuvC	Resistance to Ultra-Violet light gene C
SAVED	SMODS-Associated and fused to Various Effector Domains
SRA	Short Read Archive
ss	Single-Stranded
STING	Stimulator of Interferon Genes
tracrRNA	Trans-activating RNA
vConTACT	Viral CONTigs Automatic Clustering and Taxonomy
WYL	Tryptophan-Tyrosine-Leucine (nucleic acid binding motif)

Abstract.

Explorations of anti-phage defence systems, with a focus on CRISPR-Cas system diversity, in the past decade has enabled many significant advances in biotechnology especially in programmable gene editing platforms. Many previous investigations have employed guilt by association-based data mining approaches to discover new anti-phage defence systems. In light of these endeavours, it remains unclear to what extent undiscovered functional diversity linked to CRISPR mediated immunity remains to be explored.

In this study, I employed 3 separate yet interrelated avenues of inquiry to assess different aspects of CRISPR-Cas system diversity. I emulated past data mining approaches and constructed a computational pipeline which screened known and unknown putative CRISPR-associated genes from a large volume of assembled sequencing data. I then conducted a census of the landscape of genes co-encoded in proximity to array containing CRISPR-Cas subtypes. I found that a significant fraction of co-encoded genes possessed homology to anti-phage defence genes and mobile genetic elements even with low CRISPRicity scores, which has traditionally been used as the gold standard for proof of co-association with CRISPR-Cas subtypes. A small number of associated antiphage defence genes such as *HicAB* and *DrmB* were also found in novel association with Type VI CRISPR-Cas systems.

The second avenue I employed was to analyse the intra-subtype diversity using multi-gene gene-genome network based taxonomic representations. Focusing this approach on Type VI CRISPR-Cas systems, I observed a high degree of segregation between local clusters, with relatively few shared genes. Conversely, many additional genes predicted to be anti-phage defence related were found co-encoded at the level of local clusters. I then performed spacer mapping using spacers derived from the type VI systems CRISPR arrays. This revealed that local clusters of Type VI systems target a mixture of plasmids and phages, even in cases where the type VI CRISPR-array was found to be prophage encoded. As a consequence of red-queen competition between host-encoding CRISPR-Cas systems and mobile genetic elements (MGEs) such as phages or

plasmids, the taxonomic diversity of CRISPR associated genes and mapped sequences of their corresponding MGEs are inextricably linked. This justified constructing an analogous gene sharing network to model and analyse the diversity of these sequences. In contrast to the sequences of host-encoded Type VI systems, these sequences formed a contiguous and extensive network of many exchanged and inter-related genes.

Finally, a bioinformatic technique called “spacer distribution analysis” was employed to predict priming-like effects which have been shown to result from the coupling of interference and adaptation in certain CRISPR-Cas subtypes. This uncovered evidence of priming in type I-A and type V-F CRISPR-Cas systems. I also extended the sensitivity of spacer distribution analysis by including spacers with partial matches which increased sensitivity of spacer distribution analysis in certain subtypes. However, this increase was accompanied by differences in the spacer distribution observed.

In summary, my investigation discovered several novel gene co-associations and priming-like effects in several CRISPR-Cas subtypes which have gone unreported in previous studies. This suggests that pockets of functional diversity among known and unknown CRISPR-Cas systems remains to be explored.

Table of Contents

Chapter 1: Literature review

1.1 Phages are a potent cause of prokaryotic mortality

Among the most severe and continuous threats posed to the survival of prokaryotic organisms, both in the bacterial and archaeal kingdoms of life, are parasitic or predatory invasions of the cells by phages. Phages are viruses which infect bacteria. As with all viruses, phages consist of nucleic acids to encode instructions for their replication, which is encased in protective coat proteins to prevent degradation. Upon infecting a bacterial "host" cell, phages subjugate and repurpose the intrinsic machinery of the host cells for their own replication, usually at the expense of the cell's survival. Distinct clades of viruses also exist which infect and subjugate archaeal cells for their own replication in a functionally equivalent manner to phages.

Approximately 20% of all prokaryotic cell deaths are due to phage predation[1]. Furthermore, there are approximately 10^{31} phages extant within the biosphere[2,3] which is estimated to be up to 100 times more than the number of prokaryotes[4]. These comprise approximately 200 Mt of organic carbon or 20% of all estimated reserves of organic carbon by mass[3,5]. This makes phages the second most common organic automata present by mass on earth, behind only prokaryotic cells[3,6].

The genomes of phages are constructed from either DNA or RNA. Of these, DNA phages appear the most prolific, with approximately 90% of known phages being comprised of a double stranded (dsDNA) phages while only approximately 10% are ssDNA or RNA phages[7]. There have been several proposed explanations for this, both experimental and ecological. RNA phages are much more difficult to sequence and detect compared with their DNA phage counterparts[8]. Prior to 2019 only a dozen complete ssRNA phage genomes, and half a dozen dsRNA genomes, had been deposited on publicly available repositories[8]. This has since expanded dramatically, with tens of thousands of RNA phages now sequenced and identified[8]. A second explanation for the predominance of DNA phages is the larger diversity of mechanisms available to DNA phages to ensure

their own replication and long-term survival.

There are two broad means by which phages are able to replicate, to ensure their long-term survival (Figure 1.1). Once a phage has transfected its genome into the host prokaryotic cell, it may either opt for replication via the lytic cycle, or lysogenic cycle. In the lytic cycle, the phage genome directly repurposes the host cell replication machinery to ensure its own replication. This may be done in conjunction with phage encoded proteins such as polymerases, elongation factors, or cell-cycle control genes[9] to increase the efficiency of phage replication. An example of this is phage replication by T4, which is one of the main model phages of study, owing to its host being *E. coli*[10]. The T4 genome encodes its own DNA polymerase (*gp43*), helicase (*gp41*) and primase (*gp61*) genes, as well as RNase H exonuclease (*Rnh)* and ligase (g*p30*) genes for lagging strand DNA synthesis[10]. The genome undergoes rolling circle replication, producing a single long concatemer of copies of the T4 phage genome, which are subsequently cleaved and packaged by phage terminase proteins into the icosahedral head capsid[10,11]. The phage head, tail and fiber components self-assemble to form a mature bacteriophage[12]. Once the number of mature T4 phages reaches a critical threshold, phages then initiate lysis of the host cell membrane via phage holin proteins[13-15], enabling the mature phages to escape the host cell, and infect more *E. coli*.

In contrast to the lytic cycle, which relies on the direct replication of naked phage DNA once inside the host cell, in the lysogenic cycle, the phage instead integrates into the host genome, becoming what is known as a prophage[16]. The model phage in which this process has been most extensively studied is the ***λ*** phage, which infects *E. coli*. Integration occurs at attB sites in the host genome, and requires the *λ* integrase *(int)* protein and host encoded Integration-host factor (IHF)[16]. At a later period, after many host-cell generations, the prophage may self-excise itself from the host genome in response to certain environment stimuli such as host-genome DNA damage, forming a naked phage genome which may then escape the host-cell by entering the lytic cycle [17]. This process entails inhibition of λ repressor, the main operator gene which silences the expression of the genes required to enter the lytic cycle [18-20]. In this manner, lysogenic

phages use the host-cell as a form of shelter, tying their abundance to the fitness of the host-cell lineage and excising themselves when the viability of the host deteriorates.

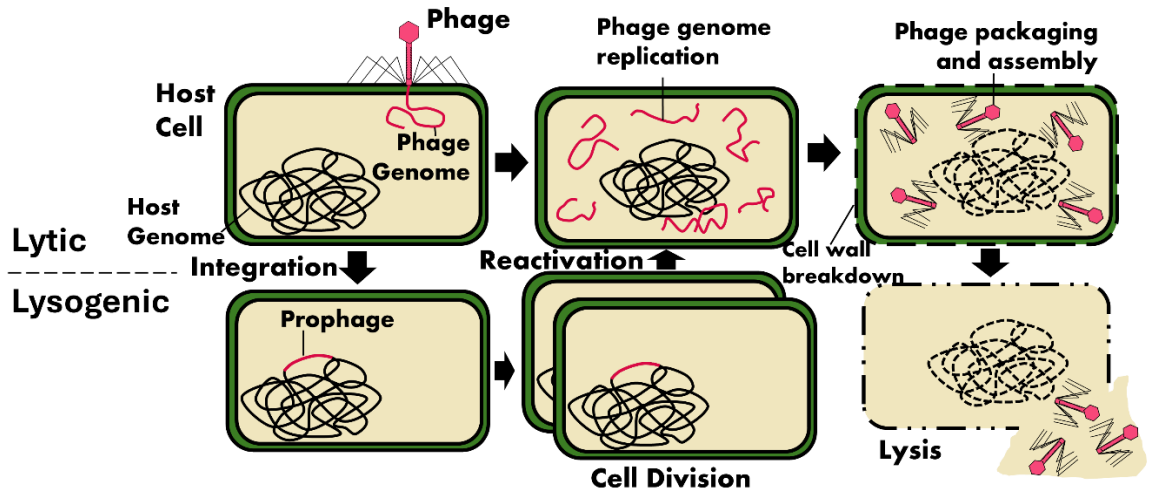


Figure 1.1: ***Overview of phage replication in prokaryotic cells.*** *(A) In the lytic cycle, an invading phage hijacks the host-cell replication machinery. This is then repurposed for genome replication and producing the core proteins required for complete virion packaging and assembly. After virion assembly, the final stage of the host cell is the disruption or degradation of the host cell membrane, killing the host-cell, and releasing virions to infect other cells. (B) In the lysogenic cycle, an infecting phage integrates it's genome into the host genome. This is then replicated along with the host, whenever the host replicates via mitosis. Some lysogenic phages may re-enter the lytic cycle by excising themselves from the host-genome and re-entering lytic replication in response to environmental stimuli.*

1.2 Diverse systems and mechanisms exist to confer phage immunity by both innate and acquired manners

In response to the threat posed by phages, prokaryotes have evolved an extremely diverse array of anti-phage defence genes. In common organisms derived from human gut microbiota, each organism has an average of 5.8 anti-phage defence systems comprising approximately 15.4 genes[21]. Given that surveys of antiphage defence genes only count known genes, this figure is likely an underestimate. These genes often occur together in large contiguous units called defence islands [22]. Despite this, these systems are highly mobile and are frequently exchanged within the same phyla, often resulting in

a diverse repertoire of systems present within the same species being investigated [23]. Despite diverse origins and functions, the mechanisms by which these systems confer immunity can be sorted into approximately three categories: Direct immune responses, Abortive infection and infection-exclusion (Figure 1.2).

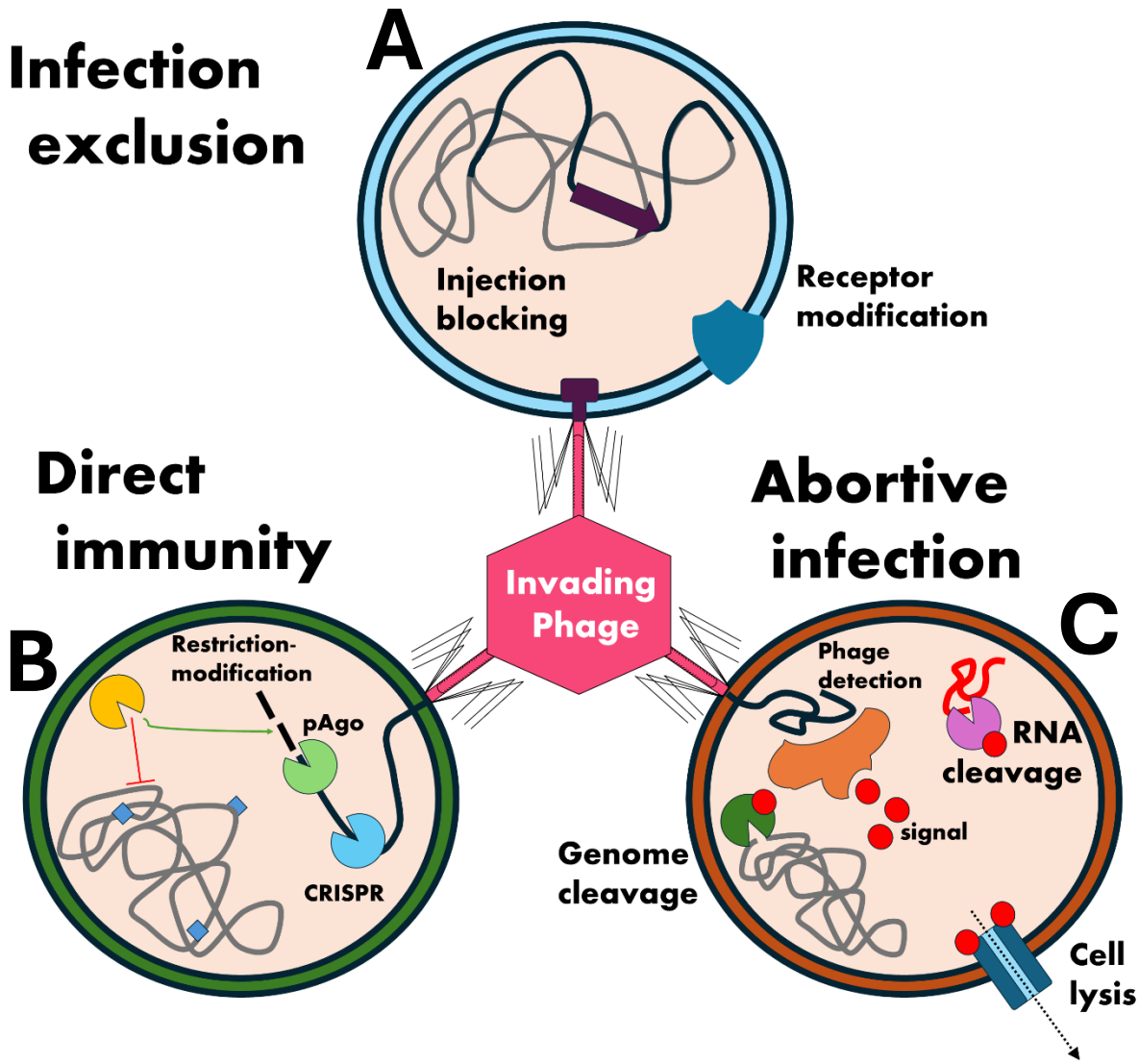


Figure 1.2: ***Outline of the main mechanisms by which antiphage defence systems confer immunity against phages and other invasive Mobile Genetic Elements.*** *(A) In infection-exclusion systems, an encoded prophage produces a protein that promotes lysogeny and prevents subsequent infection by the same species of phage. (B) Direct immune systems act to physically degrade invading phage components (i.e. by physical genome/mRNA transcript degradation). (C) To quarantine an infection and deny an invading phage the host-cell resources required to replicate and infect more cells of the same bacterial population, some anti-phages defence systems induce the deliberate destruction (abortive infection) of the infected cell.*

Direct immune responses

Anti-phage defence systems which induce a "direct" response, are those whose mechanism targets the degradation or replication of the phage genome, or phage virion components, as opposed to conferring enhanced fitness to the host cell as a consequence of intra-phage competition (Infection-exclusion) or inducing senescence or programmed cell death to quarantine an invading phage population or deny them of the resources of the infected cell for replication. Occurring in approximately 84% of genomes comprising a sample of all known taxa, restriction modification (R-M) systems are the most common class of antiphage defence systems[21]. The hallmark of R-M systems is the presence of a restriction endonuclease which relies on the methylation of the host-cell genome by methyltransferases[24 25-28]. These restriction enzymes can distinguish methylated-vs non-methylated DNA, which enables them to selectively target phage DNA lacking these modifications [24 25-28]. There are many methylation-sensitive restriction endonucleases, including *EcoKI, Hpy99I,* and *BanI* [24 25-27] which. introduce staggered dsDNA cuts based on 4-8bp recognition sites [28].

In addition to the canonical R-M systems, many systems have also been discovered which are variations on the same functionally equivalent mechanism. BREX (Bacteriophage exclusion) systems contain an adenine methyltransferase gene called P*glX* which methylates at GCTA**A**T sites[29]. Although methylation presumably protects the host genome like other R-M systems, there is no nuclease present to effect phage degradation [30]. Instead, the homology of an AAA+ ATPase gene called *BrxL* with eukaryotic DNA replication helicases has led to the hypothesis that *BrxL* interferes with the phage replication origin, preventing replication [31]. Similarly, the DISARM (Defense island system associated with restriction-modification) system encodes 4-5 genes. These consist of a helicase heterodimer (D*rmAB), PLD* domain *(DrmC)*, and an optional extra helicase (DrmD), optional unknown gene (D*rmE*), and either an adenine (D*rmMI*) or cytosine (D*rmMII*) methyltransferase [32]. Mutational analyses indicate that only D*rmAB and DrmE* were required for phage protection[32]. Interestingly, the essentiality of each gene differed based on the exact phage used, indicating different specialised responses depending on the type of invading phage[32]. Fluorescence microscopy studies observed DNA decay in invading phages, but its exact mechanism is still unknown [32,33].

Methylation is not the only means of shielding host-cell DNA from degradation. SspABCD-SspE and DndABCDE anti-phage defence systems employ phosphorothioate modification of the host genome phosphate backbone to enable endonucleases to degrade non-self DNA[34-36].[34,35]

Prokaryotic Argonaute systems (pAgos) are also notable for their ability to utilize guide sequences for the purpose of phage defence [37]. Prokaryotic Argonautes have been shown to load plasmid or phage derived RNA guides generated from the transcripts of the actively expressed genes of the invading plasmid/phage. Interestingly, a subset of pAgo guide RNAs have been shown to be derived from resected double stranded DNA breaks (DSBs) in direct cooperation with the replication execution checkpoint Rec[38]. This is consistent with observations of other pAgos which have also reported guides generated from the fragments of phage genomes [37]. While some pAgos have been shown to degrade DNA [38,39], many pAgos, especially short pAgos lack this ability, and instead recruit additional enzymes to facilitate DNA degradation[40], or rely on an alternative means of conferring immunity against phages [41,42].

Other mechanisms of anti-phage defence, though still common, are somewhat less ubiquitous. Some anti-phage defence systems localise to the cell membrane and degrade the phage genome at the point of injection. The Zorya defence system consists of 4 genes designated *ZorA-D*[22]. *ZorA* and *ZorB* express a membrane embedded complex which forms a long filament in the host cell[43]. The structure of the ZorAB complex is similar to that of other ion-driven rotary motors[43]. It is postulated that upon phage infection, the filament localizes to the infection site and spools the invading phage DNA[43]. The nuclease proteins ZorC and ZorD then facilitate degradation of the injecting DNA[43].

Abortive infection (Abi) systems

Another prominent type of anti-phage defence systems act to induce programmed cell death and dormancy (Abortive infection) in response to phage infection. The purpose of this is to quarantine the phage infection to the host cell and deny it the resources to further replicate[44] . There are many prominent anti-phage defence systems which have

been described inducing abortive infection in phage infected host-cells. One example is PARIS (Phage anti-restriction-induced system) anti-phage defence system. This system consists of two core proteins, *AriA* and *AriB*[45]. *AriA* contains an ABC ATPase domain and possesses some homology to the eukaryotic Rad50[45]. Another extremely common system is Gabija. Gabija consists of two genes, *GajA* and *GajB*[46]. These form a complex which is activated in the absence of ATP and inhibits phage replication [46].

There exists significant overlap between systems whose primary purpose is to induce programmed cell death, and those for which the primary goal is attrition. Attrition based Abi systems often deplete metabolites, such as NAD+, or degrade mRNA transcripts, preventing the use of the host-cell ribosomes for translation. Some systems, such as AbiEi-AbiEii and HicAB induce reversible bacteriostasis as opposed to programmed cell death[47]. In the HicAB toxin-antitoxin system, HicA and HicB form a complex (HicAB) which sequesters the RNase activity of HicA. [48-50] HicB also represses the expression of HicA[49]. Upon sensing phage infection, HicA is released from HicB and non-specifically degrades the host mRNA, thereby stunting the growth of the host cell and inhibiting phage replication[49].

Abortive infection mechanisms are often triggered by signalling cascades as a secondary line-of-defence [44] . R-M and other systems have been observed to possess secondary mechanisms to induce bacteriostasis or programmed cell death [44,51,52]. Among the most common mechanisms are cyclic-oligonucleotide-based anti-phage signalling systems (CBASS)[53]. The hallmark of CBASS antiphage immunity are genes which produce cyclic oligonucleotides from nucleoside-triphosphates in response to phage invasion[54]. These molecules act as signalling messengers to downstream response proteins, which induce programmed cell death via a variety of mechanisms[54]. The simplest CBASS systems (type I) consist of a cyclic GMP-AMP synthase (cGAS) and a phospholipase. Upon sensing an invading phage, the cGAS protein produces cyclic Guanine-Adenine monophosphate (cGAMP) which allosterically activates the phospholipase, triggering apoptosis via breakdown of the cell membrane[54]. This is not the only way cGAS based apoptosis takes place. Other cGAS systems instead encode a gene possessing a *SIR2* domain which becomes activated in response to cGAMP and

deplete NAD^+, slowing host-cell metabolism and denying the invading the lytic phage the resources for its replication[55].

Notably, a functional analog of CBASS based immunity exists in a subset of CRISPR-Cas systems (Type-III)[51]. These produce cyclic-oligo adenylate (cOA) in response to site-specific binding to phage transcribed mRNAs[51]. A critical concentration of cOA then allosterically activates effector proteins via binding their CARF (CRISPR-associated Rossmann fold) domains[56-59]. This appears to be accomplished via a wide variety of mechanisms, as these CARF domain proteins are often fused with other domains, such as HTH, HNH, and HEPN[60]. Such CARF-domain containing effectors include Csx1 and Csm6, both of which function by non-specifically degrading the host-cell RNA[58,59] and Cam1, which, in response to CARF-domain signalling, forms a transmembrane pore which induces cell death via membrane depolarisation[56].

Infection-exclusion systems

It has been shown that some prophage genes, such as *JBD26*, encode a Tail assembly blocker protein (Tab) which is a non-functional mimic of phage tail proteins. This is called phage self-exclusion[61]. In another case of prophage induced infection self-exclusion, the prophage encoded *SieA* gene expresses a protein which is able to directly block the injection of DNA into the host-cell by other phages[61]. This is a type of selfish adaptation by the phage to protect the host-cell at the expense of other phages.

CRISPR adaptive immunity

CRISPR-Cas are RNA-guided antiphage defence systems, which have revolutionised biotechnology by drastically simplifying the ease of gene editing. Extremely large diversity exists in the exact mechanisms by which CRISPR-Cas systems confer acquired immunity against past phage infections, which will later be discussed in more detail in Section 1.7. The process is divided into three distinct phases (Figure 1.3). In the acquisition phase, DNA fragments (protospacers) produced during the replication of an invading phage are uptaken, trimmed, and integrated by the conserved acquisition proteins Cas1 and Cas2 into tandem repeat DNA structures called CRISPR-arrays, which are encoded either within the host genome, or encoded on a resident plasmid within the host cell[62-70]. The CRISPR-array consists of three main elements: the leader

sequence, which encodes a promoter sequence for expressing the CRISPR-array as well as a partial recognition sequence to bind the acquisition proteins in addition to any cofactors required for site specific integration, such as Integration Host Factor (IHF); the CRISPR-direct repeats upon RNA transcription form hairpin secondary structures that serve as the main binding interfaces for interference proteins, while the phage DNA derived spacer sequences when transcribed act as guide RNAs (gRNAs)[63,65,71,72]. During protospacer integration, the protospacer first undergoes a pre-processing step by acquisition enzymes wherein the protospacer is trimmed to a certain size and any sequences constituting the protospacer adjacent motif (PAM) are recognized and trimmed from the end of the protospacer[72,73]. Integration then occurs at the leader-repeat junction which results in the duplication of repeat sequence, and the insertion of a new spacer flanked upstream and downstream by CRISPR-direct repeat sequences[74]. Both the spacer and repeat sequences are usually of a fixed length, which makes it possible to recognise CRISPR-array structures within host genomes irrespective of the sequences of either the repeats or the spacers[62,64]. Once the spacer has been integrated, it functions as both a record of past phage infections[63], as well as a means to build site-specific guide sequences for specific phage targeting if the same phage reinvades the host prokaryotic cell[63,65].

During the invasion/reinvasion of an invading phage, the CRISPR-array is then transcribed and cleaved into mature CRISPR RNAs (crRNAs)[71]. These RNAs may also be paired with a second RNA molecule called a trans-acting RNA (tracrRNA)[75]. In this case, the hairpin and guide RNA functions of the crRNA are segregated into two separate molecules, which form a dsRNA bridge through base-pairing interactions[75]. These crRNAs then form a ribonucleoprotein complex with effector proteins, which form a large surveillance complex to interfere with, and degrade phage RNA/DNA in a site-specific manner[76,77].

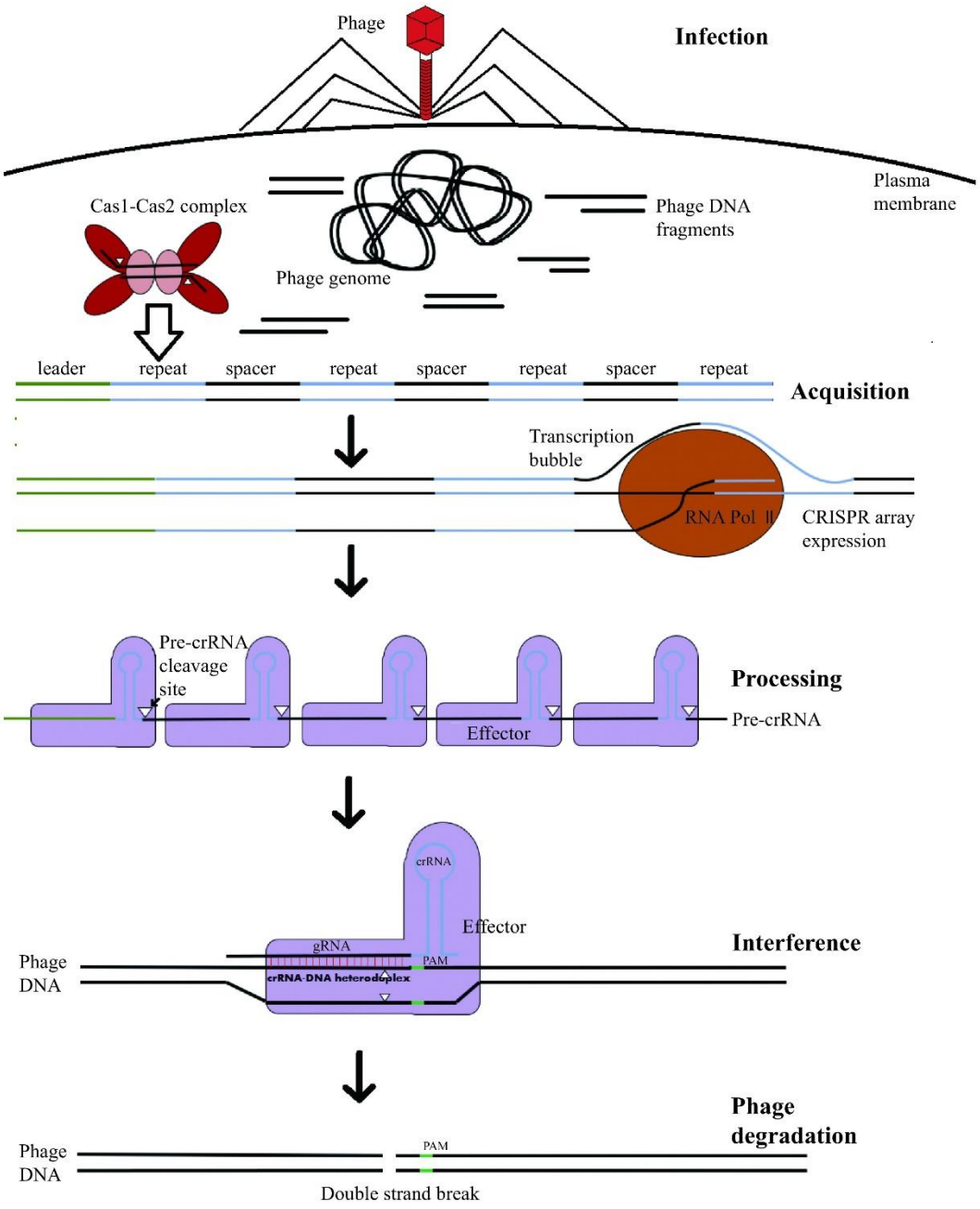


Figure 1.3: ***Generalised steps in CRISPR-Cas mediated acquired immunity.***

Acquired immunity comprises three general stages: acquisition, processing and interference. In the acquisition stage, fragments of DNA from an invading phage are uptaken by the Cas1-Cas2 protein complex and integrated at the leader-repeat junction of tandem repeats called CRISPR-arrays. During phage re-infection of the surviving cells from the original infection, these CRISPR-arrays are transcribed to form premature CRISPR RNA (pre-crRNA). Pre-crRNAs are then cleaved (processed) to mature crRNAs containing hairpin and guide RNA segments. The hairpin forms a complex with CRISPR-Cas effector proteins. These effectors then facilitate cleavage of phage nucleic acids at

sites determined by complementary to the phage derived guide RNA sequence from the original phage invasion. Note: Substantial functional variation exists in the mechanism of each stage, such as the absence/or differing compositions in the acquisition/interference proteins (see section 1.4) [57,78].

1.3 Guilt-by-association is an effective means of expanding the repertoire of known genes involved in phage defence

The most common means by which most antiphage defence genes have been discovered entails a reasonably standardised data mining approach called "guilt by association" (GBA)[21,22,57,79-89,90]. This approach depends on the observation that most anti-phage defence genes tend to be co-encoded in a single large pan-operon configuration called a defence island[22,81]. Hence if detectable homology to any protein within this island to a known defence gene exists, then the rest of the island can also be found.

The structure of a computational pipeline used to discover new genes by "Guilt-by-association" (GBA) can be generalised into five basic steps (figure 1.4)[22,81,91]. In the first step, a conserved feature or motif of the antiphage defence systems of interest called a bait or seed, is identified via pattern recognition tools which query a large block of genome sequencing data from whole sequenced genome (WGS), metatranscriptome or metagenome data repositories[22,81,90]. The motif recognition must occur with low computational time/complexity to be able to screen large volumes of assemblies. Previous GBA pipelines have employed a wide variety and number of seeds for detecting anti-phage defence systems in sequencing data[21,22,57,81-89 90]. The most common of these are Hidden-Markov Model (HMM) profiles of the most conserved genes in known systems[22,81,84,92]. However, other seeds, such as the tandem repeat motifs which comprise CRISPR-arrays s have also been employed [83,84,87-90,93-96]. Some investigations employ multiple seeds, depending on the type of antiphage defence system being searched for and the desired sensitivity/specificity trade-off [22,83,90] (using more seeds which are less conserved produces a higher false positive rate). Some recent investigations performed in the last two years have also used structure-based

searches. Since the development of Alphafold2[97], it has been possible to employ the 3D structures of domains derived from the protein translations of common genes as queries for searching on structure-based databases using tools such as Foldseek[98]. This has resulted in the discovery of additional clades of CRISPR-Cas systems from known types such as type VI, as well as expanding the known evolutionary diversity of other anti-phage defence systems such as BREX [99,100]. Although more sensitive than the equivalent HMM searches, it should be noted that the increase in sensitivity from this approach compared to HMM only resulted in the discovery of one additional clade of type VI CRISPR-Cas systems, indicating that HMMs searches already span a significant fraction of the search space[100].

After motif identification, the DNA upstream and downstream of the motif is then searched and parsed into open reading frame prediction software such as Prodigal or GenemarkS[22,83,84,91,101], to predict the set of protein coding genes encoded in proximity to the array. The translations of the predicted genes are then clustered into "putative" families[22,83,84,91]. These are then subject to further filtering criteria such as protein length, cluster size and co-associated proteins, which has the effect of reducing the incidence of falsely associated proteins or proteins unrelated to the anti-phage defence systems of interest[22,83,84,91]. This is also used to remove false positives such as highly abundant genes, and incorrect prediction of the CRISPR-array as a protein coding sequence ORFs predicted from the promoters of CRISPR-arrays which contain fragments of but are not antiphage genes of interest. This process results in a refined subset of candidate defence genes[22,83,84,91]. These genes are then further validated as anti-phage defence related via either annotation or experimental approaches[22,83,84,91]. In the annotation approach, each candidate defence gene is subject to further HMM/ structure-based searches by a database of well characterised domains, such as Pfam, or a custom database which contains a set of reference domains known to have functions in antiphage defence[22,90,99,100]. Alternatively, or in conjunction with annotation, predicted antiphage defence systems can be assayed in a generic manner at scale using xenotropic phage infection assays[22,30,32]. This involves synthesising and cloning the entire defence gene locus, or contig encoding the candidate defence protein and transforming the resulting construct into a model organism, usually *E. coli*. The ability of

the transfected system to infer anti-phage resistance is then tested by infection of the host-cell by a phage[22,30,32]. A protective function against the invading phage, or other mobile genetic element, is confirmed by improved survival of the transformed cells compared to the wild-type lacking putative defence system[22,30,32].

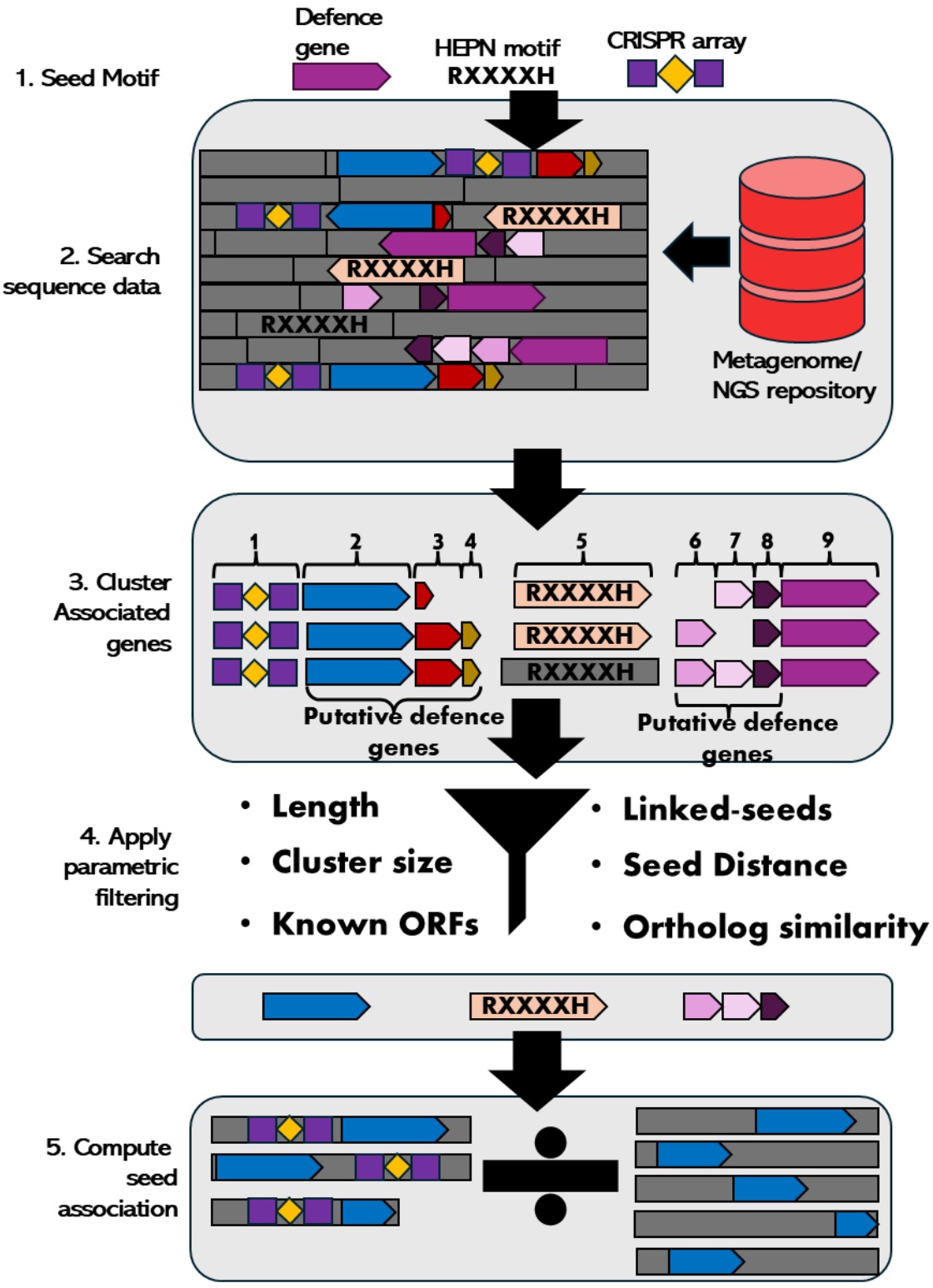


Figure 1.4: ***Generalised steps in the detection of anti-phage defence systems using the "Guilt by association" methodology.*** *A seed or bait motif is first chosen as a recognition motif to extract the target genes of interest co-encoded in proximity. The seed can comprise conserved genes, DNA elements, or a pattern (i.e. the HEPN RxxxxH motif). DNA is then extracted in a window upstream and downstream of the seed. These windows are then subject to Open Reading Frame (ORF) prediction and the resultant proteins clustered into families. After filtering these families by parametric criteria to screen for the target genes of interest a co-occurrence score is then calculated to prove co-association with the seed motifs.*

There are several important limitations to bear in mind when employing Guilt-by-association to detect CRISPR-associated genes. GBA implicitly assumes that all requisite CRISPR-associated genes are co-encoded in close proximity to the seed[90,91,102]. However, in some instances certain genes, such as Integration-Host-Factor and RNase III are critical for the functioning of type I acquisition and type II pre-crRNA processing respectively, are trans-encoded and thus not detected by GBA[74,75]. This approach also implicitly assumes accurate genome assembly, in order to be able to extract a neighbourhood of genes upstream and downstream of detected seeds, which reflect their real encoded loci in the native organisms from which the DNA was derived[90,91,102]. Sequence repeats, including the tandem repeat CRISPR-array structures, can cause genome mis-assembly[103]. Long-read sequencing data improves upon, but doesn't surmount this issue[104,105], and the majority of assembled sequence data in publicly available repositories remains assembled from short-reads, which is much more sensitive to the presence of sequence repeats[106,107].

Even when seed-associated genes are detected, the process of elucidation and assigning a functional role to associated genes is also very involved and prone to false-positive designations[91]. Parametric criteria such as CRISPRicity, average encoded distance from the seed and total abundance of a given gene in proximity to a particular seed motif used for mining are approximations rather than true predictors of co-inheritance and functional association[91]. These estimations are further distorted by the near-redundancy of many of the contigs used in the calculation of these criteria[108].

Methods such as representative sequence clustering have been shown to reduce these biases by deduplicating strongly related sequences[102]. However this method removes the ability to conduct an accurate census of the conserved frequencies any genes present within the assembled data, and the remaining sequences, when used to compute parametric scores, return estimates determined by the degree of clustering[102].

Once an association has been predicted, assigning the function of a given associated gene based on its corresponding protein domain/sequence/structural homology to domain/structural motif databases is also often error-prone. This is due to many proteins encoding the same domain yet employing this domain for unrelated functions[83,102,109,110]. Nevertheless, establishing this preliminary designation is often necessary if an investigator wishes to design and employ assays for experimental characterization[84,88,89,102,110,111].

1.4 CRISPR-Cas systems are classified via a parsimonious approach using signature genes.

Competition between phages and their prokaryotic hosts has induced a vast radiation of different systems that have evolved. A parsimonious classification scheme which distinguishes systems based on the properties of the surveillance complexes used in interference differentiates CRISPR-Cas systems into class, type, and subtype. Class I CRISPR-Cas systems utilise large multi-subunit complexes to degrade either DNA or RNA from the invading phage in a processive manner. Class II systems utilise a single monomeric protein to carry out interference (in complex with a crRNA). Below class, each system is defined by its “type”, which refers to the specific effector protein or complex used to carry out phage interference. Types are further divided into sub-types, which group orthologous effector proteins.

Origin of CRISPR systems

Different CRISPR-Cas systems utilise different effector proteins which have convergently evolved to facilitate RNA guided interference. The origin of many of these effectors is thought to stem from genes encoded in antiphage defence, or genes

encoded on transposable elements [112] (figure 1.5). Class 1 systems consist of three known types (I, III and IV)[57]. Each of these types relies on a multi-subunit surveillance complex to facilitate RNA-guided interference[57]. It is thought that this enzyme originated from an ancestral signalling system, with some similarity to currently observed cGAS-STING systems, which facilitate immunity via the induction of cAMP synthesis in response to binding of MGE-derived RNA[113]. This ancestral system is postulated to consistent of three elements: An RNA binding cOA cyclase sensor which was the primordial version of a modern Class 1 subunit, a CARF domain receptor to detect cOA and an HEPN domain which degrades RNA in response to allosteric binding by cOA[113]. Type III CRISPR-Cas systems are direct descendants of these systems and may have emerged from gene duplication and differentiation of the *Cas10* gene which then acquired the ability to bind small RNAs[113]. There are two distinct lineages, type III-A and type III-B, of RNA-guided multisubunit complexes (Csm and Cmr) in type III systems. The hallmark of Type III immunity is RNA-guided binding to transcribed phage RNA followed by co-transcriptional RNA degradation[58,114-118]. The other Class 1 types are functionally divergent ancestors from these early type III systems which have lost cOA signalling activity but retain otherwise functional domains[57,113]. Type I systems perform DNA as opposed to RNA-interference while type IV systems have evolved specific genes such as *DinG* for plasmid clearance, although the exact mechanism of this has not yet been determined[119].

Class 2 systems utilise monomeric effector proteins which have evolved from two principal domain superfamilies, RuvC and HNH. Type II and Type V systems are thought to have evolved from insertions in *TnpB* genes which are frequently encoded on *IS605* transposable elements[57]. All TnpBs possess a RuvC nuclease domain[57,120]. Interestingly, this works in concert with an RNA guide derived from the transposon 3' end to achieve site-specific strand cleavage[121,122]. This mechanism appears to pre-date an association with CRISPR arrays[122]. *TnpB* genes associate with IS200 transposable elements and use RNA-guided cleavage of the transposon excision site to ensure transposon retention across generations[121]. Interestingly, evolutionary intermediates have been isolated, which bear significant homology and similar size to TnpBs, yet utilise crRNAs[122,123]. This

demonstrated that RNA-guided DNA targeting is an activity of TnpBs that predates a specific role in phage defence[120,122,123].

Type II systems descend from IscB transposons, which are believed to themselves descend from IS605 transposons. The main distinguishing feature of ISCBs is the presence of an HNH endonuclease domain, which is postulated to be the consequence of a historical HNH endonuclease insertion event, in which a HNH endonuclease, potentially derived from an Abi or R-M antiphage defence system inserted itself into the *TnpB* gene. As with *TnpBs*, ISCBs are also capable of RNA-guided cleavage using guides encoded on the same transposon[124]. The evolutionary barrier for TnpBs and ISCBs co-opting crRNAs in place of their native guides was thus very low given that crRNAs and the guide RNAs used on these TEs have converged to possess almost the same structure[125]. This may explain why so many variants of Type V systems are thought to have evolved independently from IS605 derived TnpBs[124]. These two categories of TnpB derived RNA-guided systems are collectively designated as Obligate Mobile Element Guided Activity (OMEGA) systems[124].

In contrast to Type II and Type V systems, Type VI CRISPR-Cas systems employ an effector protein (Cas13) for RNA-guided RNA targeting[57]. This gene contains two HEPN domains which facilitate RNA cleavage. Interestingly, some Cas13 orthologs (Type VI-D) possess bystander cleavage in addition to RNA guided cleavage, which induces strong cytotoxicity consistent with an Abi system. Cas13 effectors also contain a second HEPN domain, which is thought to have been gained by a second insertion event, although the origin of this domain is as yet undetermined[100].

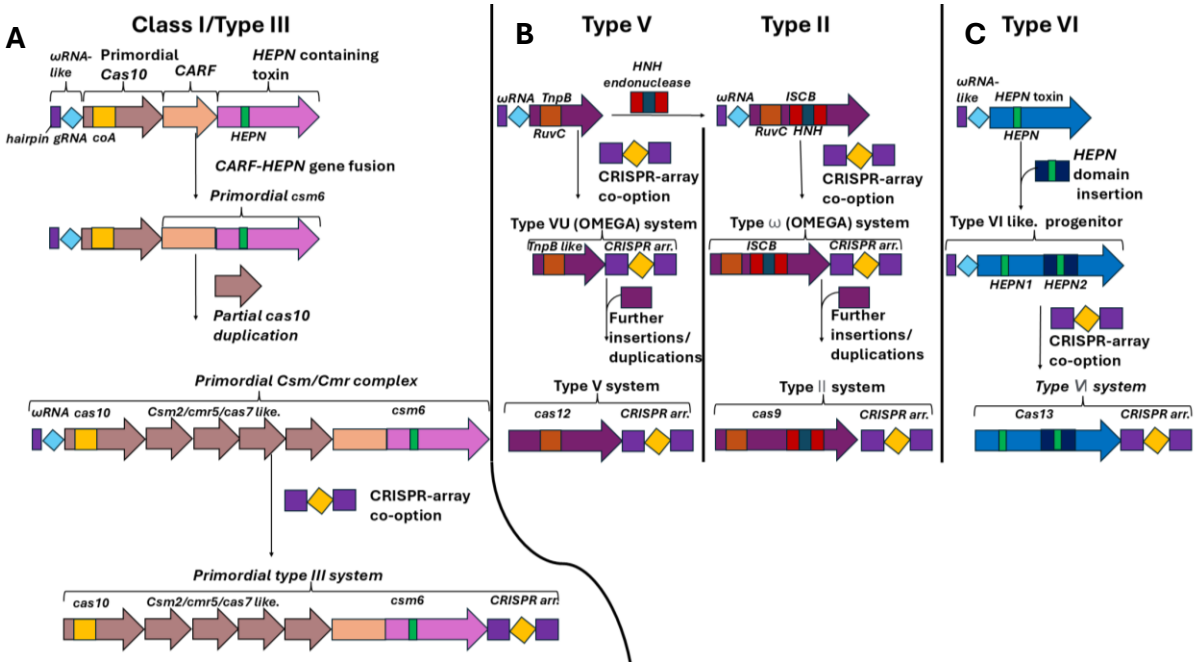


Figure 1.5: ***Origin of different surveillance modules used to facilitate interference in different CRISPR-types.*** *(A) Class I systems (Type III systems shown above) are postulated to have evolved from an oligonucleotide signalling system comprising an RNA sensing and cOA signalling gene (now encoded on the cas10 subunit of type III systems) and a cOA activated (via the CARF domain) toxin-antitoxin gene. The Csm and Cmr complexes then emerged from cas10 gene duplications and expansion followed by CRISPR-array capture and incorporation of crRNAs into Csm/Cmr target recognition. (B) Type II and type V systems evolved via gene expansion of RNA-guided IscB and IS605 TnpB genes. (C) Type VI systems evolved from the gene insertion of a second HEPN domain into an HEPN domain containing toxin-antitoxin encoding gene, producing a hybrid gene with two HEPN domains. All systems then co-opted and used CRISPR-array based RNAs which functionally displaced non-programmable small RNAs (ωRNA in Type II/Type V systems). The structure and details of this figure is based on "Origins and evolution of CRISPR-Cas systems," by Eugene V. Koonin & Kira S. Makarova, 2019, Philosophical transactions of the royal society B", Figures 3 & 5) (https://doi.org/10.1098/rstb.2018.0087) and "The Widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases", by Zhang et. al, 202, Science, (DOI: 10.1126/science.abj6856).*

Each of the seven CRISPR-Cas types can be further classified into subtypes based on the presence of co-encoded accessory genes, or variants of the conserved core-Cas genes required for acquisition, processing, or interference[57,102]. This scheme is not entirely consistent, with poorly established connections between co-encoded genes. Additionally, subtype differentiation may be performed by a novel effector gene which is functionally analogously to, but not necessarily divergent from other subtypes of the same type (i.e. different type V subtypes may both descend from TnpBs, but not necessarily a common ancestral TnpB gene)[57]. Understanding the distinguishing features of each subtype is thus best served by a review of each subtype in the context of their parent types.

Evolutionary diversity of class 1 CRISPR-Cas systems

Type I systems encode a surveillance complex which, unlike the presumptive ancestral system from which it derives, lacks any observed cyclic oligonucleotide signalling as observed in type III systems[126-128]. Instead, these systems utilize a multisubunit interference complex called CASCADE to bind dsDNA in an RNA-guided manner[129] (Figure 1.6). Upon target binding the CASCADE complex recruits Cas3, which possesses ATP-driven 3' to 5' helicase activity as well as nuclease activity mediated by an HD nuclease domain[128], which is heavily divergent from the HD domain seen in many Cas10 subunits from type III-A systems[115]. Cas3 then processively degrades the target DNA in the 3' to 5' direction producing small ssDNA fragments[128].

There are several key functionalities which vary between individual type I systems. Most type I subtypes are classified according to the genes which form the main CASCADE complex used for interference, as well as the Cas3 protein used for DNA strand degradation. Type I-D systems are a notable exception to this architecture. They encode a Cas10 protein which is an evolutionary intermediate between the Cas10 protein observed in type III systems and the Cas3 observed in type I. The CASCADE-Cas3 complex can be encoded by as many as seven genes (including the pre-crRNA processing enzyme Cas6, which is often part of the complex) in type I-A or as few as

four genes in type I-C. The mode of Cas3 strand degradation may also be monodirectional (Type I-B, I-C, I-D, I-E, I-G) or bidirectional (Type I-A, I-F). Many of these systems (Type I-B, type I-C, type I-E, type I-F) have additionally been shown to be capable of primed spacer acquisition. The directionality of acquisition tends to follow that of interference[127,130-133]. Finally, a subset of type I-B and type I-F systems have been further classified as CRISPR-associated Transposons (CASTs) [134-136]. In place of the Cas3 enzyme used for strand degradation, these systems interact with Tn7 transposons to guide the transposon integration in the RNA-guided manner [134-136].

Type III systems also encode a large multi-subunit surveillance complex (Csm/Cmr) but bind RNA targets co-transcriptionally from actively expressed genes unlike Type I systems which directly interfere with dsDNA [115,117,118]. Some Type III system variants also utilise a HD domain nuclease to perform strand degradation of the DNA on the complementary strand concurrent to RNA binding and degradation[118,137]. Guide RNA binding triggers the activation of the PALM polymerase domain (present on the cas10 subunit of the csm complex) which stimulates the production of cOA[51]. This activates other CARF domain-containing anti-phage defence proteins, leading to the recruitment of RNA strand degradation proteins which degrade the RNA strand. In some cases, additional proteins involved in abortive infection are recruited, such as Card1, Cam1, and Csx1 which can induce growth arrest in the host-cell[56,138,139].

As with Type I systems, there exists significant subtype diversity among type III, with at least five distinct type III subtypes (A-F) identified[57]. The two most abundant, type III-A and III-B, are differentiated by two distinct sets of weakly related genes, which form the Csm and Cmr surveillance complexes respectively[57]. Type III-A systems are differentiated by their strong conservation of CRISPR-arrays. Type III-B systems by contrast, tend to co-occur more strongly with proteolysis-based Abi induction enzymes, such as CHAT-SAVED domain and PCaspases[140,141]. This may imply a strong role for Abi pathways in phage defence compared with the adaptive immunity observed in type III-A[140]. Type III-C systems lack active cyclase domains, which suggests they rely on other means apart from Abi pathways to facilitate immunity. Meanwhile type III-D systems are predicted to completely lack HD nuclease domains, suggesting they rely entirely on RNA-targeting in conjunction with accessory genes to confer an immune response[140]. In

contrast, type III-F systems are predicted to possess this domain, but lack signalling and downstream effector responses, suggesting these effectors exclusively perform RNA-guided RNA/DNA targeting[57,140]. Type III-E systems are unique in that the surveillance complex is a single fusion protein comprised the Csm2, Csm3, Csm5 and Cas7 and Cas11 subunits[140]. Like type III-A systems these systems always use CRISPR-array derived guide RNAs [57,140].

Unlike type I and III systems, type IV systems are much less well characterized and rarer [57,142]. These systems are often plasmid encoded and appear to provide immunity against other types of plasmids[142]. The main surveillance complex used for interference consists of four subunit genes and is predicted to be most closely related to CASCADE[143,144]. However, these systems lack an HD domain nuclease and instead rely on an as yet undetermined mechanism to confer phage resistance[57,142]. There are five known subtypes, and a further classification level representing gene neighbourhood intra-subtype variation in type IV-A[142]. Only subtypes IV-A, IV-B and IV-D contain conserved co-encoded CRISPR-arrays [142]. Recently discovered type VII systems use a guided ribonucleoprotein complex mediated by Cas5 and Cas7 and an RNA targeting system mediated by Cas14[102]. Despite distinct cleaving activity, type VII systems are thought to have evolved from type III systems[145].

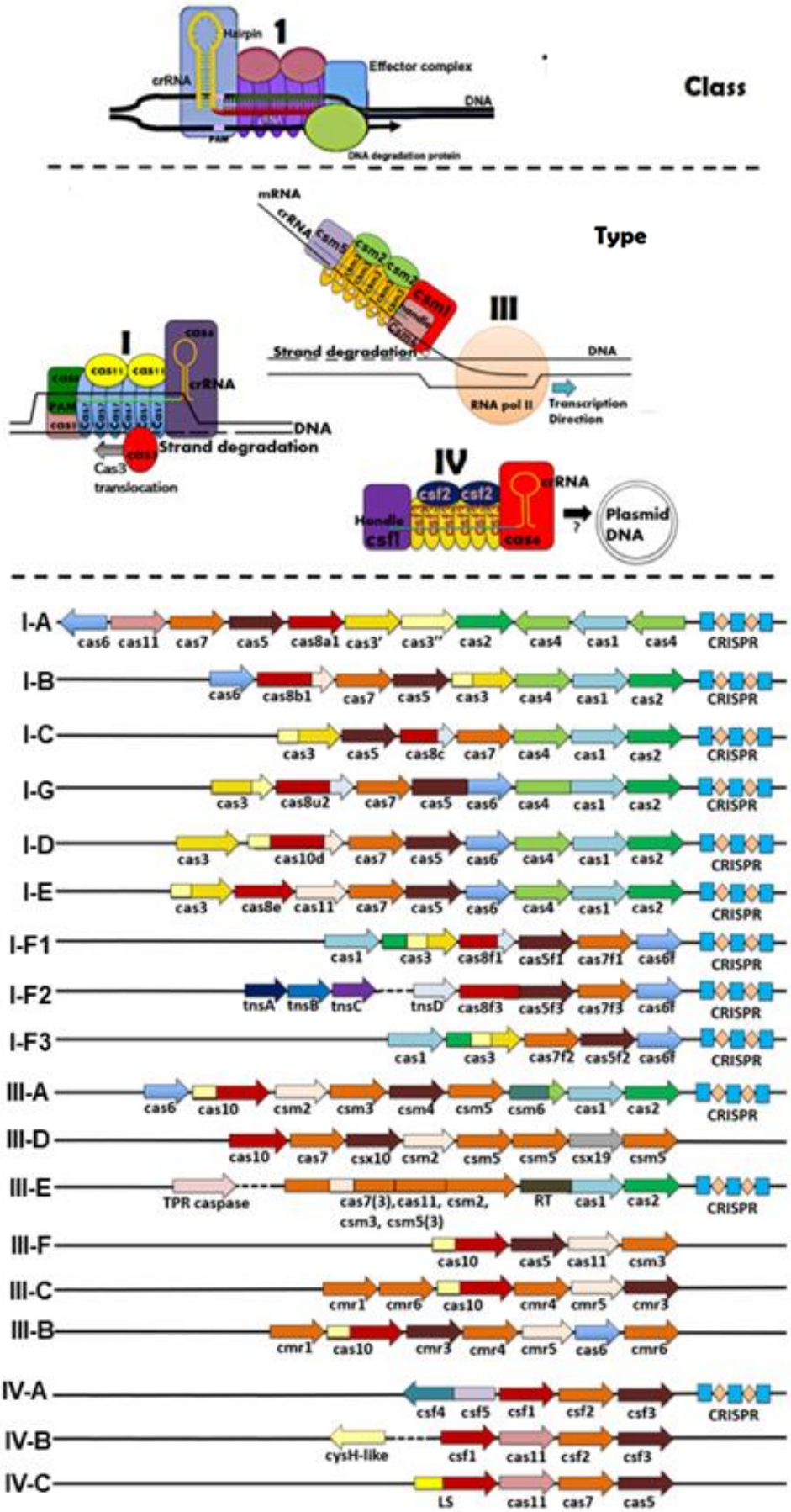

Class
1
Hairpin
crRNA
Effector complex
DNA
DNA degradation protein
Type
mRNA
crRNA
Csm5
Csm2
Csm1
III
Strond degradation
DNA
RNA pol II
Transcription Direction
I
crRNA
DNA
Strand degradation
Cas3
translocation
IV
csf2
csf2
crRNA
Handle
csf1
Plasmid DNA
I-A cas6 cas11 cas7 cas5 cas8a1 cas3' cas3'' cas2 cas4 cas1 cas4 CRISPR
I-B cas6 cas8b1 cas7 cas5 cas3 cas4 cas1 cas2 CRISPR
I-C cas3 cas5 cas8c cas7 cas4 cas1 cas2 CRISPR
I-G cas3 cas8u2 cas7 cas5 cas6 cas4 cas1 cas2 CRISPR
I-D cas3 cas10d cas7 cas5 cas6 cas4 cas1 cas2 CRISPR
I-E cas3 cas8e cas11 cas7 cas5 cas6 cas4 cas1 cas2 CRISPR
I-F1 cas1 cas3 cas8f1 cas5f1 cas7f1 cas6f CRISPR
I-F2 tnsA tnsB tnsC tnsD cas8f3 cas5f3 cas7f3 cas6f CRISPR
I-F3 cas1 cas3 cas7f2 cas5f2 cas6f CRISPR
III-A cas6 cas10 csm2 csm3 csm4 csm5 csm6 cas1 cas2 CRISPR
III-D cas10 cas7 csx10 csm2 csm5 csm5 csx19 csm5
III-E TPR caspase cas7(3),cas11, csm2, csm3, csm5(3) RT cas1 cas2 CRISPR
III-F cas10 cas5 cas11 csm3
III-C cmr1 cmr6 cas10 cmr4 cmr5 cmr3
III-B cmr1 cas10 cmr3 cmr4 cmr5 cas6 cmr6
IV-A csf4 csf5 csf1 csf2 csf3 CRISPR
IV-B cysH-like csf1 cas11 csf2 csf3
IV-C LS cas11 cas7 cas5

Figure 1.6: ***Taxonomic classification, mechanism of interference, and subtype diversity of class 1 CRISPR-Cas systems.*** *Class 1 systems are further delineated by "type" (I, III and IV) where each type represents a separate lineage of multisubunit interference complex encoding genes. Each type is further categorized in subtypes based on variations in the gene composition of each system within each type. This consists of unique genes, or fusions of several separate conserved genes in other types. For instance, some Type I-A systems contain two cas3 genes and separate cas1-cas2 acquisition genes, whereas in type I-F systems only possess one cas3 encoding gene which is a fusion of the two cas3 genes in type I-A systems as well as the cas2 gene. This is alluded to by the hybrid colour schemes of genes in some subtypes shown above. The structure of this figure was adapted from "Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants," by Koonin et. al, 2020, Nature reviews microbiology, Figure 1) (*https://doi.org/10.1038/s41579-019-0299-x*)*

Evolutionary diversity of class 2 CRISPR-Cas systems

The subtypes of many Class 2 systems, with the exception of type II systems, are underpinned by a continuum of signature monomeric effector proteins capable of interference, with only distant structural homology between each lineage of effector from the same type (Figure 1.7).

Type II systems are distinguished by the presence of Cas9, which have both HNH and RuvC nuclease domains, as well as consistently co-encoded adaptation proteins. There are three subtypes: Type II-A systems encode an additional gene, C*sn2*, which allows C*as9* to process pre-spacer fragments during spacer adaptation. Type II-B systems encode C*as4*, which has also been shown to be involved in pre-spacer processing in some organisms. Subtyping does not necessarily reflect C*as9* diversity. There is a divergent clade containing the Type II-B *Fransicella novicida* C*as9* ortholog while co-encoded C*as4* orthologs also exist in the primary clade, suggesting these do not explain the divergence. Type II-C systems are a diverse subtype comprising systems that lack pre-spacer processing genes, which is believed to be performed solely by the Cas1-Cas2 acquisition proteins.

Cas12 effector proteins of all type V CRISPR-Cas subtypes are defined by their single RuvC nuclease domain[57]. This type spans a vast continuum between large Cas12 genes consistently associated with CRISPR arrays, and small TnpB genes that may be coincidentally proximal to CRISPR arrays[57,123,124]. TnpB and ISCB orthologs that are capable of interference in an RNA-guide manner without CRISPR-arrays are designated OMEGAs and classified separately from other type II and V subtypes[124]. A *de facto* rule of Type V sub-typing is that there be no discernible sequence similarity between orthologs of different families[57,84]. However, this does not apply at the level of structural similarity, and all effectors possessing a RuvC nuclease domain may be used for DNA cleavage, RNA cleavage (Type V-G), and even crRNA processing (Type V-J) while still being counted as the same type[57]. These systems also contain conserved co-encoded CRISPR arrays[57].

The largest orthologs of cas12a, present in type V systems [type V-(A-E)], co-encode both CRISPR-arrays and adaptation proteins[57]. Many of the other subtypes discovered correspond to a CRISPR-array immobilised adjacent *TnpBs*, which lack acquisition proteins due to presumably not requiring them for transposition[84,123,124]. There is not a clear functional distinction between the attributes of many CRISPR-array immobilized *tnpB* orthologs and larger more mature Cas12 variants[124]. This explains how so many Cas12 orthologs could evolve independently from different clades of the same *tnpB* superfamily. Similar to type I systems, some type V subtypes, such as type V-K systems[134,146], have been shown to be nuclease inactivated, and instead form a complex with a *tnsB* transposons to integrate these elements in an RNA-guided site-specific manner[146].

Unlike type II and V systems, type VI systems are differentiated by the presence of the dual-HEPN domain containing Cas13 protein as the primary means of subtype differentiation[147]. These effectors cleave ssRNA as opposed to DNA, similar to type III subtypes[87-89,147]. Compared with other subtypes, the co-occurrence of spacer acquisition genes occurs infrequently, mainly in type VI-A and some type VI-D systems. Given that the CRISPR-arrays still appear to acquire new spacers, it has been hypothesised and shown in a few model systems that these systems may rely on trans-encoded acquisition genes for adaptation [148]. Some type VI systems are differentiated

based on the co-encoding of several accessory proteins [87,88]. Type VI-B proteins often encode one of two genes, *Csx27* and *Csx28*[87]. *Csx27* has been shown to be membrane bound, and represses Cas13b activity[87], while Csx28 has been shown to act as a membrane depolarisation protein, demonstrating that the combined system has Abi properties[149]. Type VI-D proteins characteristically co-encode a conserved WYL domain protein, which acts to enhance Cas13d mediated RNA cleavage [88]. Additionally, several other subtypes such as Type VI-X and Type VI-Y have been proclaimed yet later shown to be smaller, divergent variants of Type VI-B systems[150]. These were later classified as divergent Type VI-B effectors[102]. Type VI-C effectors, despite possessing clear domain homology to the other effects, have not yet been experimentally characterised, leaving the unique aspects of their activity still unknown.

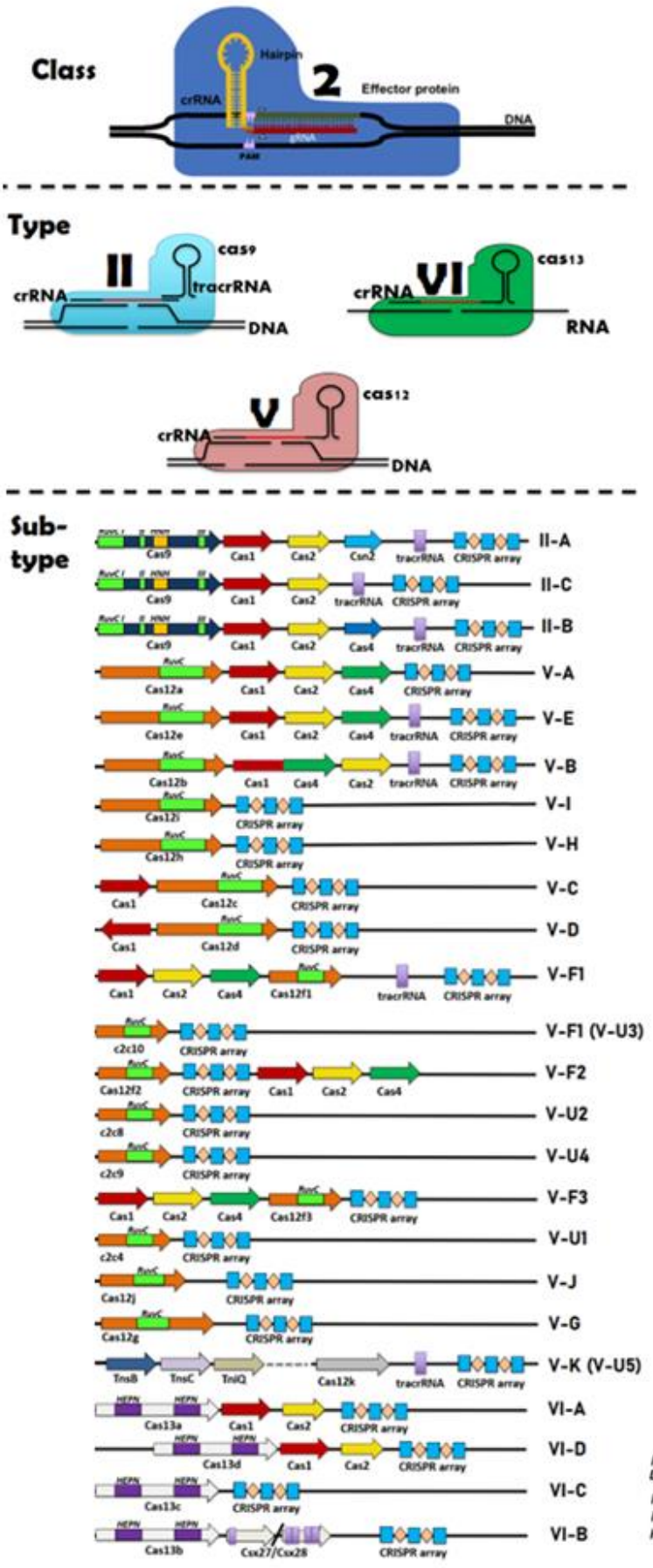

Class
2
Hairpin
Effector protein
crRNA
DNA
gRNA
PAM
Type
II
cas9
tracrRNA
crRNA
DNA
VI
cas13
crRNA
RNA
V
cas12
crRNA
DNA
Sub-type
Cas9
Cas1
Cas2
Csn2
tracrRNA
CRISPR array
II-A
II-C
Cas4
II-B
RuvC
Cas12a
V-A
Cas12e
V-E
Cas12b
V-B
Cas12i
V-I
Cas12h
V-H
Cas12c
V-C
Cas12d
V-D
Cas12f1
V-F1
c2c10
V-F1 (V-U3)
Cas12f2
V-F2
c2c8
V-U2
c2c9
V-U4
Cas12f3
V-F3
c2c4
V-U1
Cas12j
V-J
Cas12g
V-G
TnsB
TnsC
TniQ
Cas12k
V-K (V-U5)
HEPN
Cas13a
VI-A
Cas13d
VI-D
Cas13c
VI-C
Cas13b
Csx27/Csx28
VI-B
Key Domains:
RuvC
HNH
HEPN

Figure 1.7: ***Taxonomic classification of class 2 CRISPR-Cas systems.*** *There are 3 separate types (II, V & VI). These are further subclassified into subtypes. Class 1 subtypes, which are defined by unique operon level compositions of conserved co-occurring genes (figure 1.6). In addition to this means of classification, class 2 systems, particularly type V systems, are also split into multiple subtypes if the encoded interference genes possesses no detectable similarity to other orthologs yet still possess the same conserved functional domains (RuvC domain in type V systems). The structure of this figure was adapted from "Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants," by Koonin et al, 2020, Nature reviews microbiology, Figure 2) (*https://doi.org/10.1038/s41579-019-0299-x*)*

1.5 Reconciling parsimonious-subtype classification with orthology-based taxonomic classification of CRISPR-Cas systems

Inconsistencies in CRISPR-Cas subtype classification using parsimony based on co-encoded accessory genes.

The current parsimonious classification scheme for designating CRISPR-Cas subtypes assumes the presence of conserved co-evolving genes encoded in proximity (i.e. as part of the same operon) to the effector gene or cassette used for typing[57]. However, many of the individual genes of CRISPR-Cas subtypes have evolved "quasi-independently" [90]. This means that any phylogenetic construction based on a single gene in each CRISPR-Cas system will invariably fail to represent the evolutionary history of co-encoded genes[90]. The lack of co-dependency between the conserved core genes in CRISPR-Cas systems creates low evolutionary barriers to the recombination, insertion, and removal of individual genes between different CRISPR-Cas systems, which increases the number of independent origins of different genes with each system[90,91]. This means that co-encoded genes cannot reliably be used as distinguishing markers in many cases, as these genes may have a different evolutionary history from the rest of the system at large[81,90]. This also applies to conserved proteins such as the acquisition proteins Cas1

and Cas2, whose evolutionary history has been shown to depart from that of associated interference genes used for typing within certain subtypes[81]. The existence of multiple independent ancestors as a result of these horizontal gene transfer events violates the primary assumption of traditional phylogenetic methods, which root a tree based on a single common ancestor[2,151]. This requires a different classification structure to accurately model these evolutionary events.

The difficulties classifying CRISPR-Cas subtypes resembles that of phage classification

In many respects, the issues which affect taxonomic classification of phages are also pertinent to CRISPR-Cas subtyping. Phages taxa can share protein homologs, which may be used to construct an evolutionary history[151,152]. Additionally, limited phage labelled available from the ICTV poses additional challenges for taxonomy level classification[153]. Frequent gene exchange, reshuffling, and recombination between MGEs also means that many of the genes within these contigs contain genes that function independently of each other[154]. Without the underlying assumption of co-evolution, co-dependency or strong conservation between genes, it is impossible to construct an evolutionary history from a single gene using traditional phylogenetic methods which also reflects the evolutionary history of all other genes and associated elements on the same MGE[155]. This has spurred the development of classification methods which utilise the entire contig/genome for classification, as opposed to using just the most conserved elements[57,151,156].

Consistent phage classification schemes utilise unsupervised gene-genome gene cluster networks

To better represent the evolutionary history of MGEs, several alternative models have been developed which represent a rational improvement over single gene phylogenetics. Early attempts at phage classification used BLAST based identification of overlapping genes[157]. These were used to build trees from sequences of phages which showed the evolutionary relatedness between clades of phages identified from viral metagenomics[158].

More recent tools rely on bipartite gene-genome networks. In these models, the edges of different genomes sequences are linked to each other through shared genes. Edges are generated using all-by-all BLAST searches against ORFs predicted from each phage contigs[151,156]. Local communities within the bipartite network may be optionally defined by the use of Louvain or related cluster-partitioning methods, which define local clusters within networks in a heuristic manner based a recursive optimisation of the ratio of edge densities in local groups compared with edge densities outside these groups [151,159]. These network-based methods have been used to accurately classify the dsDNAphage virosphere at the family level and identify 14 conserved essential genes across 1,073 phage genomes[151].

The most modern tools developed to date incorporate several improvements upon these bipartite network-based approaches. The most recent widely used tool, vConTACT2, builds on these approaches by incorporating more efficient, recently developed search similarity search tools, such as DIAMOND and mmseqs2, in addition to BLAST [152]. It also optionally includes the ability to overlay a network-based reference set of viral sequences atop a custom generated network [152]. This enables the identification of known and unknown clades if the reference network overlaps sufficiently with the network of interest [152]. Finally, vConTACT2 is highly scalable, able to run up to one million sequences in approximately linear time complexity[152].

1.6 Investigating host-phage interactions: insights from the CRISPR arrays.

The first solid evidence that CRISPR-Cas systems contained a record of past phage and infections came from applying sequence similarity searches (via BLASTn) to assembled sequence data repositories[63,65]. Spacer mapping has been applied early in the history of CRISPR discovery to identify the MGE complement of a given CRISPR-Cas systems and investigate the function/ecology of CRISPR targeting[63,65].

To date, compelling evidence exists that CRISPR-Cas systems predominantly target dsDNAphages, ssDNAphages, jumbo phages, prophages and plasmids (Figure 1.8)[160-162]. Interestingly, outside of pre-programmed protection in *E. coli* there are no instances

of CRISPR-mediated interference against RNA phages[160]. While the capsid structure of jumbo phages has been shown to shield the replicating phage genome from DNA targeting CRISPR-Cas subtypes, transcribed genes remain susceptible to RNA targeting from type III and VI systems[160,162]. There is also conflicting evidence on the role CRISPR-Cas subtypes play in the targeting of ssDNAphages and plasmids. CRISPR spacers have been mapped to inoviruses[152] but ssDNA DNA cleavage by *Cas12a* in Type VI-A systems was shown to be insufficient to elicit immunity[163].

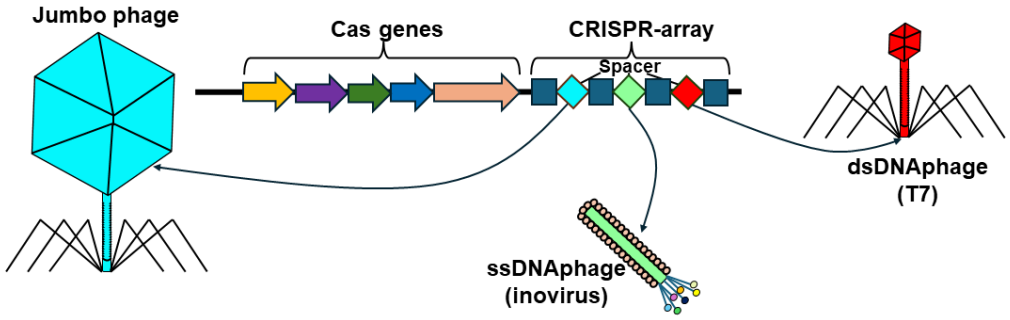


Figure 1.8: ***Different types of phages targeted by CRISPR-Cas subtypes.*** *Different CRISPR-Cas subtypes have been documented targeting Jumbo (i.e. type III systems), ssDNAphages and dsDNAphages. In contrast, RNA phage targeting by naturally occurring CRISPR-Cas systems does not occur.*

1.7 CRISPR Interference's role in the acquisition of spacers in a "primed" manner.

It has been reported that in some CRISPR-Cas systems, the process of CRISPR spacer acquisition is directly coupled to interference[164](Figure 1.9). Acquisition is stimulated by the interference activity of a pre-existing spacer matching the invading phage, in a process termed '*priming*'[67,164]. Priming enhances phage defence by adding more spacers which target regions in proximity to pre-existing spacer target sites to prevent phage escape[67,164]. Primed adaptation often still functions when the matching spacer contains several mismatches or truncations, in contrast to interference which requires

close to perfect complementarity [68,130]. In some systems, the acquisition of "slipped spacers" – spacers containing one or more PAM nucleotides, deliberately stimulates this effect, creating two classes of spacers, one intended mainly for priming, the other specifically for interference [165]. There is evidence that in certain systems priming may be enhanced when phages or other MGEs evolve single point mutations to the target sites of spacers in the PAM regions adjacent the target sites[131,165-168]. In some cases, a single region or gene in a phage is targeted by many spacers from the same array. This phenomenon is called "hyper-targeting" [169] and strengthens the CRISPR-mediated response by specialising the host-encoded spacer repertoire to provide a potent form of acquired immunity over time in response to repeated infections via the same phage. This contrasts with '*naive*' acquisition where spacers are acquired at random from a phage.

In addition to strengthening an immune response against a specific phage, primed-spacer acquisition also provides a pathway to enable CRISPR-Cas systems to discriminate self-versus non-self when acquiring spacers[131,165,170]. Because primed acquisition solely acquires new spacers from an existing target site, these will be solely phage or MGE nucleic acids which minimises the risk of self-acquisition of spacers from the host genome[127,170,171]. Given the fitness advantages priming extends in self-vs non-self CRISPR spacer acquisition, as well as increasing the difficulty of target phages developing escape mutations, one important unanswered question is whether priming only occurs in rare CRISPR subtypes or is a more universal feature of CRISPR-immunity.

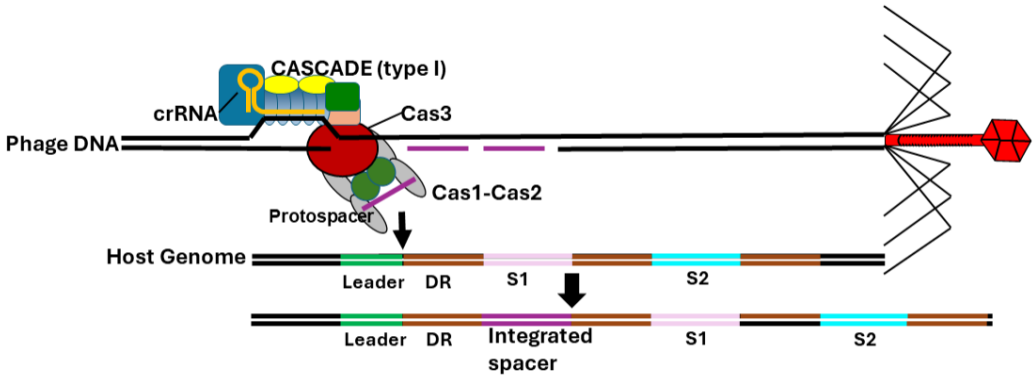


Figure 1.9: ***Acquisition and Interference are coupled in certain CRISPR-Cas subtypes.*** *This phenomenon has been most studied in type I systems*[127,130,131,172-175]*. Fragments of ssDNA produced by Cas3 mediated degradation of the non-target DNA strand are concurrently uptaken by the Cas1-Cas2 complex and integrated into CRISPR arrays. This process requires a pre-existing spacer for RNA guided interference and is termed primed spacer acquisition*[67,174]*. Priming results in a CRISPR-array possessing multiple spacers to the same target region which prevents the development of escape mutations by an invading phage*[67,164,174].

1.8 Spacer acquisition biases are generated through several independent mechanisms,

Primed spacer acquisition in Type I systems

The process by which primed spacer acquisition occurs has been the most well characterised experimentally in type I-E and I-F CRISPR-Cas systems[67,127,131,164,171,174]. In type I-E systems, the formation of an RNA-DNA heteroduplex on target site binding by Cascade triggers the recruitment of the Cas3 helicase/nuclease protein[71]. Cas3 processively degrades a single strand of DNA up to around 3-10kb from the target site[164,176,177]. Fragment production of ssDNA is monodirectional and proceeds in the 3' to 5' direction in the non-target strand [67,164]. These fragments are converted to dsDNA fragments. The Cas1 and Cas2 acquisition proteins then integrate the fragments of DNA into the host-cell CRISPR-arrays. These processes are posted to occur concurrently as part of a single complex consisting of the CASCADE-Cas3-Cas1-Cas2-IHF proteins called the primed acquisition complex (PAC) [133,172]. Acquisition *in vivo* proceeds at the leader-repeat junction of the CRISPR-array and requires integration host factor which alters the topology of the target site and enhances site-specific binding by Cas1-Cas2 to the junction [74]. The absence of IHF has been shown to result in reduced activity, increased cell cytotoxicity and integration at any direct repeat in the CRISPR-array [74]. This mechanism of acquisition is also thought to be the same in type I-B and type I-C systems[178].

Despite being closely related subtypes, there are several significant differences between how spacer acquisition functions in Type I-E systems versus type I-F, I-B or I-C.

For instance, priming in type I-F systems appears to be bidirectional [127]. In these systems, the Cas2 protein used for acquisition is directly fused to the Cas3 which facilitates strand degradation [78]. The direction of strand bias is also the opposite of that observed in type I-C, I-B and I-E systems, with new spacers being acquired 5' of the priming protospacer (PPS) on both the target and non-target strand [131]. This implies a slightly different mechanism underpinning the means by which type I-F systems achieve primed spacer adaptation.

Spacer acquisition biases as a byproduct of DSB induction

In addition to primed adaptation conferred by the interference complex of type I systems, a second means exists by which protospacers can prime the acquisition of additional spacers in the immediate vicinity of target sites. The RecBCD complex, which is critical in the initial step of homology directed repair of double strand breaks, produces ssDNA fragments during resection, which serve as protospacer substrates for Cas1-Cas2 acquisitions proteins [179]. The resection continues until recombination hotspot DNA-motif sequences known as Chi-sites are reached, at which point the resection reaction terminates[169,179]. The acquisition complex has been shown interacting with the RecBCD repair complex, to more efficiently receive and integrate them into CRISPR-arrays concurrent to the resection reaction[180]. Type II and V CRISPR systems both produce DSBs as their primary means of degrading invasive phages and other MGEs. Hence, DNA repair of these breaks on a phage produces more ssDNA fragments which are integrated into CRISPR-arrays concurrent to resection, and confer a stronger, and more escape-mutation resistant immune response if the phage reinvades the same host-cell or its descendants[165]. This phenomenon has been demonstrated in type II-A systems. However, to date these biases remain weak or unobserved in type V CRISPR-Cas subtypes [181].

Other sources of acquisition biases

In addition to primed spacer acquisition, there are several other potential reasons an apparent spacer acquisition bias may be observed, whether genuine or artefactual. A consequence of pre-spacer fragment production from RecBCD resection is that Chi-sites reduce the acquisition of spacers by preventing further resection past these

sequences [169,179,181]. This means that genome sequences rich in Chi-sites may produce less ssDNA fragments from resection [169,179]. Host genomes, which are rich in Chi-sites produce less ssDNA substrates for acquisition and are thus less vulnerable to CRISPR-induced autoimmune self-targeting compared to phage or plasmidic genomes, which contain fewer of these sites [169,179].

It has been observed that some CRISPR-Cas systems, such as type III, for which there are no previous reports of priming, acquisition biases still exist [182] [114]. This bias is localised to the start sites of many genes as well as a distinct preference for regions of the genome which encode RNAs which significant secondary structure such as replication origins and tRNAs [114,182]. In some subtypes Reverse-transcriptase (RT)-Cas1 fusion proteins have been reported to be capable of direct integration of RNA as spacers in CRISPR arrays [183]. However, many type III systems lack these RTs, and it has been shown that adaptation is not proportional or dependent in the volume of RNA transcribed [114,182,184]. Furthermore, for many CRISPR-Cas subtypes, the presence or absence of priming or other spacer acquisition biases has not yet been reported.

1.9 Thesis aims:

This review introduces key aspects of anti-phage defence systems, phage and CRISPR-Cas classification schemes, host-phage interactions and spacer acquisition biases. Surveys of anti-phage defence systems has unveiled sprawling superfamilies of diverse origin with many distinct modules which have evolved independently. This includes a theorised set of developmental pathways accounting for the emergence of the key functional domains integral to RNA-guide interference, which define each CRISPR system type. At the taxa level, CRISPR-Cas systems combine with other anti-phage defences as a single integrated immune system, capable of both innate and adaptive responses, to facilitate enhanced survival of host-cell populations against infection. There is a very extensive stratification of CRISPR-Cas at the type and subtype level, which is shaped evolutionarily by red-queen driven co-variation between these systems and their target phages. This is also reflected by the diverse mechanisms of CRISPR spacer acquisition which generate different acquisition distributions between different system types and subtypes.

Despite a plethora of different CRISPR-Cas systems and their associated complement of co-encoded genes being extensively characterised both computationally and experimentally, I hypothesise the full extent of novel variation within these subtypes and their complex relation with the phages and other defences remains largely unknown and requires a multigene (entire contig) rather than a single gene approach. The goal of this thesis is to survey the extent of interactions between CRISPR proteins, accessory genes, other CRISPR defences as well as differences in how phages are targeted by spacers.

In this study, I employed three separate approaches to further survey the diversity of CRISPR-Cas subtypes and their associated repertoire of co-encoded genes. Firstly, I emulated previous approaches which used guilt-by-association to discover new CRISPR-Cas systems to identify any additional systems or associated genes which had remained unreported. I then employed network-based methods to model the intra-subtype diversity of Type VI CRISPR-Cas subtypes and identified the complement of novel defence genes associated with local clusters. I then performed spacer mapping

and characterised the diversity set of mapped Mobile Genetic elements (MGEs) using the same network-based approach. Finally, the ability of different CRISPR-Cas subtypes to display a bias in spacer acquisition, through effects such as DNA repair and direct primed spacer acquisition was surveyed by analysing biases in the distribution of mapped spacer pairs in cases where 2 or more spacers mapped to the same MGE sequence. Each of these thesis approaches was delineated into a separate thesis research chapter:

Chapter 3: Using a computational pipeline to survey gene diversity associated with CRISPR-Cas systems

In this chapter I constructed a computational pipeline of my own to mine known and unknown CRISPR-associated genes from a large block of assembled sequencing data using guilt by association. Using the output set of CRISPR-associated genes identified by this process, I annotated and surveyed the relative makeup of co-encoded genes in proximity to CRISPR-arrays including many genes whose function remained experimentally uncharacterised. I then identified conserved genes among common CRISPR-Cas subtypes to identify any novel co-associations. The set of CRISPR-Cas subtype encoding contigs extracted by data mining were retained and further used in subsequent chapters 4 and 5 to probe intra-subtype diversity or provide a source of spacers for spacer mapping.

Chapter 4: Network based characterization of intra-subtype diversity of host-MGE interactions at the gene cluster level

After surveying the diversity of CRISPR-associated systems and co-encoded genes. I next investigated the intra-subtype diversity of type VI systems using a network-based approach. Concurrent to these efforts, spacer mapping was performed to identify the targets of each local cluster of each subtype investigated. The gene composition, shape and host-phage nature of subtype encoding contigs was then interrogated to identify novel genes and defence islands structures among local clusters. This procedure was repeated with the mapped target contigs derived from spacers of Type VI-A, VI-B and VI-D systems. Finally, interactions between local clusters of host-encoded subtypes and local clusters of their corresponding phage or plasmidic targets were illustrated to identify the tropism of different MGEs as well as the diversity of MGEs targeted by

CRISPR immunity at the local cluster level. The spacer mappings performed in this chapter, alongside their corresponding target sequences, were taken and used to estimate spacer acquisition biases in chapter 5.

Chapter 5: Characterisation of primed spacer acquisition biases across CRISPR-Cas system subtypes

In my final results chapter, I sought to identify whether different CRISPR-Cas subtypes exhibited spacer acquisition biases. I used a previously developed technique called “spacer distribution analysis”, to estimate acquisition biases and analysed an expanded range of CRISPR-Cas subtypes to see whether any exhibited acquisition biases as a consequence of a priming protospacer (PPS). I then attempted to increase the sensitivity of the technique by searching for partial spacer matches instead of mostly complete ones.

Ultimately this project has further elucidated the complex biology of CRISPR anti-phage defences in the taxa rather than single gene context.

Chapter 2: Methods

2.1: Overview

A single vertically integrated workflow was utilised to generate the data presented in all three research chapters of this thesis. There are three general layers to the overall process (Figure 2.1). The first layer (Figure 2.1A) employs a data mining approach to extract CRISPR-associated ORFs co-encoded with CRISPR-arrays, which are detected by pattern recognition tools as a single-pass operation as the data is fed into the pipeline. The second part of this layer screens candidate CRISPR-associated proteins using three metrics: distance, co-occurrence and abundance. These reduce the false positive detection rate of ORFs which are co-encoded with CRISPR-arrays but not actually genetically linked with the arrays. The second layer of the process first employs multiple CRISPR-array prediction tools to validate and expand the maximum detected size of the CRISPR-arrays, then utilises BLAST to perform spacer mapping followed by spacer distribution analysis on the mapped sequences (figure 2.1B), which is important in describing the preliminary acquisition and interference requirements for the system being annotated. The final layer employs a variety of programs to annotate individual genomes and contigs according to features such as shape, taxa and sample of origin (Figure 2.C). Separately, predicted ORFs are annotated by homology to reference HMM based protein databases.

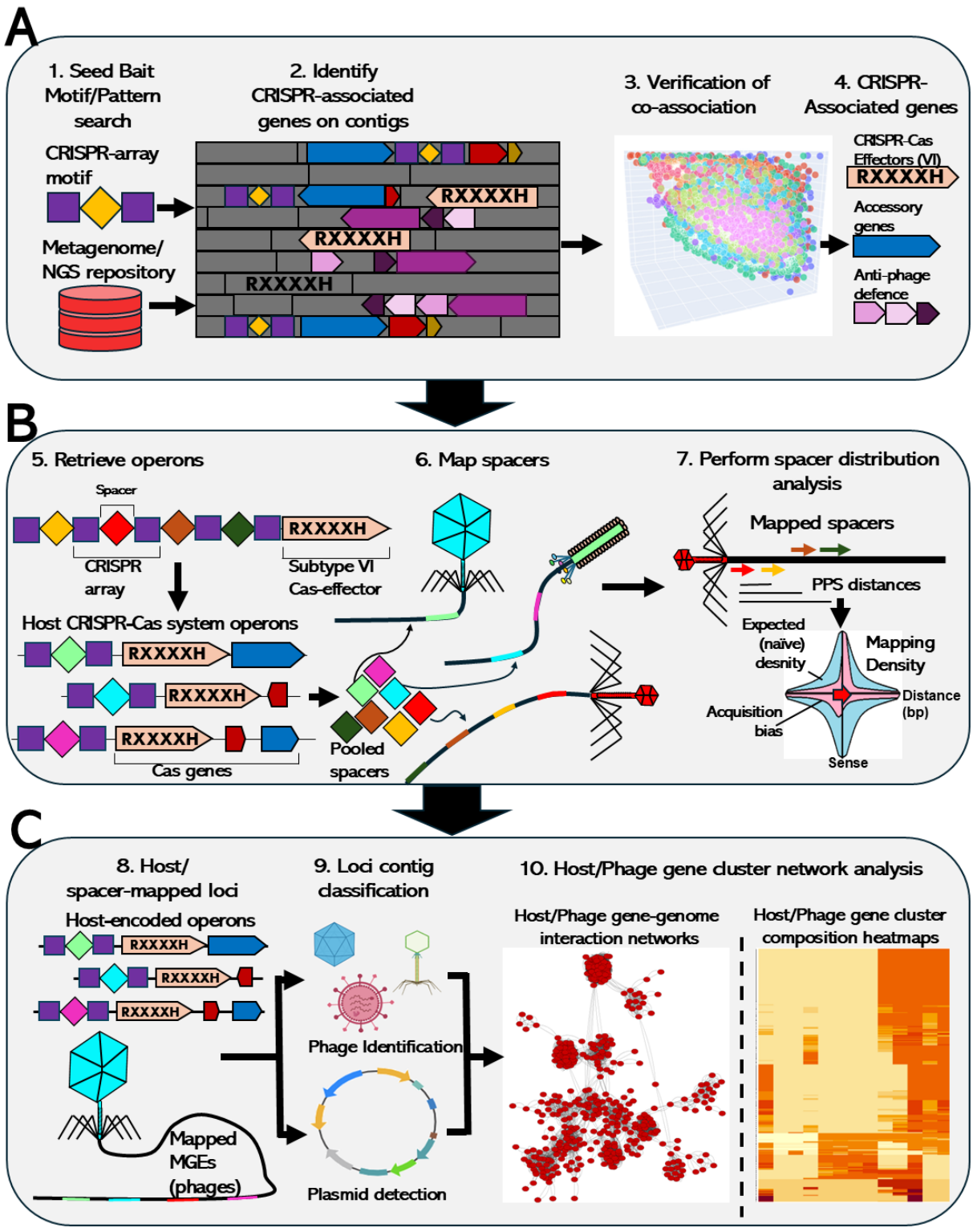

A
1. Seed Bait Motif/Pattern search
CRISPR-array motif
Metagenome/ NGS repository
2. Identify CRISPR-associated genes on contigs
RXXXXH
RXXXXH
RXXXXH
3. Verification of co-association
4. CRISPR-Associated genes
CRISPR-Cas Effectors (VI)
RXXXXH
Accessory genes
Anti-phage defence
B
5. Retrieve operons
Spacer
CRISPR array
RXXXXH
Subtype VI Cas-effector
Host CRISPR-Cas system operons
RXXXXH
RXXXXH
RXXXXH
Cas genes
Pooled spacers
6. Map spacers
7. Perform spacer distribution analysis
Mapped spacers
PPS distances
Expected (naïve) desnity
Mapping Density
Distance (bp)
Acquisition bias
Sense
C
8. Host/ spacer-mapped loci
Host-encoded operons
RXXXXH
RXXXXH
RXXXXH
Mapped MGEs (phages)
9. Loci contig classification
Phage Identification
Plasmid detection
10. Host/Phage gene cluster network analysis
Host/Phage gene-genome interaction networks
Host/Phage gene cluster composition heatmaps

Figure 2.1: ***Overview of computational workflow used for data generation throughout chapters 3-5.*** *This workflow was split into 3 general layers. (A) In the first layer, metagenomic data was mined for CRISPR-arrays and DNA sequences within ±20kb were interrogated for co-associated genes using Guilt-by-Association (GBA). (B) Contigs encoding signature effector genes for certain CRISPR-Cas subtypes were then retrieved and used to perform spacer mapping, followed by spacer distribution analysis to detect acquisition biases. (C) Mapped sequence contigs, along with their corresponding spacer-encoding CRISPR-loci were then annotated and used to produce gene-genome interaction networks, both within and between host and mapped sequence (usually phage) clusters as well as heatmap-based comparisons of the gene compositions between clusters.*

2.2: Extraction of CRISPR-associated proteins from assembled metagenome sequence data

2.2.1: Sources of assembled metagenome sequence data

Assembled sequencing data was downloaded from the JGI repositories[185,186]. To download the data from NCBI, a search was performed on September 6th 2019 against the Genbank and JGI using the entrez query "bacteria OR archaea OR metagenome" . Files were then downloaded in FASTA format. In total ~10 TB of assembled metagenome sequence was downloaded from the Joint Genomes Institute (JGI) and from the National Centre for Biotechnology Information (NCBI) on 20th February 2020. The individual assemblies were then reorganized into 38,810 'blocks' of FASTA files approximately 250Mb in size. Each genome sequence header was labelled by the corresponding block filename to enable fast retrieval in downstream steps of the pipeline process. These files were then indexed using "makeblastdb" command and "samtools faidx". This enabled the usage of these files as BLAST databases and the use of index-based retrieval of individual genome sequences within each block file.

2.2.2: Extraction of putative CRISPR-associated proteins

A computational pipeline was constructed and used to predict CRISPR-Associated proteins via the previously established guilt-by-associated approach[91]. The CRISPR-array was used as the seed for this purpose, due to being the most conserved feature of CRISPR-Cas systems. Utilising multiple seeds complicates the data mining process, and results in a higher false positive rate if the seeds being utilised are not conserved features of CRISPR-Cas systems and/or have roles in operons beyond CRISPR-immunity. PILER-CR[93] was utilised as the primary array prediction tool to detect CRISPR-arrays, using the default parameters . A 'window' of DNA up to 20kb upstream and downstream of predicted arrays (or until the end of the contig) was extracted and parsed into prodigal (with -p meta enabled) for corresponding ORF prediction [101]. In some cases, the windows subtended from two CRISPR-arrays overlap. To prevent deduplication of subsequent predicted ORFs, these windows were concatenated together and redundant overlaps deduplicated so that only one copy for each window in each assembled genome region existed. The distance between each ORF and the start position of the corresponding CRISPR-array was labelled within the FASTA headers of each ORF predicted from the windows. Predicted ORFs were then clustered using mmseqs2 [187] using the following command:

mmseqs cluster -s 7.5 –min-seq-id 0.3 -c 0.8 -e 0.00001 –cluster-mode 1 <sequenceDB> <clusterDB> <tmp>

Although the exact settings used during this step were non-critical towards the subsequent downstream steps of the computational workflow, the choice of sensitivity (-s), minimum sequence (similarity), coverage (-c), e-value (-e) and cluster mode (cluster-mode) values was driven by the need to coalesce proteins into the largest possible protein families, without also counting proteins which possessed detectable domain homology but no shared evolutionary lineage. This entailed sensitive connected-component clustering between sequences at minimum of 30% similar as well as possessing a relatively high e-value (10^{-5}) between sequences. A moderately high coverage threshold of 80% was chosen due the need to vet small fragments with low similarity from putative families. Relaxing these thresholds further significantly

increased the incidence of two putative protein families being merged. Conversely, tightening resulted in many putative protein families being split into multiple clusters.

After clustering, the average distance from the CRISPR-array was calculated within each cluster. Clusters were then filtered by two criteria. Firstly, whether ORFs overlapped significantly with CRISPR arrays, as these ORFs sometimes corresponded to promotors present within the direct repeats of the arrays which results in false-positive association between ORF and CRISPR array. Secondly, clusters lacking at least one ortholog greater than 300aa in size were removed, as most known CRISPR-associated effector and accessory proteins have usually been at least this length, and the higher false positive rate for smaller predicted proteins make their authenticity harder to verify. Representative proteins for each cluster were then generated from the cluster files using the "createsubdb" and "convert2fasta" functions within mmseqs, and pooled into batches of 50-80 queries for a tBLASTn search against the entire 10TB data block. A separate search was conducted against the file containing 20kb extracted upstream and downstream of CRISPR arrays. These searches were conducted under the same parameters, except for the e-value cutoff which was set at 10^{-5} for searches of DNA within 20kb of a CRISPR-array, and 10^{-8} for the 10TB data block (split into 38,810 units). This was to compensate for the difference in the background detection rates of weakly similar sequences because of the difference in database size, which was in the ratio of approximately 1:1000 for DNA within 20kb of a CRISPR-array compared to the 10TB data block. The command was as follows:

tblastn -query <representative_seq_file> -out <Blast_matches_amalgamated> -outfmt 10 -evalue 10^{-5}/10^{-8} -max_target_seqs 10000000 -max_hsps 1

Hits to individual sequences were then identified by splitting the result files (in .csv format) into individual files for each query. For each candidate protein with an average encoding distance <10kb from the CRISPR array, a CRISPRicity (co-occurrence) score[90,91], was then computed for each representative sequence from each cluster (representing a putative protein family) by dividing the number of hits to CRISPR-window sequences, (distance ≤ 20kb to CRISPR-arrays) with the total number of hits in the original 10TB data block [91]. There was some minor distortion from setting the largest number of returning highest scoring pairs to 1 (-max_hsps) given that multiple BLAST

hits to a single sequence were possible. This may have reduced the CRISPRicity scores for some sequences. However, these scores still established co-association between predicted protein sequences with nearby CRISPR-arrays. Each representative protein (representing each putative protein family detected) was then tabulated by designated name, total number of tBLASTn hits ≤ 20kb from CRISPR-arrays (abundance), average distance from the CRISPR-arrays (distance) and co-occurrence (co-occurrence). The correlation between each of these variables was then visualized in 3 separate scatterplots using ggplot2. To generate the density contours, the geom_density2d function was used with bins=25.

2.2.3 Calculating F-statistics for each dimension used for screening of CRISPR-associated proteins

To calculate and compare the variance between CRISPRicity (co-occurrence), distance and abundance, each dimension was first normalised by the highest recorded value in the dataset [co-occurrence = 1, distance = 10,000 abundance=1,323,936 (highest recorded abundance)]. For CRISPRicity and abundance, the negative base 10 logarithm was then also taken. The variances of each normalized dimension were then calculated and the ratio used to calculate a F-statistic (Chapter 3: Table S3.2).

2.2.3: Annotation of representative proteins using HMM and DEFLOC searches

Representative CRISPR-associated proteins were utilized as queries to HMM based searches using HHblits against the Pfam database and Hmmscan against the DEFLOC database. The hit with the lowest e-value was used for annotation. A relatively permissive e-value was chosen for both HHblits and Hmmscan searches to enable the detection of remote homology to putative genes. Phage defence genes stratify at relatively rapid evolutionary rates[92,188,189], so a reasonably high threshold was selected to

ensure that any genes with homology to antiphage defence domains were detected. The command used was as follows:

Pfam:

hhblits -i <input_sequence> -o <output_sequence> -e 0.001 -maxseq 10 -d <path to pfamA_35.0 database files> -v 0

DEFLOC:

hmmscan -o <output_sequence> <DEFLOC_database_models.HMM> <input_sequence>

Due to the lack of a standardised ontology describing the domains of the putative CRISPR-associated proteins, HMM/BLAST annotations were manually curated and each putative CRISPR-associated protein representative was binned into one of nine categories.

1. **"CRISPR-Cas effector"** – Homology to known CRISPR-Cas effector protein
2. **"Accessory"** – Homology to known protein which is not part of the conserved canonical pathway for CRISPR immunity detailed in Section 1.1, but which is conserved within certain CRISPR-Cas subtypes.
3. **"Defence-island"** – Protein with no known role in CRISPR-Cas immunity, but which does have homology to genes either known or suspected (by domain) to play an important role in anti-phage defence.
4. **"Hypothetical protein"** – Protein with no characterised function/domain but which has a match in the Pfam/BLAST database.
5. **"Known protein with no CRISPR function"** – Protein with no known function in CRISPR immunity but which contains a domain that plays a known functional role in other contexts.
6. **"Phage/Transposon"**- Protein with homology to known phage/transposon proteins
7. **"Acquisition + processing"** – Protein with homology to known proteins involved in CRISPR spacer acquisition or pre-crRNA processing

8. **“Known RNA with no CRISPR function”** – Homology to ORF translation of an RNA molecule with no known role in CRISPR immunity, but with a conserved function in a different context.
9. **“No similarity”** – No detectable homology to any profile/sequence in the database

Each representative protein was then tabulated and represented as a vector of five attributes [Identifier, abundance, distance, co-occurrence, category]. This was then plotted in three dimensions using the **scatter_3d** method from the plotly package.

2.2.4: Observation and ranking of CRISPR-Cas/phage gene cluster composition and abundance

ORFs from each CRISPR-Cas subtype were first annotated by the procedure detailed in Chapter 2.3.5. The pfam database, and an in-house database (DEFLOC) comprised of a collection of HMM profiles corresponding to known CRISPR-associated proteins aggregated and derived from CRISPRCasfinder [190], Defence Finder [86] and PADLOC [191]. The HMM profiles were then used as a *de facto* homology-based ontology to match, annotate and quantify the number and identity of each predicted protein. Sometimes multiple matches to different Pfam/DEFLOC families were undercounted if the proteins were from the single common ancestor. In these cases where the profiles were clearly related, protein domains and motifs were manually merged to include both proteins in the same profile.

Conversely, to assess whether divergent orthologs were co-encoded on the same subtype, arising from two unrelated proteins sharing an overlapping domain, the ancestral lineage of the representative proteins in each cluster was determined by phylogenetic tree construction using IQtree2 [192]. The maximum likelihood trees were performed with modelfinder and 1,000 bootstraps under the default parameters, on a multiple sequence alignment of proteins in each cluster as input, generated by Clustal Omega (v1.2.4) [193,194]. Representatives were excluded if annotations from orthologs were determined to be from two completely separate clades of orthologs. Common

annotations shared between representatives of each putative co-associated protein were then tallied and tabulated by subtype. A conservation score was derived from these tallies, as the fraction of instances a protein was detected in a subtype divided by the total number of contigs in a given subtype. A protein which occurs once in each contig for a given subtype is defined to have a conservation score of 1. It was possible for a protein to have a score higher than one if the protein has more than one copy present, on average in the contig set. The tabulated conservation scores were then used as an input matrix. This matrix was used to generate a heatmap using the pheatmap function from the pheatmap software package in R.

2.3: Comprehensive annotation, differentiation and phylogenetic analysis of host-phage interactions at the gene cluster level

2.3.1: Consensus CRISPR-array validation, orientation prediction and spacer numbering

To perform spacer mapping for each CRISPR-Cas subtype, which was used in Chapter 4 to further perform host-MGE interaction analysis, and in Chapter 5 to study primed spacer acquisition, a set of CRISPR-array encoding contigs were required for each subtype. These were retrieved using a tblastn search (e-value=10^{-7}) against the set of CRISPR-array encoding DNA sequence contigs ± 20kb from the array using a strongly conserved gene for each subtype specified in Table S4.1.

The spacer mapping and annotation of individual genomes was conducted in two distinct steps. The first steps consisted of using three CRISPR array prediction tools to validate the CRISPR-encoded arrays on each contig from each subtype:

1. PILER-CR[93]

2. CRISPR-CRT[94] as part of the CRISPRleader package [195] (modified to process and output multiple CRISPR-array containing files, instead of a numbered list of arrays)

3. CRISPRDETECT[95]

To maximise the size of the CRISPR arrays, the union comprising the largest possible

array given non-redundant overlapping spacers was taken. When deciding which spacer to keep a hierarchy of CRISPRdetect > CRT > PILERCR predictions were retained, as CRISPRdetect is more precise than CRT, and although PILERCR is slightly more precise than CRT, only the CRISPRleader program, of which CRT is a subcomponent, predicts orientation, which was necessary for downstream calculations. Hence, to ensure an orientation for all arrays CRISPRdetect and CRISPRleader-CRT were preferred above PILERCR, which was mainly used for array expansion and cross validation.

After array expansion, the consensus orientation was determined from the CRISPRleader and CRISPRdirection/CRISPRstrand subprograms of CRISPRdetect[196,197]. respectively. In cases where orientation predictions were contradictory, predictions by CRISPRdetect were preferred over CRISPRleader, due to the higher accuracy of CRISPRdetect predictions. However, CRISPRdetect was not able to give a prediction for every array, whereas CRISPRleader always returns a prediction, hence CRISPRdetect predictions were considered superior to CRISPRleader predictions, which were used when CRISPRdetect was unavailable.

After array expansion and orientation prediction, spacers within the expanded arrays were numbered in descending order from the spacer furthest from the promoter. The first spacer in the array was assumed to be the first spacer integrated in the array and was thus assigned as the priming protospacer (PPS) which was further used to spacer acquisition biases investigated in Chapter 5.

2.3.2: Spacer mapping and deduplication

Spacers identified from the consensus arrays validated in the previous section were next mapped against the 10TB data block used to mine for and survey CRISPR-linked genes. BLASTn was used to identify matches between spacer queries and assembled sequences in the 10TB data block (split into 38,810 approximately equal sized smaller files used for searching). Relatively high identity and coverage (-qcov_hsp_perc) thresholds for each highest scoring pair were required to prevent returning a massive number of false positives, stemming from the fact that the average CRISPR-spacer size was 20-40bp, and a minimum spacer size of >12-14bp is required to return mostly unique matches. The search parameters used for spacer mapping are given below:

blastn -query <spacers.fasta> -db <assembled sequence data block> -outfmt 10 -out <output matches> -perc_identity 0.95 -max_target_seqs 100000000 -max_hsps 1 -qcov_hsp_perc 90

To enable the removal of spacer hits to their own CRISPR-arrays, both to the contig from which the spacer was derived, as well as other metagenome data which contains the same, or homologous arrays, 10bp handles consisting of the upstream and downstream direct-repeat sequences adjacent each spacer were constructed. These were then queried against the same 10TB metagenome data block. The identity threshold was reduced slightly to further enhance the detection of direct-repeats. This was performed with the following search parameters:

blastn -query <spacers.fasta> -db <assembled sequence data block> -outfmt 10 -out <output matches> -perc_identity 0.90 -max_target_seqs 100000000 -max_hsps 1 -qcov_hsp_perc 90

The rationale for this was to compare matches for the direct-repeat-spacer constructs with the original spacers (Figure 2.2A). The DR-spacer constructs were expected to map to parts of CRISPR arrays with perfect complementarity (~100% similarity) but map to other positions lacking the direct repeats with a much lower proportion of similarity than the original spacers. This approach utilising DR-spacer fusions was analogous to the approach employed by CRISPRtarget[198]. It was thus possible to use this method to filter out spacers targeting matches to parts of the direct repeats in their own CRISPR arrays, leaving only spacer matches exclusively to non-CRISPR loci (Figure 2.2B). These were kept for subsequent analysis, while the rest of the matches were removed from the dataset.

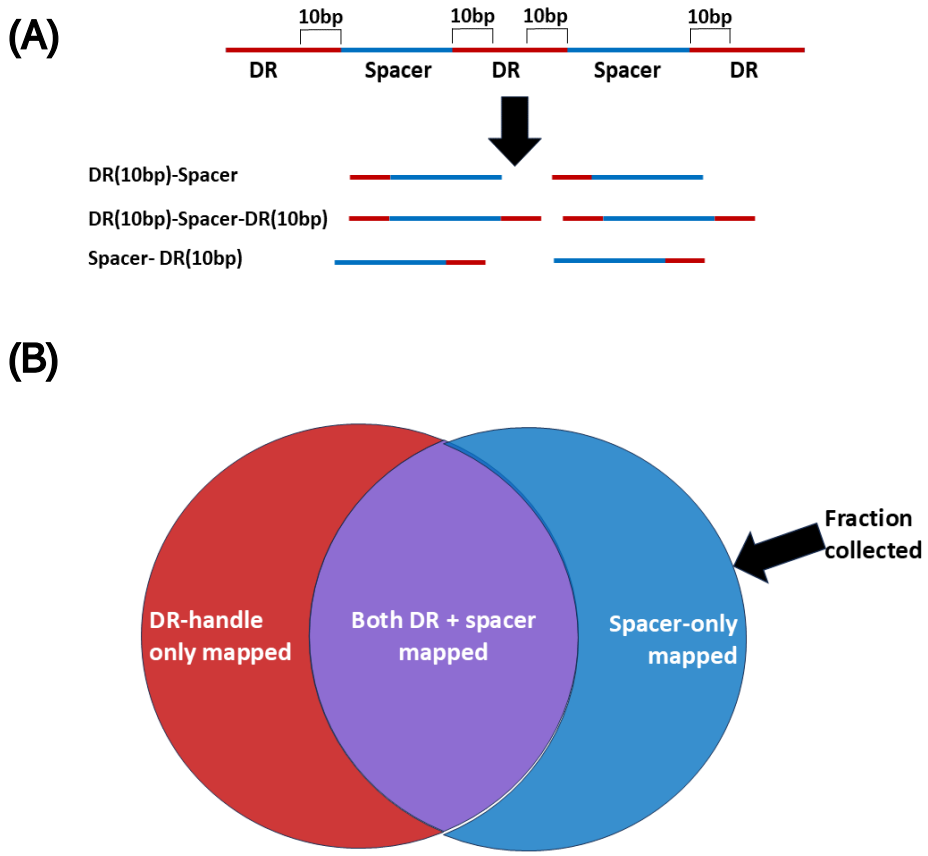


Figure 2.2: ***Concatenation of Direct Repeat (DR) distal and proximal ends to spacer sequences to filter self-targeting spacer matches from matches to other loci (i.e. phage genomes).*** *(A) 10bp handles composed of the upstream and downstream Direct Repeats encoded adjacent to each spacer in the CRISPR-array were barcoded to the spacers and used to perform a BLAST search under slightly more relaxed parameters. (B) Only spacer matches which did not have an equivalent match by the DR handle-spacer sequences were kept for subsequent use as these did not match any direct repeat elements which were the main distinguishing signature of the CRISPR arrays compared to other spacer targets.*

After mapping CRISPR-array derived spacer sequences to targets, one issue that arose was the large number of hits to contigs which were complete functional duplicates within the dataset. To reduce this bias, each row containing each spacer match was indexed as a set only if the start and end coordinates of the matching spacer did not overlap. Genome Identifiers from spacer matches to contigs identified by BLASTn were then used as keys to retrieve the corresponding sequence in the 10TB block using

seqkit[199]. In a significant fraction of cases, sequences were mapped to prophages or prokaryotic genomes, or contigs considered too large to perform downstream protein annotation and network-based analysis. As a result, a maximum size of 10kb of the DNA upstream and downstream of the mapped site was retrieved. These sequences were then concatenated together as a single FASTA file for subsequent analysis.

2.3.3 Genome type and shape annotation of host and spacer mapped sequences

The subtype-specific set of CRISPR array encoding DNA sequences up to 20kb from a CRISPR-array along with their spacer-mapped sequences were annotated by Virsorter and PLASMe

Retrieved host contigs and non-redundant phage contigs were annotated to determine their genome shape (circular/linear), type (cell/phage) and species. Genome shape was determined using PLASMe.[200] The coverage (-c),identity (-i), and transformer probability scores were relaxed somewhat to improve the sensitivity of plasmid detection. This likely also increased the false-positive rate. However, given that Plasmid-detection was performed on all sequences rather than just single sequences in each cluster, this improved the confidence of plasmid detection when the results were converted to a heatmap representation. This was performed via the following command:

python3 <path to PLASMe>/ PLASMe.py <input_contigs> <output_file> -c 0.6 -i 0.6 -p 0.5 -t <num_threads> --temp <temp_dir>

Virsorter2 was also utilized, using the default parameters to predict genome shape, but also scored sequences based on propensity to be ssDNAphage/dsDNAphage/RNAphage/Jumbo phage /Virophage(Lavidaviridae). The default cut-off was chosen to maximise the sensitivity of viral prediction. The prediction of these additional phage categories, in addition to ssDNAphage and dsDNAphage (the default) was specified by the following command:

virsorter run -l <input_contigs> -w <output_folder> --include-groups dsDNAphage,NCLDV,RNA,ssDNA,lavidaviridae -j <num_threads> --tmpdir <tmp_dir> --rm-tmpdir

2.3.4 Gene annotation of host and spacer mapped sequences

A short mini-pipeline was developed for predicting and annotating individual contigs gene composition. Firstly, ORFs were predicted from both host contigs and mapped sequences for each CRISPR-Cas subtype investigated. These proteins were then subject to three separate HMM based searches against three separate reference databases using either HMMscan or HHblits[201] (table 1). These databases were PfamA_35.0 (HHBLITS), DEFLOC (HMMSCAN) and PDB70 (HHBLITS). Relatively high e-value and coverage cut-offs were set to enable domain assignment for proteins which contained only distant homology to protein sequences in the databases. This did increase the false positive rate of some of the assignments. However, this increased rate was mitigated via the comparison of annotations derived from Pfam, and those from derived from DEFLOC sequences, which were a subset of HMM profiles specific to known antiphage defence domains. In concert they provided high sensitivity and high specificity in being able to assign an annotation based on homology to almost any ORF yet also containing a description specific enough to infer potential functions. In each case, the top search result from the homology search was the only annotation retained for each protein as this was considered the most likely homology according to the overall probability-score.

REFERENCE DATABASE	**PFAMA_35.0**	**DEFLOC**	**PDB70**
SEARCH TOOL	HHBLITS	HMMscan	HHBLITS
COVERAGE CUTOFF	0	Default	0
MAXSEQ	15000	default	15000
E-VALUE CUTOFF	0.001	default	0.001

Table 2.1: Key (non-default) parameters utilised in HMM based searches used to annotate predicted proteins from individual host and spacer mapped contigs.

2.3.5 vConTACT2 Network generation for host contigs and spacer mapped targets

To show the parsimonious differentiation of different CRISPR-Cas systems and their mapped spacer targets at the operon level, the set of sequences encoded a CRISPR array in addition to the DNA sequence ± 20kb upstream and downstream for each subtype, as well as the corresponding set of spacer-mapped mapped sequences ± up to 10kb from the target site, was parsed as input to vConTACT2 as two independent processes. As a precursor to running vConTACT2, the contigs utilised as the input file first underwent ORF prediction by prodigal using the following command:

prodigal -i <CONTIG_INPUT_FILE.FASTA> -o <CONTIG_INPUT_FILE.FASTA_OUTPUT.TXT> -a <CONTIG_INPUT_FILE.FASTA_aa_raw.fasta> -p meta

The output files from prodigal were then parsed to vConTACT2[152]. DIAMOND was used as the main search tool due to it's relatively low computational cost, which was desired given the poor time-complexity of the all-by-all sequence-search step required for network generation. A relatively high e-value (0.001) between protein clusters was set to maximise the formation of edges between distantly related nodes and minimize the number of separate networks generated. The main parameters for running vConTACT2 are given below:

python3 vcontact2 --raw-proteins <PROTEIN.FASTA> --db None --rel-mode Diamond --pcs-mode MCL --pc-inflation 1.2 --pc-evalue 0.001 --proteins-fp <GENE2GENOME FILE> --output-dir <OUTPUT_FILE> --c1-bin /g/data/va71/vConTACT2/MAVERICLab-vcontact2-c0413a6c92e8/bin/cluster_one-1.0.jar -t 52 -e "cytoscape"

The output of vConTACT consisted of a folder for each CRISPR subtype investigated, as well as a corresponding folder for each set of spacer-mapped sequences. This folder contained the network file (c1.ntw) as well as a gene-by-gene-overview.csv file containing the assignments of each input sequences as a cluster, overlapping contig, singleton or outlier. There was also a file (c1.clusters) containing a list of the viral clusters produced by vConTACT2, which was later used for bipartite host-MGE network generation. Operons for which no detectable similarity could be found between contigs were assigned as singletons or outliers.

2.3.6 vConTACT2 Network visualisation

To visualise the vCONTACT2 generated networks the first step was to convert the c1.ntw output file produced by vCONTACT2 into a graphML format using cytoscape. Each of the networks was partitioned into clusters using the leiden cluster partitioning method[202]. The resolution parameter (0.01) was set in order to maximise community size while still retaining distinct modular groups (table S4.2). This value was kept constant to enable a direct comparison between the host and mapped sequence communities. However, as the threshold for what constitutes a community within the network is somewhat arbitrary, a different resolution could be used[202]. Each cluster was assigned a randomly generated colour. A file containing the sequence identifier, the partitioned cluster number, and its randomly generated colour was then saved separately for use in heatmap and bipartite network generation. The monopartite network was then rendered using the Fruchterman-Reingold layout algorithm[203]. The weight parameter corresponded to the k value in the algorithm. The list of weights used for rendering is given in table S4.2. The reason for the different weight values was to attempt to set the repulsive field between nodes strong enough to enable the observation of distinct sub-clusters within a single large component. The repulsive field strength was inversely proportional to the k value. Setting the repulsive field too low resulted in nodes within the network being clustered too closely together to resolve. Conversely, too high a repulsive field resulted in a homologous network with points equidistant. Consequently, local clusters within the network between related sequences could no longer be observed. Because each network was generated from a different pool of sequences from each subtype, which had differing levels of similarity,

different weight parameters were required to achieve a similar level of discernment between nodes.

2.3.7 IQtree tree generation for each CRISPR-Cas subtype

In addition to network generation using vConTACT2, phylogenetic tree generation was also performed using IQtree2[192]. Each bait/seed protein utilised to recognise and retrieve CRISPR encoding contigs and assemblies corresponding to a particular system subtype was employed to perform a BLASTp search against a prodigal-predicted set of all ORFs found within 20kb of a CRISPR array, which were extracted from the original 10TB block. Each protein ortholog was retrieved if it matched above an e-value of 1e-7 and was then sorted into clusters at least 70% similar using mmseqs2. A single representative sequence from each cluster was then pooled into a single file and aligned using clustal omega under the default parameters. The Multiple sequence alignments (MSAs) were then parsed to IQtree2 which generated a tree utilising the following command:

iqtree2 -s <input_MSA> -st AA -m MFP -T <number of threads> -alrt 1000 -B 1000 -bnni

Extensive (1000) bootstraps of the trees were performed during tree generation to ensure the branches were approximately accurate.

2.3.8 Calculating the conservation scores of genes, genome shapes and genome types

Genome shape and genome type information produced in Chapter 2.3.4 was associated with the sequences in each of the clusters generated by Leiden partitioning in Chapter 2.3.6. The conservation score outlined in Chapter 2.1.4 was then employed to calculate the proportion of assembled sequences in each cluster predicted to be (linear/circular) by PLASME and Virsorter2 and (ssDNAphage/dsDNAphage/Lavidaviridae/NCLV/RNA phage) by Virsorter2. An important difference between PLASMe and Virsorter2 was that PLasME always returned a linear/circular assignment (due to the design of PLASMe as a binary-classifier), whereas assignments were only made by Virsorter2 if homology could be established

above an e-value based threshold. As a consequence of this, a host_cell category was assigned to sequences not possessing homology to any phage category. Similarly, DNA not assigned as circular by Virsorter2 was designated as linear. This meant that designations made by Virsorter2 were an underestimation of the true phage and plasmid diversity present in the sample. The phage and plasmid predictions for each cluster were then aggregated together into a matrix. This matrix was then used to generate a heatmap using the pheatmap() function as part of the pheatmap 1.0.12 package in R.

Each of the predicted Pfam and DEFLOC annotations (Chapter 2.2.5) assigned to each host-encoded subtype sequence encoding the CRISPR-array ± 20kb DNA sequence upstream and downstream, as their corresponding set of spacer-mapped phage sequences (for each contig ± 10kb from the target sites) was also associated with each of the partitioned clusters by sequence identifier. Conservation scores (performed as per Chapter 2.1.4) were then calculated for each predicted gene in each cluster. These were aggregated into a matrix of genes in each cluster for each host-encoded subtype/mapped sequence. This matrix was then visualised using pheatmap(). For the heatmaps generated and presented in Chapter 4, an additional filtering criterion was applied. For each gene to be included in the map at minimum conservation of at least 30% in one cluster or 15% in two or more clusters was required.

2.3.9 Computing cluster-level diversity scores

To calculate the samples/species diversity index in each partitioned cluster within each monopartite network generated, the sequence annotations were retrieved from GOLD a via master table containing information on all sequences within JGI and NCBI for which a GOLD analysis ID exists (URL: https://gold.jgi.doe.gov/download?mode=site_excel) . For sequences derived from NCBI a two-pronged approach was utilised. Firstly, sequences were retrieved from the Genbank assembly summary files found on the

NCBI FTP repository located at https://ftp.ncbinlm.nih.gov/genomes/genbank or directly retrieved using the Entrez efetch tool, from the Bio.Entrez module within Biopython 1.79.

To identify differences between individual clusters generated by vConTACT2, the clusters were numbered and differences in the species/sample diversity; proportion of sequences with a given genome shape and/or a given genome type (cell/phage) were calculated based on the fraction of contigs identified with one of the above properties divided by the total number contigs for which an annotation could be detected. In the case of PLASMe, a putative linear/circular assignment was always given, due to the design of PLASMe as a binary-classifier where the only two categories are circular/linear DNA. In the case of virsorter however, the proportion of contigs labelled linear/circular and cell/ssDNAphage/dsDNAphage was the total abundance of each detected category divided by the total number of contigs in each cluster. Importantly, this score was almost certainly an underestimate, due to the denominator representing the number of detected cell/host based contigs plus the number of contigs for which a phage-like annotation could not be assigned. This could be due to weak/below threshold similarity detected by Virsorter profiles (false negative) or the genuine absence or any plasmid/phage homology.

The normalised Shannon diversity index was then computed for each host-encoded partitioned cluster within each network. The formula employed to compute the Shannon diversity is given below:

$$S = f(n) = \begin{cases} 0 & , n = 0 \\ 1 & , n = 1 \\ \dfrac{nln(n) - \sum_{i=1}^{k} f_i \ln(f_i)}{nln(n)} & , n \neq [0,1] \end{cases}$$

Where:
S = Shannon Index
n = total number of species/sample categories
K = number of each individual species/sample

The diversity scores for both host CRISPR-Cas system encoding systems, and spacer-mapped putative phage sequences, were each aggregated together into a single matrix

then visualised using a heatmap plot, via the pheatmap() function as part of the pheatmap 1.0.12 package in R.

2.3.10 Host-phage network generation between selected CRISPR-Cas subtypes and spacer-mapped putative phage targets

Each of the viral clusters produced by the output of vConTACT2 when generating host and mapped sequence monopartite networks were used as the main nodes in a corresponding host-MGE bipartite network. These nodes were related by a spacer-mapping table detailing which host-encoded spacers mapped to which target sequences. A table was constructed which associated these mappings with the identities of sequences in each cluster. The colours of partitioned clusters in the networks generated in Chapter 2.3.5 were also included in this table. These mappings were then used to form an edgelist from which a bipartite network was constructed using igraph [204] in R. These were visualized using the Fruchterman Regingold layout algorithm as in Chapter 2.3.6. The weights used to set the repulsive field strength are given in (table S4.2, S4.3 chapter 4). The colours of each node were inherited from the colours of each partitioned cluster generated by the procedure detailed in Chapter 2.3.5.

2.4 Characterisation of spacer mapping patterns across CRISPR-Cas systems

2.4.1 Generation and deduplication of redundant of PPS-spacer pairs

To characterise the spacer mapping patterns, mapped spacers for each putative CRISPR-Cas subtype identified by the running of the procedure specified in Chapter 2.3.2 were first filtered to take only spacers of which at least two or more spacers from the same array mapped to the same target phage contig. The spacer furthest away from the promotor in the CRISPR array was designated as the PPS. The distance between this

spacer and the PPS was then computed for each additional mapped spacer to the same mapped sequence. As units, these were designated PPS-spacer pairs.

The next step involved removing redundant spacer pairs. Firstly, the PPS-spacer pairs were grouped together when they mapped to the same contig. If the distance values for these pairs were less than 20bp then these pairs were grouped together as if from the same CRISPR-array. Then, each array group was considered compared to the set of mapped sequences targeted. If the pairwise difference in distances between the PPS-spacer pairs was less than 20bp, then only one distance value was allowed. The effect of this was to deduplicate cases where either multiple spacer pairs from different CRISPR-array targeted the same, or almost the same target site on the same contig, or cases where the same PPS-spacer targeted multiple contigs, which were redundant versions of the same contig. A side effect of this method was a reduction in the resolution of the resultant spacer distribution. It also meant that distances shorter than 20bp could not be visualised. This approach was similar to that which was employed when this technique was first developed [175]. However, it was more consistent and less case-based compared with this earlier approach increasing the reliability and reducing the risk of biases or artifacts being created during deduplication.

2.4.2 Generation of spacer mapped distribution plots

A mapped spacer distribution was defined as a kernel density estimation (KDE) of the non-redundant PPS-spacer distance pairs subtended from the PPS from unrelated CRISPR-arrays grouped by subtype. To compute the KDEs for the observed and expected distances, the density() function, from S3 base R (version > R 4.3.1) was used. A smoothing window of 50bp was set along with n=512. These parameters were chosen as a compromise between optimising the density and resolution of the peaks displayed by the KDE. Importantly, the range was set to [-5000,5000]. This limit was set because most detectable priming/spacer acquisition biases only occur within this range. A statistical approach utilizing the Kolmogorov-Smirnov test (KS-test) was then applied to determine whether the mapped spacer distribution generated by the PPS-distances reflected an acquisition bias compared to the background probability of observing a given PPS-spacer distance on a contig by chance. This is equivalent to the probability of

observing a distance D over a contig of length L. It can be proved mathematically that the average distance between two points on a line, positioned at random is (L/3). Using this fact, the expected spacer distance for each contig could be calculated as a function of contig size. The mapped spacer distances were then scaled by the length of the spacers. This meant that the distribution of scaled spacers was expected to fluctuate about a constant value with equal probability in a manner independent of the size of the spacers. The probability of a spacer acquisition bias could therefore be determined by comparing this scaled spacer distribution to a p-uniform distribution using the KS-test. The observed vs expected distances for each strand were then visualized using ggplot2() using a mirrored line plot chart.

To test for any significant bias in the strand directionality of acquisition, a binomial test was utilized which asked whether the number of observed spacers in a given quadrant was significantly different from the probability (1 in 4) of a spacer falling randomly into any of the other quadrants. This approach was also employed to identify significant differences in the number of spacers mapped onto the target and non-target strands based on binomial probability (1 in 2). Both tests were computed using the EMT library, available as a package for R 4.3.1 or greater. To determine the statistical power of the test, 10,000 Monte Carlo simulations of the KS-test against a uniform distribution with different sample sizes were performed. Sample sizes which returned greater than 80 % of the simulated KS-test probabilities at a significance threshold <0.05 were retained as 1-β threshold values.

2.4.3 Kmer-based searches to identify partial spacer matches

To map partial spacer matches against phage sequences, each CRISPR array and mapped sequence containing at least with one spacer match from spacer mapping (Chapter 2.3.2) was used to produce a sliding window of kmers. Kmer sizes of 10, 12 and 14 were used but size=14 was used exclusively for later spacer distribution analysis due to much lower false-positive rates from background spacer mapping to contigs by chance. These kmers were then mapped against the same sequences as the original match using a window up to 20kb of DNA ± 10kb from the original mapped target site. To

prevent matches by kmers to the original mapped site, the original target sites were masked by substituting these nucleotides with 'X' characters. Duplicate spacers in the same CRISPR-array were also removed. To eliminate any kmer-hits targeting their own CRISPR-arrays which had escaped previous filtering methods, the direct repeats bounded upstream and downstream by 5bp spacer overhangs were also kmerised and matches to the same loci on the same mapped phages were removed from the dataset.

2.4.4 Testing kmer vs. complete match enrichment

To test to enrichment of partial spacer matches against complete spacer matches, kmers produced from spacers derived from six CRISPR-Cas subtypes (Type V-A, Type V-B, Type V-F1, Type VI-B, Type I-B, type I-D) were mapped against a set of contigs identified during the spacer mapping step described in Chapter 2.3.2, which contained a least one match to a known spacer and deduplicated as described above in Chapter 2.4.1. The number of partial matches was then compared to two controls. Firstly, each of contigs mapped was subjected to 1,000 Monte Carlo simulations. The original target site on the contig was also masked to prevent any possible kmers formed from the original mapped spacer mapping to the original target site. Partial matches were then deduplicated by the construction and mapping of 5bp spacer handles barcoded onto either side of the adjacent direct repeat followed by filtering of any matches which were shared by the original kmers, and those including the direct repeats and spacer handles. This filtering step was analogous to the procedure described in chapter 2.3.2. A p-value was calculated based on the number of times partial matches for a given kmer size after Monte Carlo simulations were higher than the original kmer-matches to the same contig containing the complete spacer match.

As a separate control, kmers generated from spacers from a given subtype were mapped to a contig of the same size as the contig containing the original mapped spacer, but with a completely random sequence of nucleotides. This was done to establish the background level of kmer-matches at random based on a set of contigs of the same total size as the original contigs. Although no replicates were performed, the

level of background kmer mapping was much lower than any of the reshuffled kmer matching runs, from which it is virtually certain that the background level of kmer mapping by chance was much lower than that observed. These results were then aggregated for each subtype and each control run (original-kmer mapping, reshuffled-randomised and randomly simulated negative control) and plotted using ggplot2.

2.4.5 Computation and generation of spacer mapped distributions using partial spacer matches

Spacer mapping from "complete" matches for each subtype identified via spacer mapping in chapter 2.3.2 were merged with the kmer-matches (size=14 only) identified in chapter 2.4.3. Due to the kmer mappings only occurring on target sequences within a window up to 20kb in size, the expected distances were revised downwards to be a function of this reduced contig length. PPS-spacer distances were computed then deduplicated using the same procedure specified in chapter 2.4.1. A spacer mapping distribution and binomial strand directionality test were then both calculated as described in Chapter 2.4.2.

Chapter 3: Using a computational pipeline to survey gene diversity associated with CRISPR-Cas systems

3.1 Background

Antiphage defence systems are abundant across the archaea and bacteria to protect prokaryotes against mobile genetic elements. CRISPR-Cas systems are adaptive defence mechanisms under continual selection pressures to provide phage resistance by specifically binding and/or degrading viral DNA or RNA[48,63,147,205]. To survey the genetic landscape and discover new mechanisms of defences, big data searches on large volumes of prokaryote genomes and metagenomes assemblies have been undertaken, pioneered from the Koonin laboratory [81,87,88,90,102,110,206-208]. Guilt by association approaches have been employed to uncover new CRISPR-associated genes. This strategy led to the discovery of new effector and accessory genes, mostly near the CRISPR-array (less than 5kb on average) with co-occurrence (measured by CRISPRicity or weighted naive scores) with conserved core CRISPR-Cas system constituents, such as CRISPR interference genes or CRISPR arrays[81,87-90,102,110,206-208]. However, less attention has been given to associated genes with lower co-occurrence with core CRISPR-Cas system components. There is a presumption that the degree of co-occurrence of a co-encoded protein with CRISPR-Cas is proportional to the indispensability of its function in CRISPR mediated immunity. However, this assumption has been recently challenged by a series of discoveries that CRISPR-Cas systems form part of larger defence islands, where multiple anti-phage defence operons are co-localised within the same contig, and exert synergistic functions to provide protective effects against intruders [22,23,53,209,210].

Many past large-scale surveys of metagenomic data at the terabyte scale set high co-occurrence thresholds, and stringent parametric filtering criteria such as requiring translated genes to be of a minimum length (e.g. 300 or 750aa) or minimum distance from the CRISPR-array or known CRISPR-linked genes (< 5-10kb)[22,57,81,82,88,89,102]. However, the observation that defence islands synergise with distant CRISPR-Cas systems raises the possibility that a significant number of potentially important genes are more weakly

associated with CRISPR-Cas systems, which remain to be discovered via the use of more permissive screening criteria (i.e. lower co-occurrence scores, and greater distances from the core CRISPR-Cas system components).

To discern whether these additional antiphage defences and accessory proteins exist that are associated with CRISPR-Cas subtypes, and play conjoined roles in phage defence, I constructed a computational pipeline emulating previous seed/bait-based designs to extract CRISPR-associated proteins. I reasoned that a significant number of novel associations may remain unreported, as past investigations only interrogated a small part of the CRISPR-associated gene landscape[81,87-90,110]. Such studies do not report the false positive discovery rates of experimentally validated CRISPR-associated proteins, which are selected by arbitrarily chosen filtering thresholds.

I hypothesised that additional associations could be discovered with more inclusive filtering criteria applied to more expansive datasets. After screening ~TB of genomic data for putative CRISPR-associated genes, a conservation score was applied to place associated genes within the genomic context of existing CRISPR-Cas subtypes. The intermediary and filtered set of candidate novel CRISPR-associated genes were used to address several key questions:

1. What is the extent of the bioinformatic and experimental characterisation of the CRISPR-Cas sequence space obtainable by utilising terabyte-scale big-data mining approaches on publicly available sequence data?
2. Are there additional unreported genes in novel associations with CRISPR-Cas systems?
3. Do the gene structures support a possible role for these proteins in CRISPR-immunity, even at distal positions from the CRISPR-array?

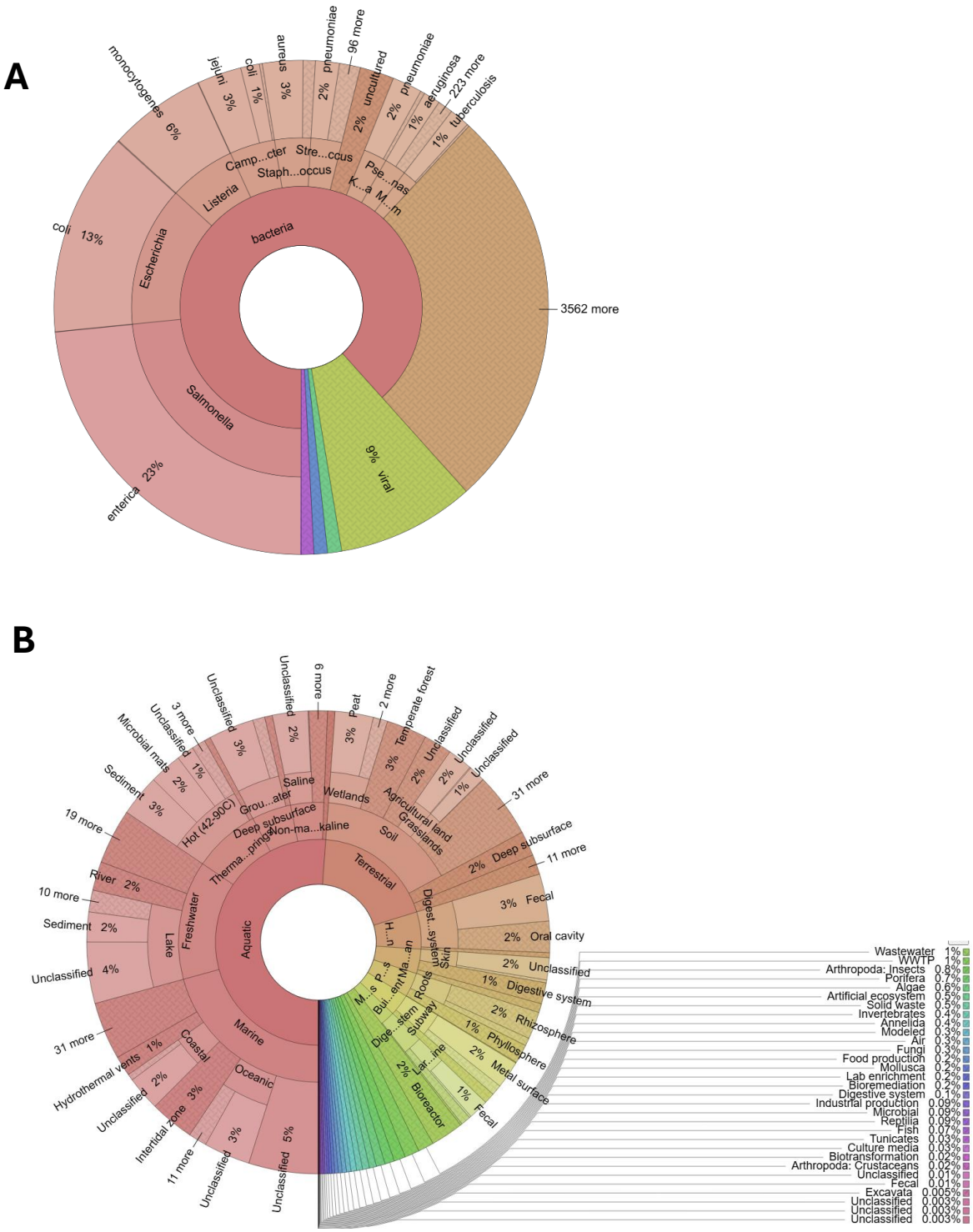
3.2 Survey of CRISPR-Cas systems based on guilt by association

To conduct a survey of the space of known and unknown CRISPR-associated proteins, approximately 10TB of assembled genomes and metagenomes was downloaded from the Joint Genome Institute (JGI) and the National Centre for Biotechnology Information (NCBI) Genbank prior to March 2020. A breakdown of the origin of the assemblies along

with the estimated composition of data from each repository is given below (Figure 3.1A, Table S3.1).

3.2.1 Composition of assembled sequence data used for mining and identifying CRISPR-associated genes

Data downloaded from JGI were exclusively from uncultivated metagenomes, while NCBI assemblies were approximately 89% cultivated genomes. 97% of the data analysed from JGI was derived from just 13 different biotypes including aquatic, terrestrial, and gut microbiota (Figure 3.1A). Of the remainder, 8% were viral in origin, while approximately 1% were derived from archaea, and 1% were unassigned. The large number of NCBI sequences were from individual taxa of disease significance, such as *Salmonella, Campylobacter, Escherichia,* and *Staphylococcus*. The NCBI/JGI dataset compositions introduced a large substructural bias towards mammalian gut microbiota communities (Figure 3.1B). This bias remained in the dataset when NCBI and JGI sequences were subsequently aggregated together into a single 10TB datablock for detection and extraction of CRISPR-arrays and their associated proteins.



<u>Figure 3.1:</u> ***Aggregated assembly, scaffold and contig species and environmental composition of terabyte data block used as the feedstock for subsequent guilt-by-association based data mining.*** *Sequences sourced from the NCBI-Genbank database (A) were largely from cultivated microbes and hierarchically delineated by*

kingdom, genus and species respectively. Annotations for sequences derived from the Joint Genome Institute (B), were hierarchically classified top-down based on the biome, local environment and the material the sample was composed of.

3.2.2 A computational pipeline for mining and identifying CRISPR-associated genes from assembled sequence data at scale

To screen the assemblies for CRISPR-Cas systems, sequences were then parsed into a computational pipeline for detection, screening and annotation (Figure 3.2). In general, the parameters used in this pipeline emphasized enhanced sensitivity, as the goal was to survey and capture the metagenomic landscape to assess the fraction of CRISPR-associated genes discovered, rather than assessing the integrity of specific genes or systems. PILER-CR was selected as the primary CRISPR array prediction tool due to its reasonable sensitivity, high prediction speed and stable performance (lower chance of mis-identifying other tandem repeats as CRISPR arrays) compared to other CRISPR array detection tools [93-96,211]. Using this tool, 3.76 million CRISPR-arrays were detected, and a 20kb 'window' of DNA sequence upstream and downstream of the array was then extracted, which was in-line with the maximum window sizes used in previous investigations[88,89]. In cases where the windows overlapped, reflecting the presence of two separate CRISPR arrays within 20kb of each, the windows were merged into a single continuous contig, to prevent duplication of the overlapping regions. To retrieve CRISPR effectors and accessory proteins, ORFs on each contig were predicted using Prodigal [101] and clustered into putative families of proteins using mmseqs2 [187]. Approximately 12 million ORFs were predicted, which clustering reduced to ~70,000. Further parametric filtering criteria, such as requiring each family to have at least three representatives, a minimum size of longer than 300aa, and no overlap with the CRISPR-array were then added to remove artefacts/false positives from downstream analysis. This approach was a trade-off. It removed many clusters whose largest member was fragmented yet also removed many small CRISPR-associated proteins. Importantly, no adjacency criteria were imposed, such as mandating the protein to be encoded less than five ORFs from the CRISPR-array, which in conjunction with the larger window size, is a departure from previous works[81,87-90,102,110]. This enabled the detection of ORFs at significantly greater average distances from the CRISPR array. For the remaining ~31,000 proteins,

the distance between each protein and the corresponding CRISPR-array encoded on each contig was then calculated, and the average distance from the CRISPR-array taken for each cluster. A single representative sequence (centroid) from each cluster was then used to compute the CRISPR-icity score via blast search. The CRISPR-icity score for these ~15,000 query proteins was the proportion of the total number of hits in the entire 10TB datablock, compared to hits in the extracted windows around CRISPR-arrays [88-91,102]. From this calculation, I deduced a vector of the CRISPR-icity scores corresponding to the average distance from the CRISPR-array and the total abundance of each putative Cas gene within 20kb of the CRISPR-array.

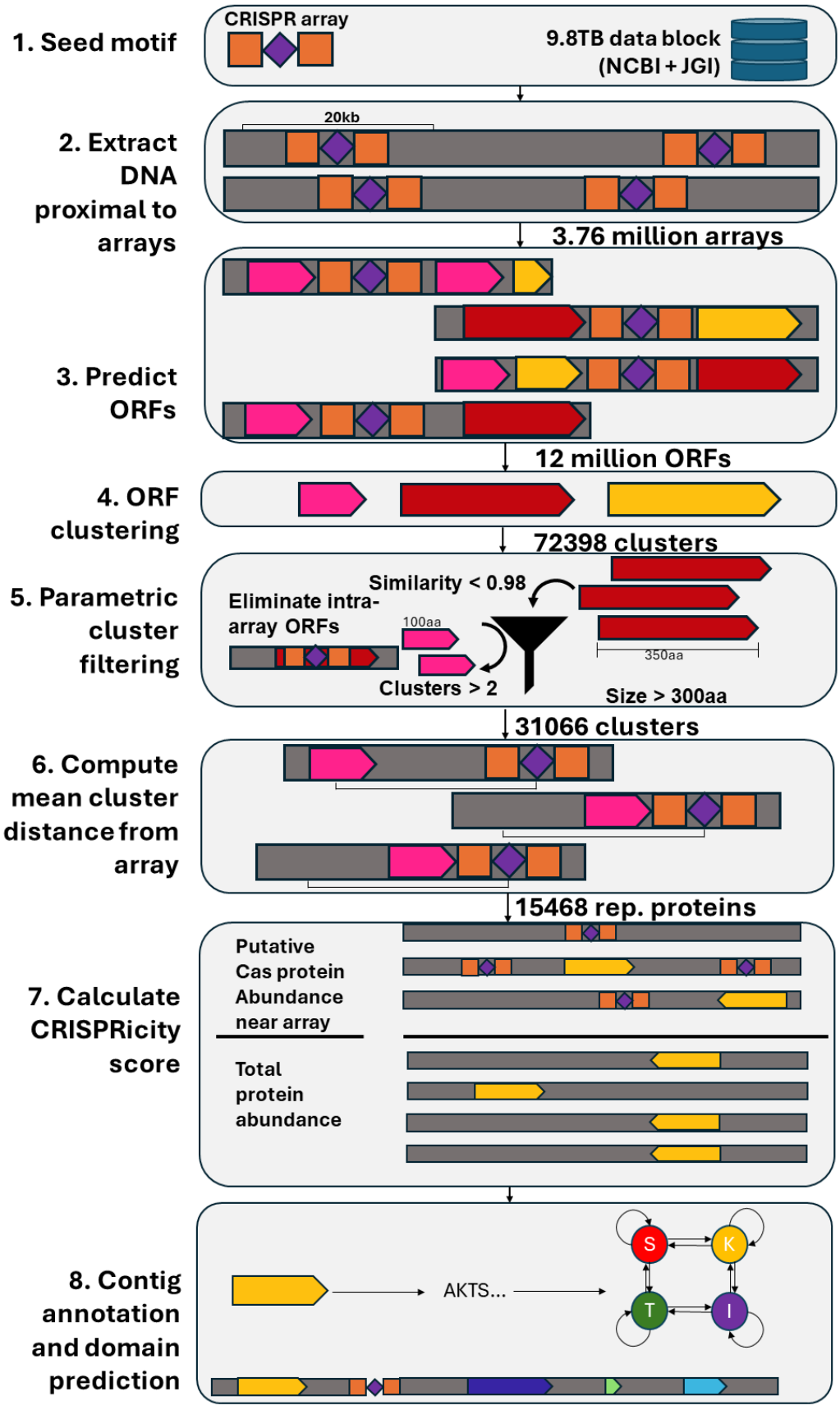

1. Seed motif
CRISPR array
9.8TB data block
(NCBI + JGI)
2. Extract DNA proximal to arrays
20kb
3.76 million arrays
3. Predict ORFs
12 million ORFs
4. ORF clustering
72398 clusters
5. Parametric cluster filtering
Similarity < 0.98
Eliminate intra-array ORFs
100aa
350aa
Clusters > 2
Size > 300aa
31066 clusters
6. Compute mean cluster distance from array
15468 rep. proteins
7. Calculate CRISPRicity score
Putative Cas protein Abundance near array
Total protein abundance
8. Contig annotation and domain prediction
AKTS...
S
K
T
I

Figure 3.2: ***Schematic of computational pipeline used to uncover new CRISPR-Cas systems using the CRISPR-array as the seed/bait/motif to detect putative systems and their associated proteins.*** *(1) CRISPR-arrays were first detected from assembled scaffolds and contigs. (2) The DNA 20kb upstream and downstream of the predicted arrays was then extracted. (3) Open reading frames (ORFs) were then predicted from the extracted DNA sequences using prodigal. (4) ORFs were then clustered into putative protein families using mmseqs2. (5) Clusters were filtered using simple parametric criteria such as largest sequence size, cluster size, effect cluster size once artefacts/redundant sequences were removed etc. (6) average distance from the CRISPR array for each cluster was then computed. (7) A CRISPRicity score was then calculated for a subset of representative sequences corresponding to each cluster. This was computed as the proportion of sequence orthologs within 20kb of a CRISPR-array compared to the total number of orthologs in the original 10TB data block. (8) Each screened representative was then subject to HMM based annotation to assign a putative function.*

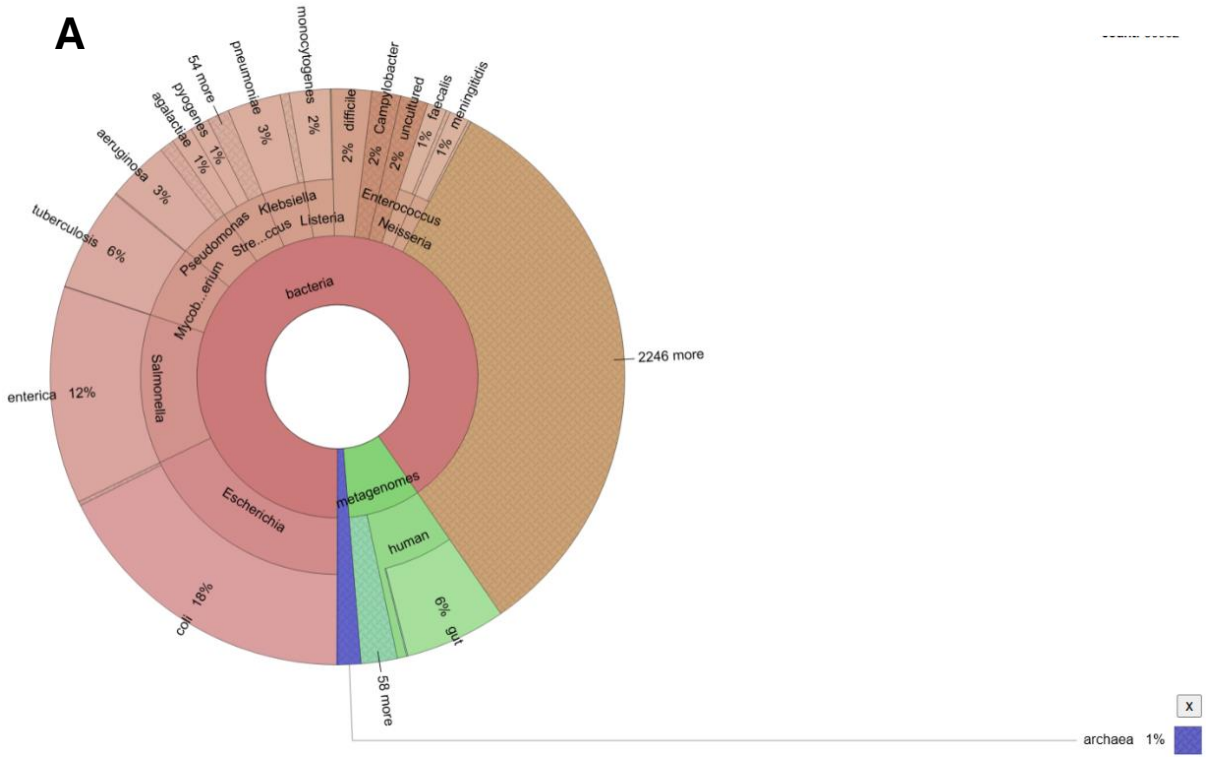
One of the most striking observations made during the running of the computational pipeline was that the taxonomy of CRISPR-array-enriched contigs differed from the initial dataset, suggesting an over or under representation of CRISPR system in certain taxa. When the taxonomic composition of CRISPR-array containing windows was analysed (counted by individual species/sample diversity) from the NCBI repository, the pre-existing bias towards human-gut microbiota datasets was preserved (Figure 3.3A), constituting at least 80% of all sequences. However, the most represented taxa such as *Escherichia coli* and *Salmonella enterica* contained fewer samples encoding CRISPR arrays relative to less well represented taxa. (Figure 3.3A). The proportion of metagenome samples containing CRISPR-arrays was also six-fold elevated, indicating that CRISPR-arrays may be more concentrated in uncultivated microbes than highly represented taxa in assemblies. Taxonomic diversity of CRISPR-array containing contigs in JGI assemblies was however taxonomically congruent with the starting JGI dataset (Figure 3.3B).

Interestingly, this consistency only held when NCBI accession or GOLD analysis ids were counted only once. When the data block was analysed using proportions

measured by the raw abundance of each CRISPR-array from each phylum, stark biases in the data appeared. Approximately 70% of sequences containing CRISPR-arrays were uncultivated metagenomes, of which almost half the CRISPR-arrays sourced from the NCBI repository were derived from gut human microbiota (Figure S3.1A). In the JGI derived dataset approximately 65% of CRISPR-array containing sequences were derived from freshwater samples, of which around 20% came specifically from thermophilic environments (Figure S3.1B).

There were several potential causes for the observed substructural bias in both datasets. The NCBI Genbank dataset features certain taxa which were heavily oversampled relative to other species in the dataset, prior to 2020. Furthermore, because the GenBank dataset is a synchronised aggregation of European Nucleotide Archive (ENA) and DNA databank of Japan (DDBJ) database sequences[212], in addition to directly deposited sequences, additional duplication of certain sequences also often occurred[106]. Similarly, uneven sampling of uncultivated JGI data resulted in a disproportionate abundance of sequencing data for certain environments. These disparities in both datasets were amplified when CRISPR-arrays disproportionately occurred in high concentrations of sequencing data from uncommon but well-sampled environments (i.e. Hot springs from Yellowstone national park) (Figure S3.1B). Nevertheless, these biases did not invalidate my observation that CRISPR-arrays are detected in greater concentrations in sequencing data of microbial communities from human gut microbiota and hot-spring environments.

.


A
bacteria
metagenomes
Escherichia
Salmonella
Pseudomonas
Klebsiella
Listeria
Enterococcus
Neisseria
Campylobacter
monocytogenes 2%
difficile 2%
uncultured 2%
faecalis 1%
meningitidis 1%
pneumoniae 3%
54 more
pyogenes 1%
agalactiae 1%
aeruginosa 3%
tuberculosis 6%
enterica 12%
coli 18%
2246 more
human
6% gut
58 more
archaea 1%
x

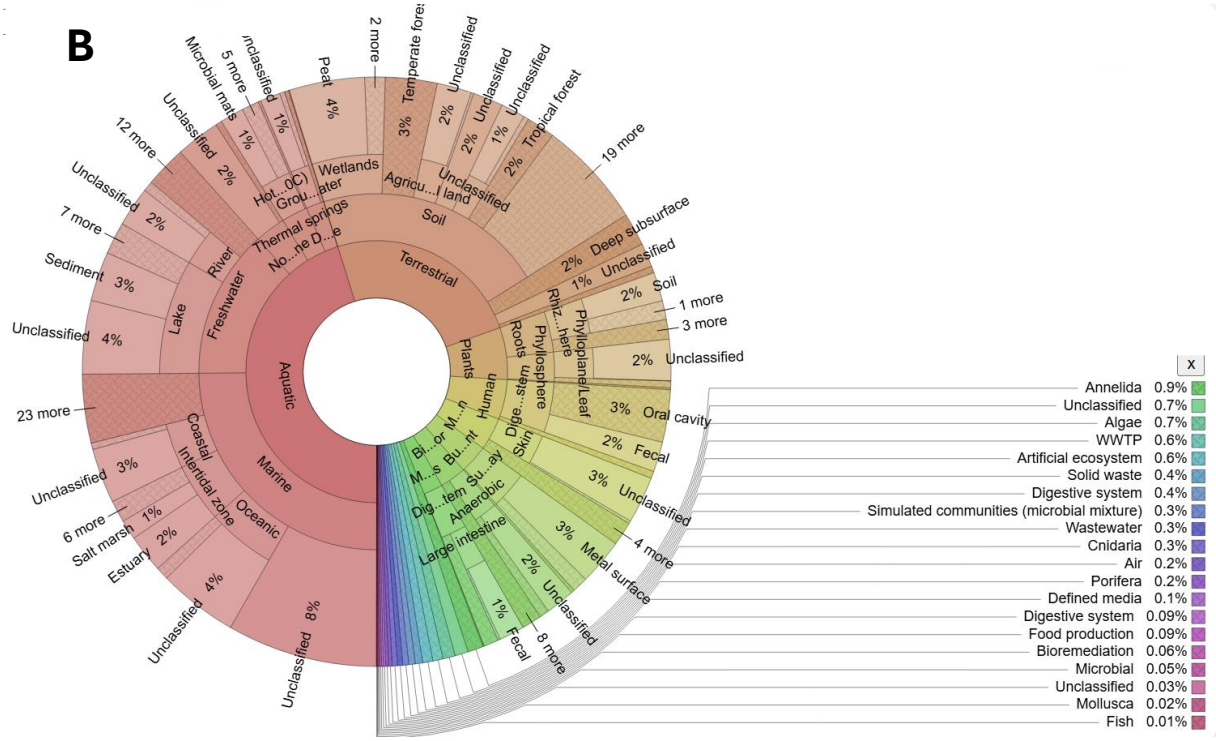

B
Terrestrial
Aquatic
Soil
Plants
Human
Wetlands
Thermal springs
River
Lake
Marine
Freshwater
Coastal
Oceanic
Intertidal zone
Peat 4%
2 more
Temperate fores
3%
Unclassified 2%
Unclassified 2%
Unclassified 1%
Tropical forest 2%
19 more
Agricu...l land
Unclassified
Deep subsurface 2%
Unclassified 1%
Soil 2%
1 more
3 more
Rhiz...here
Phyllosphere
Roots
Dige...stem
Skin
Phyloplane/Leaf
2% Unclassified
3% Oral cavity
2% Fecal
3% Unclassified
4 more
3% Metal surface
2% Unclassified
8 more
1% Fecal
Large intestine
Anaerobic
Dig...tem
M...s
Bi...or
M...n
Bu...nt
Su...ay
Microbial mats 1%
5 more
Unclassified 1%
Unclassified 2%
12 more
Hot...0C)
Grou...ater
Unclassified 2%
7 more
Sediment 3%
Unclassified 4%
23 more
Unclassified 3%
6 more
Salt marsh 1%
Estuary 2%
Unclassified 4%
Unclassified 8%
x
Annelida 0.9%
Unclassified 0.7%
Algae 0.7%
WWTP 0.6%
Artificial ecosystem 0.6%
Solid waste 0.4%
Digestive system 0.4%
Simulated communities (microbial mixture) 0.3%
Wastewater 0.3%
Cnidaria 0.3%
Air 0.2%
Porifera 0.2%
Defined media 0.1%
Digestive system 0.09%
Food production 0.09%
Bioremediation 0.06%
Microbial 0.05%
Unclassified 0.03%
Mollusca 0.02%
Fish 0.01%

Figure 3.3: ***Taxonomic diversity of extracted CRISPR-array containing windows.*** *Windows were extracted from (A) NCBI GenBank derived sequences and (B) uncultivated JGI sequences. Classification was hierarchical and delineated as per Figure 3.1. In (A) only one non-redundant root GenBank accession was counted when computing the proportions of each category. Analogously, in (B) only one GOLD Project analysis ID was counted when computing the proportions. To see the direct proportions of each taxa by the gross number of windows extracted, see supplementary figure S3.1.*

3.2.3: Screening putative CRISPR-associated genes by abundance, distance and co-occurrence with respect to CRISPR-arrays

Screening putative CRISPR-associated proteins entails setting appropriate thresholds for abundance, co-occurrence with an existing CRISPR effector, and the mean distance from the nearest CRISPR-array. Previous works have identified CRISPR-associated proteins that are required for CRISPR effectors to interfere against phages with maximum efficacy and are encoded in close proximity to a CRISPR effector[56,60,88,139,149,213-216].

I postulated that identification of CRISPR accessory proteins based on CRISPR-icity, distance, and abundance, would enable the distinguishment of CRISPR accessory proteins from unrelated proteins. To determine the best parameter(s) to classify the putative CRISPR associated proteins, and optimise the threshold values between CRISPR associated proteins and false positives; the co-occurrence, distance, and abundance scores for each associated protein were plotted and statistically assessed using F-tests (Figure S3.2 and Table S3.2). The aggregate correlation of CRISPR-associated proteins across the co-occurrence, abundance and distance dimensions was also compared to a set of existing CRISPR-associated proteins as positive controls (Figure 3.4) to distinguish at which thresholds clusters of putative CRISPR-associated proteins were the same as known proteins. While abundance was by far the strongest varying dimension when compared with distance (F-ratio 15.52, Figure S3.2A and Table S3.2), the magnitude of the observed variance was high. I observed that a small

outgroup of highly abundant putative proteins was proportionally more CRISPR-associated (Figure 3.2A-3.2B, abundance > 10,000). However clear cluster-based separation between CRISPR associated proteins and false positives were not observed (Figure 3.4).

By contrast, co-occurrence was a better differentiation marker of CRISPR-association than either distance or abundance. This finding was in line with previous studies [90,91]. Similar to abundance, variation of co-occurrence scores was much greater than distance (F ratio=14.27, Figure S3.2C, Table S3.2) The variation in scores in both co-occurrence (CRISPR-icity) and abundance was exponentially distributed, which necessitated conversion to a logarithmic scale to visualise these representatives. Two separate clusters were observed at co-occurrence thresholds of 0.1 and 0.001 respectively and abundances from 1-100,000 (Figure 3.4B)). The high co-occurrence distribution corresponded directly to strongly conserved CRISPR-associated proteins (Figure 4B). The lower distribution observed may be partly attributed to false positive proteins which are ubiquitous across many prokaryotes, such as enzymes required for metabolic reactions; or may correspond to other anti-phage defence systems co-occurring with CRISPR systems (toxin-antitoxin, restriction-modifications), or may represent less conserved proteins which remain nevertheless associated with CRISPR arrays.

Distance to the CRISPR array was the weakest metric for measuring association of a putative CRISPR protein to a CRISPR effector. Although the CRISPR-icity score with various CRISPR-Cas systems weakens with increased distance, no sharp decline was observed for CRISPR accessory proteins over a 0-10kb interval (Figure S3.3A, S3.3C) This suggests that additional CRISPR-associated proteins may occur at mean distances greater than 10kb from the array (outside the measurement window).

.

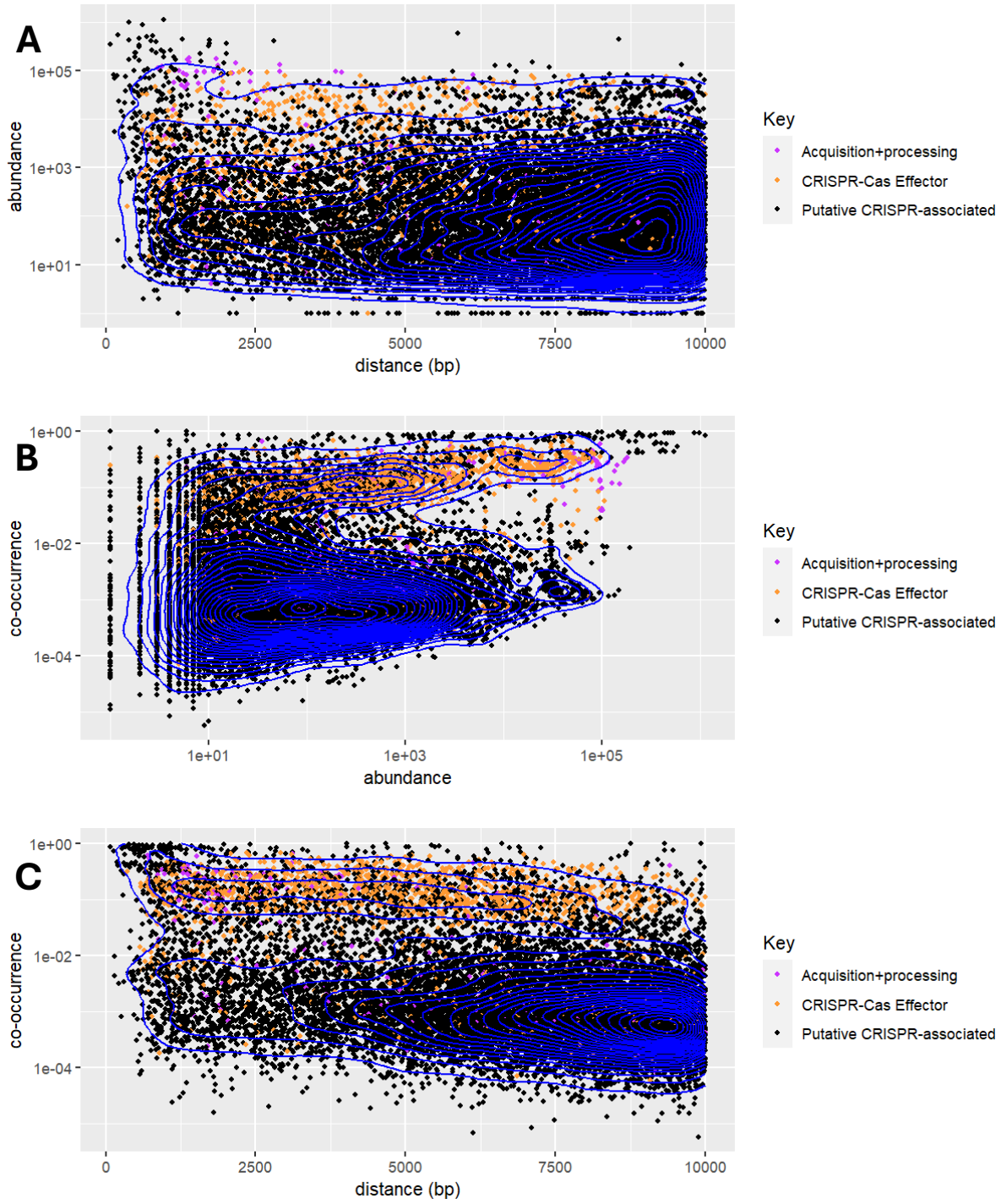


Figure 3.4: ***Distribution of candidate CRISPR-associated proteins based on abundance, CRISPR-icity (co-occurrence) and the mean distance from a CRISPR-array.*** *Relations between co-occurrence and distance from the CRISPR array (A), abundance within 20kb of a CRISPR-array and distance (B) and co-occurrence and distance (C). Each point represents one putative CRISPR-associated protein. Contours assigned each point into one of 25 bins based on their density relative to each other.*

3.2.4: Census of families of predicted CRISPR-associated genes

To identify each cluster, representatives from each putative Cas family were subject to HMM based searches using either the Pfam database (Figure 3.5), or custom-made profiles from CRISPRCasFinder, PADLOC, and Defence finder databases[86,190,191,217] (Figure S3.4A). This database was termed "DEFLOC" and was used alongside pfam due to being specific to just CRISPR and anti-phage defence genes, while Pfam labelled genes with a more general non-specific domain annotation. Annotations from each representative were then classified into nine categories (Table S3.3). Representative proteins were plotted based on distance, abundance and CRISPRicity scores to distinguish CRISPR-associated sequences from associated sequences based on the aggregation of point clusters compared with points from known CRISPR-associated proteins. As anticipated, most CRISPR-Cas effectors displayed high (0.01 -1) CRISPRicity scores (Figure 3.5A). However, only 2-15% of proteins co-encoded with 10kb from CRISPR-arrays corresponded to known CRISPR-Cas effectors, depending on whether Pfam profiles or DEFLOC was used for protein annotation (Figure 3.5B, Figure S3.4A). Similarly, the total abundance of known acquisition and processing proteins was less than 1% of all proteins co-encoded within 10kb of CRISPR-arrays with known CRISPR-accessory proteins only constituting approximately 1-2% of all proteins (Figure 3.5B, Figure S3.4A). CRISPR-Cas effector, acquisition, pre-crRNA processing proteins tended to occur in a constrained space, at co-occurrence values of between 0.01 to 1, and within approximately 6-8kb of the array, before a noticeable drop-off in abundance was observed (Figure 3.5A and Figure S3.4A). This pattern was not observed for other known CRISPR-accessory proteins, where a disposition to co-occur with CRISPR-arrays in high abundance and small mean distances was much weaker. This indicates that many accessory proteins do not co-associate strongly with CRISPR-arrays compared with the core CRISPR-machinery, despite important functional roles being shown to exist[88,90,149,218].

Interestingly I observed a distinct band of proteins which were predicted to be antiphage defence genes (restriction-modification, Abi), as well as conjugation factors from plasmids, both of which have no known role in CRISPR immunity. These proteins

tended to have lower co-occurrence scores compared to the core CRISPR-Cas component proteins (0.01-0.001) (Figure 3.5A). A subset of defence island proteins occurred in a distinct cluster with co-occurrence, abundance, and distance coordinates (0.005,10^3-10^4,0-4kb) respectively (Figure 3.5A, S3.4A , S3.4B). The overlap between this cluster and acquisition/pre-crRNA processing proteins may imply a more interdependent role for this subset of antiphage defence proteins in CRISPR-immunity than the other non-CRISPR defence proteins detected.

Intriguingly, at all CRISPRicity scores ranging from 1 to 0.001, a majority of associated proteins either were not detected or undescribed for CRISPR-Cas immunity (Figure 3.5, S3.4). This was true both at high abundance, distance, co-occurrence scores as well as much lower values. This indicates that these sequences are not merely incidentally co-encoded sequences but uncharacterised proteins with potentially important functions in both CRISPR-immunity and non-CRISPR antiphage defence systems in defence islands. Furthermore, the presence of large numbers of antiphage defence, phage, and plasmid related genes at the lowest measured co-occurrence and abundance values suggests that this cluster of protein is not composed primarily of background noise or false positives, as has been proposed, but are disproportionately defence-island genes and mobile genetic elements.

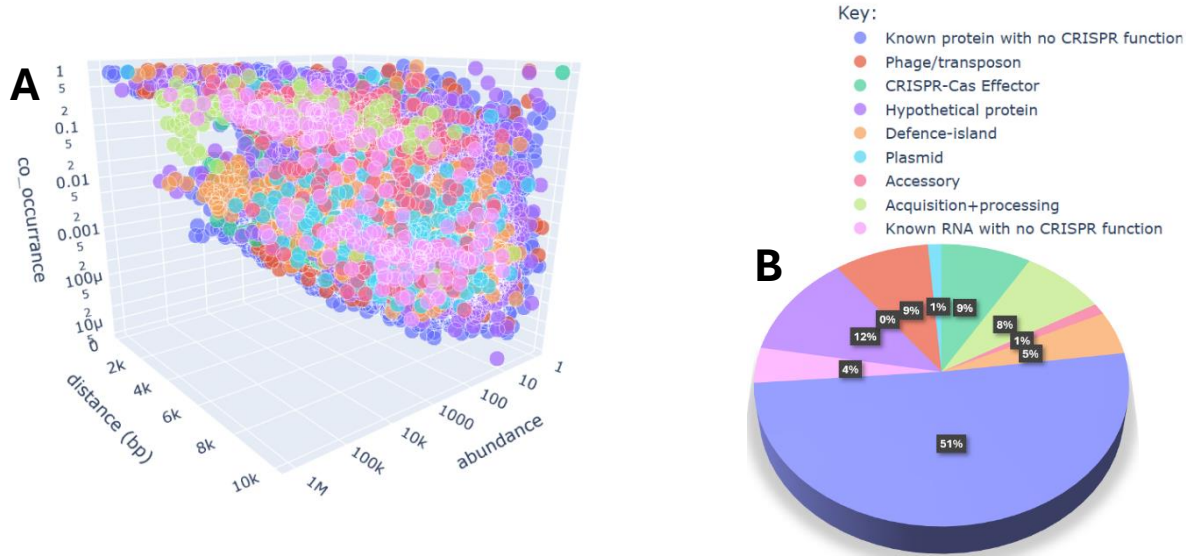


Figure 3.5: Different categories of CRISPR-associated proteins have both well conserved and highly variable co-occurrence, distance and abundance scores. (A) *Each CRISPR-associated protein was represented in (distance, CRISPRicity, abundance) space and labelled by category, which was derived from protein sequence annotation. (B) the proportions of CRISPR-associated proteins were quantified by total abundance. Abundances <1% were absent from the chart. Annotation was determined by homology to pfamA_35.0*

3.2.5: Conservation of CRISPR-associated genes in individual CRISPR-Cas subtypes

Based on the finding that a subset of anti-phage defence proteins with no known function in CRISPR immunity were detected in an adjacent region of the co-occurrence-abundance-distance space to acquisition and pre-crRNA processing proteins, I postulated that some of these proteins may be co-encoded proximate to CRISPR-Cas systems which would suggest an overlapping role for these proteins in facilitating CRISPR-based immune responses.

To determine whether screened putative CRISPR-associated proteins, including anti-phage defence proteins co-occurred with known CRISPR-Cas subtypes, I selected 500 of the most abundant CRISPR-associated proteins within each subtype. Proteins were identified and annotated by Pfam or DEFLOC, to cross-validate predictions from both profiles, and to take advantage of different specificities of labelling and expand the total number of sequence-profile matches. A 'conservation score' was then calculated; defined as the proportion of each co-encoded protein detected in the total set of contigs of each CRISPR-Cas subtype investigated. This method was considered more accurate than total protein abundance, and suitable for assembled aggregate metagenomic data, unlike other recently developed co-association scores which require mostly complete genome assembly of the organisms for which the anti-phage defence genes are investigated[23]. Known CRISPR-associated proteins, which formed the core acquisition, processing, and interference were excluded from this list except for a small number of CARF-domain containing proteins which, alongside known CRISPR-associated accessory proteins, such as WYL, were included as positive

controls. The conservation scores were then cross-correlated with each CRISPR-Cas subtype and visualised to identify highly conserved proteins both within and across different subtypes (Figure 6).

I uncovered 496 protein representatives, co-encoded with at least one CRISPR-Cas subtype with a conservation score of at least 5%. Unfortunately, many of these proteins were either distant homologs of known CRISPR-associated proteins, or proteins possessing a highly abundant conserved sequence or domain, neither of which constituted a novel co-association but were false positives. To eliminate false positives, several parametric filtering criteria were developed. These included excluding candidate novel systems by homology to common domains such as ATPase-AAA and ribosomal subunits, prioritising proteins with a conservation score greater than 20%, and excluding candidates where conservation was proportional to the abundance of each subtype. Of the subset of conserved proteins present, annotated by Pfam, there were 48 proteins, with a conservation score above a 20% cutoff threshold (Figure 3.6A, Figure S3.5A). Interestingly, less than half a dozen proteins representative domains detected, such as WYL, which is known to function as DNA/RNA-sensing transcriptional regulators or Helix-turn-Helix (HTH) domains, which occur in a variety of CRISPR-associated proteins, including cOA activated CARF domains as well as sometimes fused to WYL, have previously established roles as CRISPR accessory proteins across multiple subtypes[88,214,215,219-221] . Repeating this approach by annotating the same set of candidate proteins with DEFLOC instead of Pfam revealed an additional 11 proteins above this threshold (Figure 3.6B, Figure S3.5B). To further validate the conservation scores underlying these candidate proteins, phylogenies of each candidate protein family were constructed to determine to confirm the monophyletic relatedness between orthologs and remove candidate families which multiple origins (Figure S3.6). This further eliminated 25 of the candidate proteins. This left a set of 34 potentially novel co-associated protein representatives. To understand the genomic neighbourhood under which these of these proteins are co-encoded and further strength the evidence for a potential role in association with CRISPR-Cas systems, several operons encoding the *HicB*, and *DrmB* genes from the HicAB and DISARM anti-phage defence systems respectively, were examined in detail.

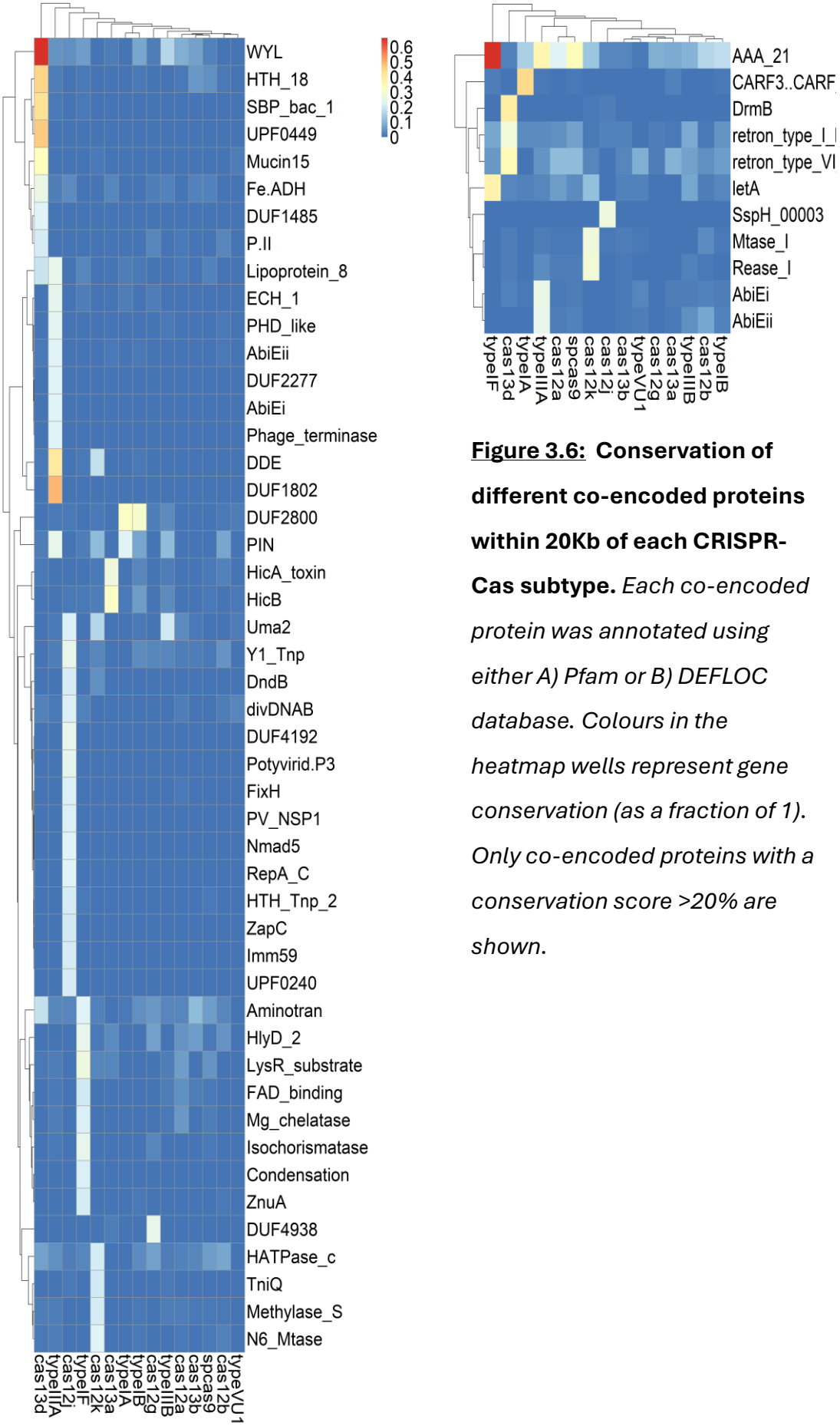


Figure 3.6: Conservation of different co-encoded proteins within 20Kb of each CRISPR-Cas subtype. *Each co-encoded protein was annotated using either A) Pfam or B) DEFLOC database. Colours in the heatmap wells represent gene conservation (as a fraction of 1). Only co-encoded proteins with a conservation score >20% are shown.*

Chapter 3.1.5: Identification novel of CRISPR-associated genes co-associated with type VI CRISPR-Cas systems

Of the 34 protein clusters which showed strong conservation within at least one CRISPR-Cas subtype, *HicB* and *DrmB* were illustrated in more detail as their high conservation score with Cas13 subtypes and close proximity to the CRISPR-Cas effector proteins was suggestive of a more direct interaction between these genes and the CRISPR-Cas system and the antiphage defence cassette. The genomic neighbourhood of the antiphage defence gene "*HicB*" supports a role as an accessory or co-associated component of a subset of Type VI-A CRISPR-Cas systems. The *HicB* gene, in concert with *HicA*, is a known toxin-antitoxin system, which is normally inhibited by HicB mediated transcriptional repression of the *HicAB* operon[222] . When examining three of these operons derived from different samples (Figure 3.7), I found detectable *HicA* and *HicB* homologs were consistently co-encoded in close proximity to the CRISPR Cas13a effector and its associated CRISPR-array. From the observation that the sense direction of *HicAB* and Type VI-A CRISPR genes is the same, it is likely these are expressed under a single operon. In two of the three operons (Figure 3.7A,7C) an additional GajAB anti-phage defence system was identified, also co-encoded in the same direction. I also observed that in one system the *HicB* gene was co-encoded twice in the same operon (Figure 3.7B). This was reflected in my observation that *HicB* tended

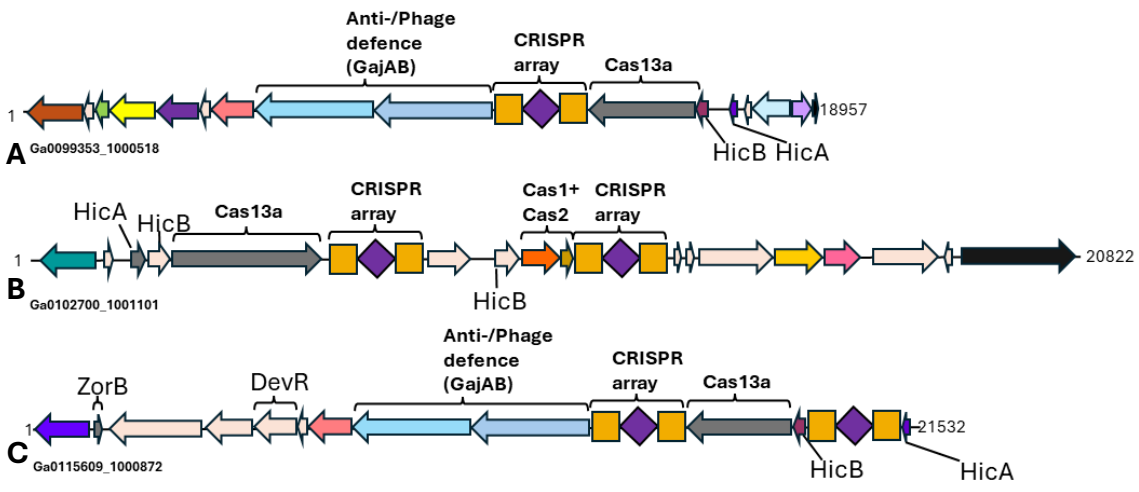


Figure 3.7: Genomic context of Type VI-A (Cas13a containing)-HicB genes. *Each gene cluster (A-C) was retrieved from a different organism. The functions of different genes/clusters were estimated using Pfam, PADLOC+ and PDB70 HMM based domain annotation. This annotation was absent if an approximate function could not be determined.*

to be more conserved than *HicA* (Figure S3.6). This may be a consequence of the importance of *HicB* to the host-cell survival as an antitoxin which must be constantly expressed to repress *HicA*.

Similarly, in Type VI-D CRISPR-Cas system, homologs with distant similarity to the *DrmB* subunit gene of the DISARM phage defence system were identified encoded in the neighbourhood of Cas13d. The DISARM phage defence system has previously been postulated as a methylation dependent restriction modification system. Interestingly, only the *DrmA*/*DrmB* subunits were co-encoded, while other canonical subunit genes such as *DrmC* and *DrMII* were absent (Figure 3.8B-C). Even this was not a conserved feature, with some standalone *DrmA* or *DrmB* being detected in many instances (Figure 3.8A). This, in conjunction with a multiple sequence alignment, and phylogeny of these homologs indicate a strong divergence from homologs which form the canonical DISARM system, although the proteins still appear to be part of the same family and have a single strongly conserved domain. This may imply an alternate function for the DrmAB homologs observed, or alternate supporting methylation and restriction proteins than reported in the canonical DISARM system. Furthermore, although both proteins were observed encoded adjacent Cas13d, these were encoded in opposite sense directions. This may mean that keeping transcription of these proteins separate is somehow important for their function, or important for the survival of the host cell.

Overall, the conservation of HicAB and DrmAB with Cas13a and Cas13d respectively, showed two independent instances of a novel genomic context for non-CRISPR antiphage defence genes in the genomic context of Cas13 effectors. Although the genomic context of other highly conserved associated proteins with various CRISPR-Cas subtypes was not examined, analysing this context in other predicted proteins which were strongly co-conserved with various CRISPR-Cas subtypes may yield additional associations. This demonstrated that high conservation score thresholds are

sufficient to identify uncharacterised gene configurations between putative CRISPR-associated proteins identified by guilt-by-association style data mining methods and known CRISPR-Cas subtypes.

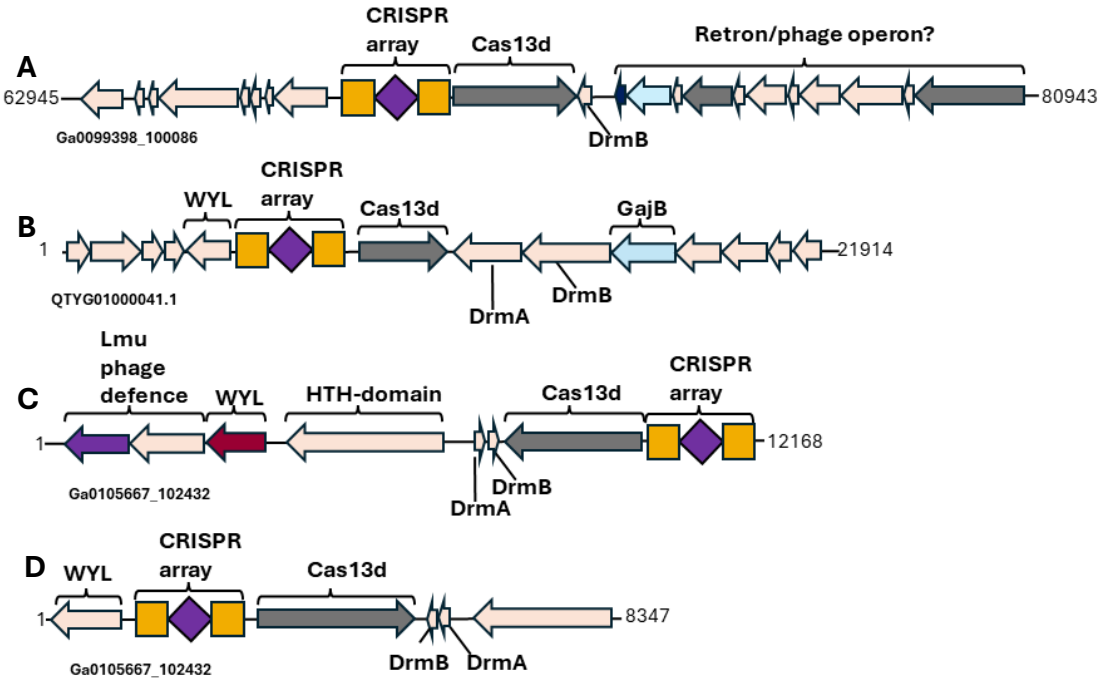


***Figure 3.8:* Illustrations of [Type VI-D]-DrmB co-encoded gene clusters.** *The functions of different genes/clusters were estimated using Pfam, DEFLOC and PDB70 HMM based domain annotation. Genes were not labelled with annotation if the annotation gave a hypothetical protein designation*

3.3: Discussion

Since 2015, many different computational pipelines based on Guilt-by-associated architectures have successfully identified new CRISPR-Cas systems and their associated proteins[81,87-90,102,110,206-208]. It is significant that this study, which was undertaken using a computational pipeline developed prior to 2020 and updated using several redundant HMM profile databases sourced in November 2021 [223] and 2022 [86,191], revealed that between 50-70% of the CRISPR-associated sequence space within 20kb from the CRISPR-array remains uncharacterised. Further analysing the conservation of co-encoded proteins to several known CRISPR-Cas subtypes revealed additional

antiphage defences which are likely associated to known subtypes. The genomic context and proximity of the HicAB and DISARM to Cas13 orthologs suggests a much more direct association of innate antiphage defences with CRISPR-Cas than previously assumed [57,102]. Taken together, these findings suggest that CRISPR-associated genes and accessory components are still prevalent and present in a large proportion of metagenome contigs in concert with CRISPR-Cas systems. Furthermore, co-occurrence with innate immune systems within a defence island to known CRISPR-Cas effector proteins remains high in metagenome contigs.

3.3.1: Evaluation of the effectiveness of CRISPRicity, abundance and distance in isolated CRISPR-associated genes and accessory modules

While guilt by association approaches have been effective at discovering new CRISPR-associated proteins [81,87-90,102,110,206-208], the sensitivity of CRISPRicity, abundance and distance was not sufficient to distinguish true association with CRISPR arrays based on a set of thresholds alone[91,224]. Surprisingly, despite being implied in past investigations to be an important variable separating CRISPR associated versus non-associated genes, my investigation demonstrated that the average distance of a co-encoded protein from the CRISPR-array up to 10kb did not significantly separate these two groups into clusters. Only a moderate decline in the incidence of CRISPR-associated proteins was observed at distances greater than 8kb. This implies that genes may co-occur with CRISPR-Cas systems even when separated by greater distances than the mean distance maximum of 10kb. This observation was supported by one recent investigation which found no strong correlation between the co-occurrence of antiphage defences, including CRISPR-Cas, and the proximity to which they were encoded[102]. Despite abundance, distance, and co-occurrence serving as the main metrics of separating CRISPR-associated from non-associated proteins, only strongly co-occurring proteins (measured by CRISPRicity) which form the core Cas genes, can be reliably selected via this workflow. In past investigations, a minimum CRISPRicity score of 15-70% was used to screen for novel CRISPR-Cas effectors [88,89,102]. Notably, although fewer CRISPR-associated proteins occurred with CRISPRicity scores below approximately 5%, I did not detect a clear cutoff below which CRISPR associated proteins were not detected. Past investigations have emphasised the use of guilt by

association-based data mining approaches as a tool to discover new CRISPR-Cas effectors and associated genes for experimental characterisation [22,60,81,87-91,102,110,206-208,214], but could not rule out functional co-interplay between genes even when CRISPRicity scores were low[90,91]. This was especially true for genes playing an accessory role which were much less well separated by CRISPRicity[90,91]. This may be due to these genes possessing distantly related orthologs with functions unrelated to CRISPR immunity which has been shown in previous studies of antiphage defence systems outside their associations with CRISPR-Cas [22,23,32,53,209,210,222].

3.3.2: Plasmid and anti-phage defence genes account for a significant fraction of co-encoded genes near CRISPR arrays despite lower CRISPRicity scores

An important finding of my survey of the CRISPR-associated sequence space up to 20kb from the CRISPR-array is that sequences co-encoded at greater distances and with lower CRISPRicity scores from the CRISPR-array were not false positives but were potentially mobile genetic elements and antiphage defence related genes associated with CRISPR immunity in defence islands. Approximately 30-50% of genes detected have a recognisable domain or motif. Of these, 15-22% were either plasmid, phage or anti-phage defence gene related compared with 2-3% of genes were annotated as the core adaptation, processing, and interference genes. While the domain classifications of some of these genes may be unrelated to defence, the use of two largely independent HMM profiles, particularly the more specific DEFLOC profiles, verified that these cases constituted a very minor fraction of all detected co-associated genes. It is possible that the number of anti-phage defence genes detected is an underestimate, given that only clusters with at least one member greater than 300aa was included in the analysis, and many anti-phage defence genes are known to be smaller than this[22]. At least some of the uncharacterised protein designations may be due to extreme sequence divergence exceeding the HMM sensitivity to detect remote homology. Advances in structure-based homology search have recently uncovered new anti-CRISPR and Cas13 orthologs by increasing the ability to detect these distant homologs[99,225]. However, because these advances, underpinned by Alphafold2 based structure prediction, use HMM profiles

generated by JackHmmer as the raw input to construct a 3D tertiary structure from sequence,[97] there are likely limitations to the extent to which predicted structures extend detection sensitivity.

Although my study lacked a dedicated category for membrane proteins, the findings in these past publications were concordant with my findings on the overall composition of anti-phage defence genes and plasmid encoding genes[83,169,224,226]. This strongly supports the hypothesis that the majority of genes found encoded in proximity to CRISPR-Cas perform related immune functions such as phage defence, defence gene propagation or sensing the infection.

3.3.3 Examining the conservation of conserved genes in type VI CRISPR systems revealed two additional genes which co-occurred within a subset of these systems

Previous works mainly ascertained gene co-association through parametric filtering criteria based on CRISPRicity [81,88,89,149]. To improve upon CRISPRicity based scoring, several recent investigations have taken different approaches to attempt to refine the accuracy of scores which discriminate associated from non-associated data[23,102]. Protein sequence clustering algorithms which run in linear time and possess a reasonable level of sensitivity, have been developed to reduce both the putative CRISPR-associated and total genome/protein sequence space required to compute CRISPRicity scores. Because a high level of sequence redundancy exists within sequencing data repositories, this both improves the accuracy of the resultant effective CRISPRicity score and reduces computation time. Measuring associations between predicted proteins and CRISPR-Cas systems or other anti-phage defence genes using a conservation score does not assess the relatedness or completeness of the contigs used for the calculation. This contrasts with more recent work[23] which weighted genomes by their phylogenetic distance, and computed a score between all antiphage defence systems, including naturally occurring Type I-E and Type I-F systems in *E. coli* based on the observed vs. expected pairwise distances between antiphage defence systems. However, this method also suffers drawbacks, such as require complete genome assemblies, that would have made it unsuitable for highly fragmented

metagenome datasets. The number of contigs for some subtypes was prohibitively high to construct a well-bootstrapped phylogenetic tree to weight sequences with respect to each other. Computing the distances between all defence systems to show co-associations required complete genomes as opposed to contigs, which formed the majority of sequences in my dataset. Using high conservation scores and the genomic context of putative associated proteins with known CRISPR-subtypes was a simple yet surprisingly effective means of uncovering potential new CRISPR-associated proteins. A conservation score circumvented some of the limitations of CRISPRicity, by assessing only the proportions of genes co-encoded with a single subtype and was most effective when the sequences used for computation were relatively closely related without any highly divergent phylogenetic branches whose position was hard to establish with respect to the other sequences. This is a possible explanation for why I detected HicAB and a variant of the DISARM system associated with Cas13 subtypes which have not been previously reported, as these were mostly encoded on short contigs which were unsuitable for methods of asserting co-association which require whole genomes[23].

Examining the genomic context of the two genes associated by conservation, *HicB* and *DrmB*, revealed three further lines of evidence for association with Type VI-A and Type VI-D systems respectively. These were the small average distance between the co-encoded proteins and the Cas13 effector, probable expression of these genes under the same operon as a consequence of co-encoding in the same sense (*HicB*) or antisense (*DrmB*) direction and the presence of a second co-encoded gene adjacent the first corresponding to a second subunit required for both antiphage defence systems to function. Furthermore, although HicAB has never been previously reported as associated with Cas13a, a previous study has shown a single instance of HicAB to be associated with Cas9[102], demonstrating that this toxin-antitoxin can associate with class II CRISPR-Cas effectors. However, it is notable that in some of the other conserved genes, such as *DrmI*/*DrmII*/*DrmE* and *DrmC*, which have previously been shown as required for DISARM to confer phage resistance were absent[32]. Despite this, the weak conservation of these missing genes increases the likelihood these systems are still potentially functional but utilise a different set of proteins transcribed from different genes to compensate. This was supported by the presence of a

methyltransferase on some contigs which may substitute for the role of *DrmC*. Additionally, *HicB* is known to encode a transcriptional repressor[222] of HicA induced non-specific mRNA degradation. This makes it compatible to transcriptionally repress the activity of Cas13a as well. This raises the possibility that both HicAB and DrmB are true accessory proteins to their respective Cas13 effectors and might play an important role *in vivo* modulating their function.

3.3.4: Limitations of the approach employed in this investigation

These findings should be understood in the context of a number of intrinsic limitations of this study. Foremost among these was the large population structural bias present within whole genome and assembled metagenome contigs from the NCBI and JGI repositories. Particularly with the NCBI Genbank repository, there was both high levels of redundancy, as well as a disproportionate amount of the data belonging to single organisms, which were frequently sequenced as part of human gut microbiota in the course of disease diagnosis. Although the makeup of the Genbank database has diversified further in the past few years since the data for this investigation was downloaded, this bias remains an entrenched feature of this repository, which sometimes resulted in artificially inflated co-occurrence or conservation scores. More even sampling and sequencing of uncultivated microbes at high levels of depth would alleviate these biases and likely result in the discovery of yet more gene-associations with existing CRISPR-Cas system subtypes.

3.3.5: Summary of findings

Overall, my study demonstrates that many of the genes co-encoded with CRISPR-Cas systems have homology to known plasmid and anti-phage defence systems, even with low CRISPRicity scores. Additionally, I discovered two additional genes which appear to co-occur with type VI-A and VI-D systems. This illustrates how despite many previous interrogations of CRISPR-Cas diversity using guilt-by-association approaches, many genes with potential functional roles in direct association, or in concert with CRISPR-Cas systems still potentially remain to be discovered.

To further expand upon this approach, I focused my investigation in the next chapter on analysing differences in gene conservation at the intra-subtype level. This involved

constructing gene-genome networks between sequences from the same subtype. Mirroring this, the gene conservation of the mapped sequences of spacers drawn from each subtype was also analysed. These approaches uncovered local communities of sequences with conserved genes in association with type VI CRISPR-Cas subtypes, which were not observable when co-association was measured at the subtype level in this chapter.

Chapter 4: Network based characterisation of intra-subtype diversity of host-MGE interactions at the gene cluster level.

4.1: Background

CRISPR-Cas systems have tremendously diversified throughout evolution. The immense evolution and diversity of all CRISPR-Cas systems is a consequence of Red-Queen competition between phages and host-encoded antiphage defences. This has resulted in a superfamily of CRISPR-associated genes, a by-product of both divergent and independent evolution from several ancestor systems such as casposons, RNA-guided signalling systems, mobile genetic elements (TnpB, IscB), and HEPN domain antiphage defence antitoxins. To classify this diversity in a robust and self-consistent manner, a sophisticated structure and accompanying nomenclature is required to effectively model the evolutionary relationships between CRISPR-Cas subtypes. Additional considerations are their context as part of larger defence islands, as well as their reciprocal evolutionary effect on the corresponding taxonomic diversity of both phage and plasmidic spacer targets.

The current classification of CRISPR-Cas systems is based on the presence or absence of signature genes within each CRISPR-Cas system [57]. This scheme was developed as an improvement to earlier subtype classification schemes which relied on direct orthology [78]. The key advantage of this approach is that, unlike single gene orthology, it does not require complete conservation of a single feature or gene among all CRISPR-Cas systems [152]. It also tolerates large divergences between orthologs of Cas genes by eschewing the requirement for the direct establishment of sequence homology between orthologs [57,152]. Despite these improvements, this approach does not completely solve outstanding issues in subtype assignment, due to cases where constituent genes in a subtype have different evolutionary histories despite being functionally similar [57,90,226]. This may occur as a consequence of horizontal gene transfer

and cassette exchange by mobile genetic elements, which is known to be a pathway by which CRISPR-Cas systems proliferate between species [227]. It is also arbitrary in designating a threshold at which the signature accessory proteins are considered conserved enough to be used for subtyping. As shown in Chapter 3 Figure 3.4, as well as additional previous works [22,23,86,102], there is clear evidence that many genes may be associated with certain subtypes in addition to the canonical ones used for subtyping. This illustrates the need for alternative approaches for CRISPR-Cas subtyping which avoid these shortcomings.

Many issues observed in CRISPR-Cas subtyping methodology are also observed when attempting to taxonomically label phage sequences from metagenomic sequence repositories. Establishing evolutionary relationships between phage genomes is impossible by orthology alone, due to strong divergent selection of all encoded phage genes in most lineages[151,152,156]. Extensive genome rearrangement or horizontal gene transfer also occur in many phage lineages which prevents an accurate representation using tree-based networks because the individual genes often have multiple points of origin rather than a single common ancestor[57,151,154]. The International Committee on the Taxonomy of Viruses (ICTV) has endorsed a revised classification approach based on gene structure and homology[153]. It differs from the Baltimore classification in attempting to construct the phage equivalent of the Linnaean taxonomy used to represent the tree of life. There are 15 different ranks a given phage genome may be classified under [153,228]. Changes to the ICTV classification allows features from viral metagenomic sequences to be used as the sole means of a taxonomic classification [155]. However, this approach does not specify how these genomic characteristics should be used to demarcate different taxonomic classification levels or quantify the evolutionary distances between different phage sequences [153].

To better classify different CRISPR-Cas subtypes, a new more unsupervised taxonomic classification method is required which incorporates both differences in genes between subtypes as well as the sequence similarity between orthologs. The most sophisticated tool in common usage which has been developed to date is vConTACT2[152]. This tool works by constructing a distance-based network of phage genomes from the predicted Opening Reading Frames (ORFs) within each genome. The similarity between proteins

and viral genomes are then transformed to a set of distance-based edges between protein clusters for each viral genome. These genomes form the nodes of a monopartite graph representing the evolutionary relationship between phage sequences. This approach blends more traditional orthology based classification, which relies on highly conserved genes, with parsimonious classification based on the presence or absence of specific co-encoded genes. This has been shown to improve phage sequence classification[83,152] and may also improve upon existing approaches for CRISPR-Cas subtyping.

Once developed, an integrated approach jointly modelling both host-encoded CRISPR-Cas and phage contigs evolutionary relationships in a representative manner can be applied to further elucidating both the sub-subtype gene level diversity of CRISPR-Cas operons as well as the diversity of the types and regions of phage genomes which are targeted by CRISPR-Cas RNA-guided interference ribonucleoproteins. Although shared gene-genome bipartite networks have previously been utilised to identify conserved genes in CRISPR-Cas subtypes in an unsupervised manner, a systematic elucidation of the joint relationships between local clades of CRISPR-Cas subtypes and their mapped target sequences to understand the intra-subtype diversity of CRISPR-Cas and their corresponding mapped phage repertoire has not yet been performed.

To explore the advantages of unsupervised taxonomic classification in representing CRISPR-subtype diversity, I first compared vConTACT2 based unsupervised gene cluster-level taxonomic classification with a traditional tree-based single-gene phylogeny. I then built upon my prior analysis of conserved genes within CRISPR-Cas subtypes (section 3.2.5), by focusing on the intra-subtype gene composition and diversity of local clusters within networks of type VI CRISPR-Cas subtypes, as well as their spacer mapped target sequences. Host-MGE interaction networks were constructed from the viral clusters generated by vConTACT2 to illustrate the tropism and relations between subtype specific host and mapped sequence communities.

I found that the gene content of many local clusters also tended to be very segregated, with few conserved genes across local clusters. In contrast, a larger number of genes from mapped sequences, corresponding to either phage or plasmid contigs contained shared genes. Host-phage interactions between local clusters of CRISPR-Cas subtypes

revealed considerable diversity in number and composition of MGEs which predate them. This work expands upon the known range of known MGEs which type VI systems have evolved to target which underpins the continuous evolution of these systems as RNA-guided RNA targeting interference complexes.

4.2: Results

4.2.1: Outline of workflow to perform spacer mapping and annotation of CRISPR-Cas subtypes

A pool of CRISPR-Cas subtype orthologs, which was prerequisite to apply unsupervised taxonomic classification to both CRISPR-Cas subtypes and their corresponding mapped phages, was identified from the 10TB data block of GenBank and JGI metagenomes retrieved for the purpose of uncovering new CRISPR-associated genes in Chapter 3.1. Using a signature protein which has been previously shown to define each CRISPR-Cas subtype analysed in this investigation (see table S4.1), a tBLASTn[229,230] search against the set of CRISPR-array encoding CRISPR "windows" extracted in Chapter 3, enabled the retrieval of contigs encoding mostly complete CRISPR-Cas subtypes as well as several genes up to 40kb from the CRISPR array in the neighbourhood (Figure 3.1). Whichever contig encoded the CRISPR-Cas subtype was designated the 'host encoded' CRISPR-Cas system, even in cases where the contig was phage derived. Arrays previously identified on each contig were then validated using predictions from PILER-CR[93], CRISPRdetect [95] and the CRISPR-CRT version included with CRISPRleader [195], with minor modifications to enable CRISPR-CRT to possess multiple input genomes simultaneously. To increase the number of spacers, the largest combined array of either CRISPR-CRT and CRISPRdetect plus one other array prediction tool was taken to maximise the number of spacers while also predicting the orientations of each array, which was later used for spacer distribution analysis in Chapter 5. Spacers were then extracted from each array and used to search the 10TB datablock assemblies for matching sequences. A 7bp handle upstream and downstream of each spacer, containing the adjacent direct repeat sequence to each spacer was appended and utilised to query the data block concurrent with spacers without the 7bp handles.

Un-appended spacer mappings for which a corresponding match also occurred by the spacer-DR queries were removed from the set of mapped sequences as these were self-mapping to spacer sequences encoded in CRISPR arrays. Protein translations of each of the remaining contigs in the dataset for each CRISPR-Cas subtype were then used for subsequent vConTACT2 prediction. Importantly, contigs were parsed to vConTACT2 with relatively little filtering. This was because vConTACT was intrinsically able to filter out contigs containing false positive or unrelated CRISPR-Cas subtypes as distinct clades.

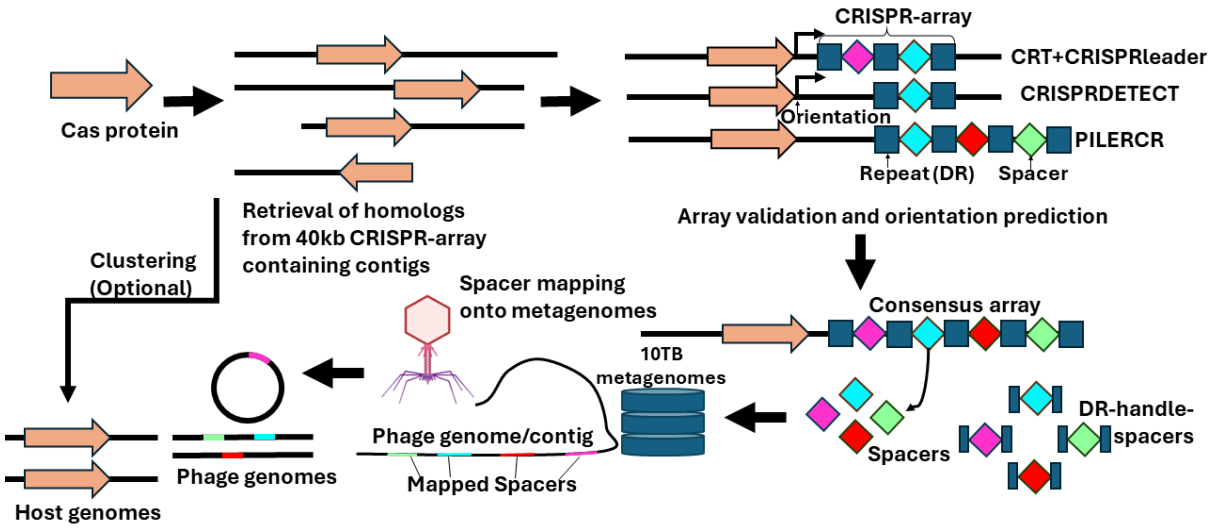


Figure 4.1: ***Overview of workflow utilized to perform CRISPR-array validation and expansion, spacer mapping and the retrieval of a set of mapped Phage (or other MGE) genomes.***

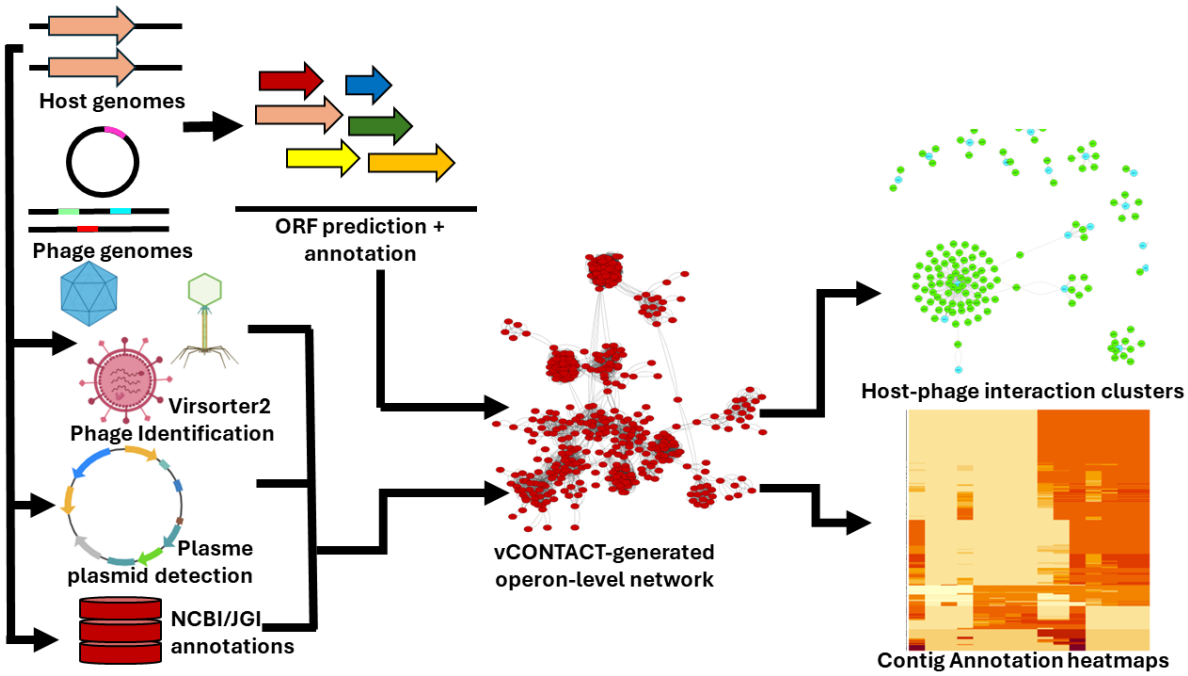
4.2.2: A network-based representation of CRISPR-Cas intra-subtype diversity

The resultant set of host-encoded CRISPR-Cas subtype specific contigs, as well as the set of mapped spacer-target sequences from each CRISPR-array, were then used to generate a monopartite network depicting the relationships between contigs via gene cluster similarity (Figure 4.2). This was accomplished by translating each contig into ORFs using prodigal, then parsing these contig-ORF mappings into vConTACT2.

Concurrent to this, both host and phage contigs were annotated to predict similarity to ssDNA, dsDNA, RNA, Lavidaviridae and NCLDV (jumbo) phages using Virsorter2[231]. Whether contigs were circular or linear was separately predicted using both Virsorter2 and PLASMe[200]predictions. The network was then partitioned into clusters using the Leiden cluster partitioning algorithm [202]. This algorithm was chosen based on its ability to partition the graph into groups of non-overlapping modules. This is an optimisation of the related Louvain method[159]. Unlike the Louvain method, Leiden clustering is not affected by the resolution limit, and is more reliable than the Louvain method, which sometimes produces poorly-connected communities[202]. These modules were the network-based equivalent of clades in phylogenetic trees. This enabled the analysis of the composition of these local groups.

Sequences from each cluster resulting from the graph partition were then matched to corresponding annotations. Heatmaps were generated showing either the protein-coding gene composition or the proportion of sequences in each cluster, which were predicted to be host, plasmid and phage derived. ORFs from each contig in each cluster were annotated via HMM searches to either the DEFLOC or Pfam databases. In cases where a matching entry was found to both databases, the annotation from DEFLOC was used preferentially as it provided a more accurate description of the protein compared with Pfam profiles. Together these maps illustrated the intra-subtype gene cluster and contig compositional differences between CRISPR-Cas loci as well as the same differences between the corresponding spacer-mapped target sequences for each subtype.

To illustrate the joint gene composition and tropism between host-encoded CRISPR-Cas subtypes and their corresponding spacer targets, the host-subtype and mapped sequences were used to form a single host-MGE bipartite network. An initial bipartite network based on single sequences was found to produce a very large number of nodes and few trophic phage-host interactions. To expand upon the number of interactions, I assumed that host or mapped sequences from the same viral clusters generated by vConTACT2 were closely related. Hence by making these clusters into the main nodes in the bipartite network, I could show the propensity of groups of MGEs to predate different local clusters of the same CRISPR-Cas subtypes.



<u>Figure 4.2:</u> Overview of workflow used for network generation and gene composition annotation. *Input contigs used for ORF prediction and gene/contig shape annotation were parsed to vConTACT2 for network generation. Generated host or phage monopartite networks were integrated with ORF, species diversity, plasmid and phage annotations to produce a heatmap showing the composition of each cluster.* These networks were also used to produce a *host-phage interaction network between clusters of host-encoded contigs and their mapped sequences for each CRISPR-Cas subtype.*

4.2.3: Comparison of network and traditional phylogenetic representations of intra-subtype diversity

I postulated that a multi-gene network-based representation of CRISPR-Cas intra-subtype diversity should be more informative than the equivalent single gene phylogeny generated from a subtype-defining effector signature protein, due to the ability of the network to identify associations between orthologs co-associated with the core CRISPR-Cas genes in addition to the proteins which form the main complex used for interference. To assess the effectiveness of using vConTACT2 generated gene cluster networks in modelling evolutionary relationships between contigs, each network generated for each host-encoded CRISPR-Cas subtype was compared to maximum likelihood trees generated for each CRISPR sub-type. vConTACT2 network generation

only includes nodes for which at least one pairwise edge, equivalent to detectable pairwise homology between ORFs on at least two contigs, exists. In some cases, the sequence similarities scores between gene clusters, which delineated a single edge in the monopartite graph, were too low to assign homology. In these instances, the graph instead consisted of multiple components. The network was then further partitioned into clusters. Each corresponding subtype tree was formed by first clustering each input set of sequences using mmseqs2 at a minimum sequence identity of 90% then taking a representative protein from each cluster. This was done to eliminate closely related sequences, which biased the diversity of orthologs towards the most sampled environments or closely related species. Maximum Likelihood tree generation was then performed using IQtree2 [192] with Modelfinder and 1,000 bootstraps [192] (see methods). To illustrate the relative positions of similar contigs predicted by vConTACT2, and their equivalent positions on the tree, the top 20 partitioned clusters were labelled and the corresponding signature CRISPR effector protein encoded on the same contig found in the network was highlighted on each tip of the phylogeny (Figure 4.3).

The congruence between network and tree representations of most CRISPR-Cas subtypes was illustrated and analysed (Figure S4.1). There were three main categories observed, of which Type I-D, Type V-A and Type I-B serve as paradigms (Figure 4.3). In Type I-D systems (Figure 4.3A) I observed almost complete congruence between the main clusters which form the subtype, and their corresponding positions into distinct self-contained clades in the phylogenetic tree. In contrast, in Type V-A systems (figure 4.3B), several clusters such as cluster 76 and cluster 106, were located at disparate positions in the vConTACT2 network, despite being located within the same clades in the tree, suggesting that other genes or non-coding RNA present in the neighbourhood play an important role in V-A interference. Type I-F system (Figure 4.3C) clusters were intermediate between these cases. Several clusters such as 31 and 42, or 5 and 30, displayed congruence with corresponding clades on phylogenetic tree. However other clusters such as cluster 5 and 55, were located on distant clades despite a closer relationship depicted by the network. In all cases, the network representations more clearly separated the groups of sequences into discrete local clusters, compared to the

clades of the equivalent single-gene phylogenetic tree.

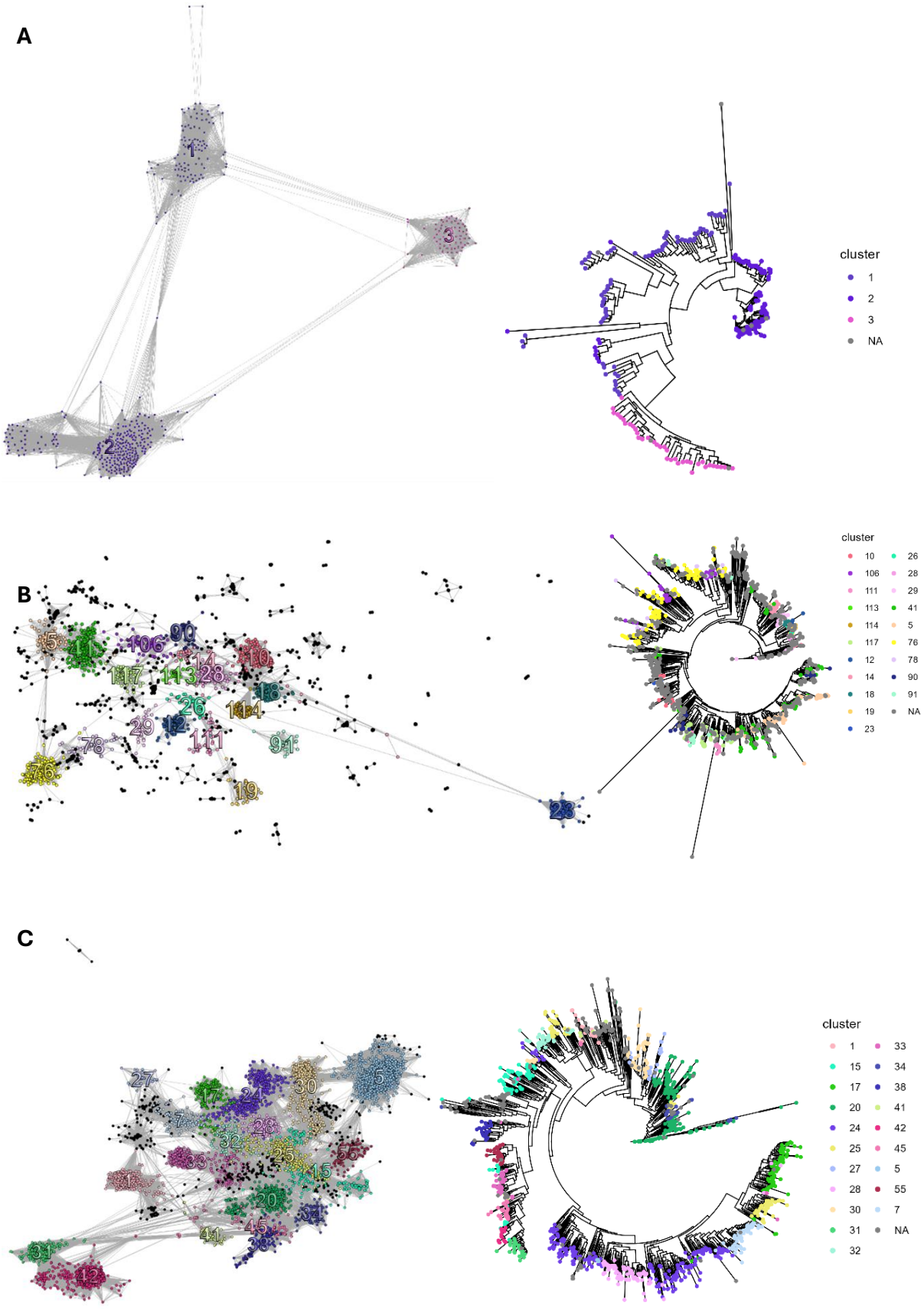


Figure 4.3: ***Comparison of network and single gene phylogenies when visualising intra-subtype diversity.*** *The top 20 clusters by size were labelled and compared. For the full comparison of clusters see supplementary Figure S4.1. Networks were generated using vConTACT2 from a pool of assembled contigs encoding a single CRISPR-Cas subtype plus up to 20kb of assembled sequence upstream and downstream of the CRISPR-array: (A) Type I-D, (B) Type V-A, (C) Type I-F. Trees were generated using iqtree2 from a pre-clustered set of orthologs of the most conserved subunit of the effector proteins of each subtype: (A) Cas10d, (B) Cas12a, (C) Csy1.*

4.2.4: Interrogation of host-encoded intra-subtype diversity in Type VI systems.

Next, I aimed to understand the population substructural differences in operon configuration between different instances of the same subtype. In the previous chapter (Chapter 3 section 3), I discovered several uncharacterised genes associated with Type VI systems. However, the degree to which these genes, as well as known accessory genes such as WYL, Csx27 and Csx28[87,88,149], are conserved within their respective subtypes has only been partially evaluated. Due to their rarity, Type VI systems have been found only in a relatively small number of genera [111,160]. To understand the differences in the gene composition of Type VI systems, as well as clarify the intra-subtype conservation of Type-VI associated accessory genes, I generated gene cluster-based networks of Type VI-A, Type VI-B and VI-D systems (Figure 5.4A-C) and analysed the gene composition of each resultant cluster of contigs (Figure 5.4D-F). Type VI-C, VI-X and VI-Y systems were not analysed due to their rarity [111]. I combined the sequence information with taxonomic annotations from NCBI and JGI repository metadata and calculated the normalised Shannon Index[232] for each cluster, to estimate sample and species diversity. Lower Shannon index scores reflected higher sample/species diversity. Finally, the proportion of contigs in each cluster predicted as viral, plasmid or unassigned (presumably host encoded) was also calculated to detect whether any type VI system clusters were phage or plasmid encoded.

All Type VI gene cluster networks generated (Figure 4.4A-C) comprised several components. The gene-gene dissimilarity between clusters was too high for vCONTACT to be able to establish any interactions despite detectable homology between the conserved Cas13 effectors. Only a small number of genes were found to be shared across clusters in addition to the conserved Cas13 effector. The normalised Shannon index scores for most clusters were > 0.5 and plurality were equal to 1, indicating that in many cases the sequences were derived from a single environment and/or annotated as a single species. This suggests that a significant amount of the diversity observed between genes clusters was partially of a function of taxonomic/environment barriers to horizontal gene transfer of CRISPR-Cas cassettes between taxa.

Based on my past discovery of HicAB and DISARM genes and additional antiphage defences encoded in the vicinity of Type VI-A and Type VI-D arrays, I postulated I would observe these genes across multiple clusters. From the cluster representation, Cas13 effector subtype was the only gene completely conserved across all clusters. This served as a quality control which ensured that all retrieved contigs for each cluster were from the correct subtype. In some clusters of Type VI-A and Type VI-B (Figure 4.4D, 4.4E) more than one copy of a homolog of the Cas13 effector protein was observed encoded on each contig as a result of gene truncation which is reflected in a conservation score for these genes higher than 1. Next most conserved were the acquisition proteins Cas1-Cas2. Across three subtypes, these were strongly observed in some clusters yet absent in others. There was no clear correlation between the sample/species diversity of each cluster (Figure S4.2A-S4.2C) in any type VI subtype. Furthermore, with the exception of cluster 22 in type VI-D systems (Figure 4.4F), there were no CRISPR-Cas systems of other types which could plausibly share their acquisition machinery with type VI-D arrays which has been reported in at least prior instance as one mechanism by which type VI systems acquire further spacers [140]. This was further evidence that the conservation of the acquisition genes may be cluster specific.

Among known CRISPR-associated accessory genes, the most conserved gene was WYL in Type VI-D systems (Figure 4.4F) which was co-encoded in representatives from all clusters and was strongly conserved in the majority of them. It is possible that for

contigs where WYL was not found, the protein was encoded in the native organism outside the sequenced contig boundary. In contrast, Csx27 and Csx28 were selectively and mostly mutually exclusively co-encoded in specific clusters of type VI-B systems (Figure 4.4E) with several clusters completely lacking these genes. This suggests that the role of these proteins as regulators of Cas13b was more dispensable compared with the function of WYL in Type VI-D immunity.

There were several other genes which were also shared across more than one cluster. Notably, the HicA-HicB toxin-antitoxin system identified in Chapter 3, which was shown to be co-encoded with type VI-A systems, was strongly conserved across two clusters (Figure 4.5D). This contrasted with *DrmB*, which was identified with Type VI-D, but was strongly conserved only in a single cluster (Figure 4.4F, cluster 11). In addition to *DrmB*, several other notable genes were shared across more than one cluster in type VI-D systems. [90,218] Orthologs possessing a CARF domain protein were also detected co-encoded with most Type VI-D clusters. These group comprised several different genes, including the *Csx1* and *Can2* nuclease gene, which has previously been observed in type III systems[138,233]

While most clusters of Type VI systems were predicted to be host genome-encoded (Figure S4.3A-C), a single cluster of Type VI-B systems were encoded within prophages (Figure 4.4E and Figure S4.3B). Additionally, it was notable that in type VI-D systems, several type IV secretion genes, such as VirD1 and T4SS-DNA transferase, were detected in clusters 7, 19, 21 and 11. However, plasmid predictions did not identify any circular DNA in these clusters which were therefore designated as host-genome (figure S4.3C).

Overall, these results uncovered a large number of genes specific to each cluster with little overlap. This shows that many co-encoded genes in type VI systems are local-cluster, and likely taxa specific.

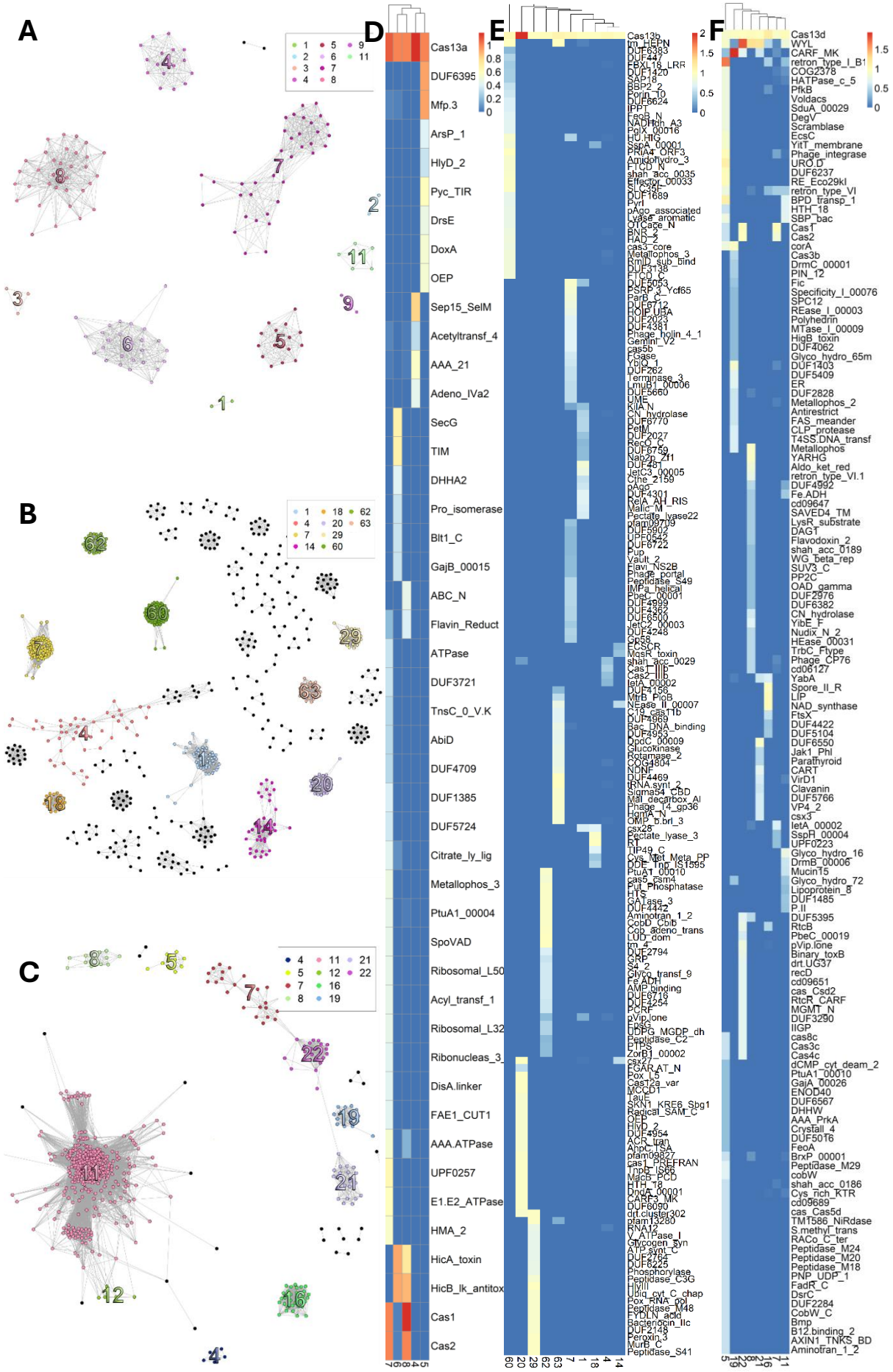

A
B
C
D
E
F
Cas13a
DUF6395
Mfp.3
ArsP_1
HlyD_2
Pyc_TIR
DrsE
DoxA
OEP
Sep15_SelM
Acetyltransf_4
AAA_21
Adeno_IVa2
SecG
TIM
DHHA2
Pro_isomerase
Blt1_C
GajB_00015
ABC_N
Flavin_Reduct
ATPase
DUF3721
TnsC_0_V.K
AbiD
DUF4709
DUF1385
DUF5724
Citrate_ly_lig
Metallophos_3
PtuA1_00004
SpoVAD
Ribosomal_L50
Acyl_transf_1
Ribosomal_L32
Ribonucleas_3_
DisA.linker
FAE1_CUT1
AAA.ATPase
UPF0257
E1.E2_ATPase
HMA_2
HicA_toxin
HicB_lk_antitox
Cas1
Cas2
Cas13b
tm_HEPN
DUF6383
DUF447
FBXL18_LRR
DUF1420
SAP18
BBP2_2
Porin_10
DUF6624
IPPT
FeoB_N
NADHdh_A3
PolX_00016
HU_HIG
SspA_00001
PRIA4_ORF3
Amidohydro_3
FtcD_N
shah_acc_0035
Effector_00033
SLT_2nd
DUF1689
Pyrl
pAgo_associated
Lyase_aromatic
OTCace_N
RNR_2
HAD_2
cas3_core
Metallophos_3
RmlD_sub_bind
DUF3138
FtcD_C
DUF5053
PSRP3_Ycf65
ParB_C
DUF812
HOIP_UBA
DUF2023
DUF4381
Phage_holin_4_1
Gemini_V2
cas5b
FbaseYbiQ_1
DUF262
Terminase_3
LmjB1_00006
DUF5660
UME
KIAA
CN_hydrolase
DUF6770
RetM
DUF2027
RecO_C
DUF6759
Nas2p_Zf1
DUF481
JetC3_00005
Cthe_2159
pAgo
DUF4301
RelA_AH_RIS
Malic_M
Pectate_lyase22
pfam09709
DUF5902
UPF0545
DUF6722
Pup
Vault_2
Flavi_NS2B
Phage_portal
Peptidase_S49
IMPa_helical
PbeC_00001
DUF4999
DUF4362
DUF6500
JetC2_00003
DUF4248
Gp58
ECSOR
MqsR_toxin
shah_acc_0029
Cas7_TIIIB
Cas2_TIIB
letA_00002
DUF4156
MtrB_PioB
NEase_II_00007
C19_cas11b
DUF4969
Bac_DNA_binding
DUF4953
DpdC_00009
Glucokinase
Rotamase_2
COG4804
NDNF
DUF4469
tRNA_synt_2
Sigma54_CBD
Mal_decarbox_Al
Phage_I4_gp36
HgmA_N
OMP_b-brl_3
csx25
Pectate_lyase_3
RT
TIP49_C
Cys_Met_Meta_PP
DDE_Tnp_IS1595
PtuA1_00010
cas5_Csm4
Cpt_Phosphatase
HTS
GATase_3
DUF4442
Aminotran_1_2
CobD_CbiB
Cob_adeno_trans
LUD_dom
tm_4
DUF2794
CRP
S4_2
Glyco_transf_9
Fe_ADH
AMP_binding
DUF6715
DUF4294
PCRF
pVip.Ione
EpsG
UDPG_MGDP_dh
Peptidase_C2
PTPS
ZorB1_00002
csx27
FGAR_AT_N
Pox_L5
Cas12a_var
MCCD1
T4uE
SKN1_KRE6_Sbg1
Radical_SAM_C
OEP
HlyD_2
DUF4954
ACR_tran
AnpC_TSA
pfam09827
cas1_PREFRAN
Tnp_IS66
MucB_RcD
HTH_18
DnaA_00001
CARF3_MK
DUF6090
drt.cluster302
pfam13280
RNA12
V_ATPase_I
Glycogen_syn
ATP-synt_C
DUF2784
DUF6225
Phosphorylase
Peptidase_C3G
HlyIII
Ubiq_cyt_C_chap
Pox_RNA_pol
Peptidase_M48
FYDLN_acid
Bacteriocin_IIc
DUF2148
Peroxin-3
MurB_C
Peptidase_S41
Cas13d
WYL
CARF_MK
retron_type_I_B1
COG2378
HATPase_c_5
PfkB
Voldacs
SduA_00029
DegV
Scramblase
EcsC
YitT_membrane
Phage_integrase
URO-D
DUF6237
RE_Eco29kI
retron_type_VI
BPD_transp_1
HTH_18
SBP_bac
Cas1
Cas2
corA
Cas3b
DrmC_00001
PIN_12
Fic
Specificity_I_00076
SPC12
REase_I_00003
Polyhedrin
MTase_I_00009
HigB_toxin
DUF4062
Glyco_hydro_65m
DUF1403
DUF5409
ER
DUF2828
Metallophos_2
Antirestrict
FAS_meander
CLP_protease
T4SS.DNA_transf
Metallophos
YARHG
Aldo_ket_red
retron_type_VI.1
DUF4992
Fe.ADH
cd09647
SAVED4_TM
LysR_substrate
DAGT
Flavodoxin_2
shah_acc_0189
WG_beta_rep
SUV3_C
PP2C
OAD_gamma
DUF2976
DUF6382
CN_hydrolase
YbiE_F
Nudix_N_2
HEase_00031
TrbC_Ftype
Phage_CP76
cd06127
YabA
Spore_II_R
LIP
NAD_synthase
FtsX
DUF4422
DUF5104
DUF6550
Jak1_Phl
Parathyroid
CART
VirD1
Clavanin
DUF5766
VP4_2
csx3
letA_00002
SspH_00004
UPF0223
Glyco_hydro_16
DrmB_00006
Mucin15
Glyco_hydro_72
Lipoprotein_8
DUF1485
P_II
DUF5395
RtcB
PbeC_00019
pVip.Ione
Binary_toxB
drt.UG37
recD
cd09651
cas_Csd2
RtcR_CARF
MGMT_N
DUF3290
IIGP
cas8c
Cas3c
Cas4c
dCMP_cyt_deam_2
PtuA1_00010
GajA_00026
ENOD40
DUF6567
DHHW
AAA_PrkA
Crystall_4
DUF5076
FeoA
BrxP_00001
Peptidase_M29
cobW
shah_acc_0186
Cys_rich_KTR
cd09689
cas_Cas5d
TMT586_NiRdase
S_methyl_trans
RACo_C_ter
Peptidase_M24
Peptidase_M20
Peptidase_M18
PNP_UDP_1
FadR_C
DsrC
DUF2284
CobW_C
Bmp
B12_binding_2
AXIN1_TNKS_BD
Aminotran_1_2

Figure 4.4: ***Conserved genes of Type VI systems at the individual gene cluster level of monopartite similarity networks.*** *vCONTACT2-based network generation was performed for the 3 most prominent type VI subtypes. (A) Type VI-A, (B), Type VI-B, (C) Type VI-D. The conservation scores of each gene in each of the up to 10 largest partitioned clusters > 5 sequences is shown. Gene composition by local cluster in: (D) Type VI-A, (E) Type VI-B, (F) VI-D was computed from annotations performed by HMM based comparisons against the DEFLOC database.*

4.2.5: Analysis of the gene composition of the mapped targets of CRISPR-spacers from Type VI systems

After analysing the clusters of networks representing gene cluster similarity between Type VI CRISPR-Cas subtype encoding contigs, I next postulated that each type VI system targets a specific repertoire of mobile genetic elements with a characteristic gene composition. I generated a network for each Type VI subtype formed from the set of matching target sequences derived from spacer-searches. Each network was then partitioned into clusters using the Leiden algorithm[202] in the same manner as the corresponding subtype network. The gene composition of the ten largest clusters was then analysed to identify conserved genes among the sequences targeted. Unlike host-encoded subtype contigs, sampling biases were more challenging to determine from metadata-based annotations as a significant fraction of mapped sequences were detected as prophages. This bias was contrasted by large number of phage contigs compared to subtype encoding contigs in each sample, an effect which predominated over the prophage sampling and increased the overall diversity of mapped sites. The proportions of viral and/or plasmidic contigs within cluster was also determined to survey which type of MGEs were most subject to targeting by each type VI subtype (Figure S4.4).

In the mapped sequence set of all three subtypes, a significant number of clusters had at least some detectable homology to known viral sequences. Most of these were dsDNAphages or plasmidic sequences (Figure S4.4A-4.4C). In general, there was greater numbers of shared genes between MGEs compared with the gene compositions of host-encoding CRISPR-Cas subtypes.

In type VI-A mapped sequences (Figure 6A), the vConTACT2-generated network revealed a single large cluster containing a plurality of phage sequences (Figure S4.4A), comprising local clusters 22,28 and 33 (Figure 4.5A). Interestingly, some members of these three clusters were distinguished by encoding endonucleases such as HNH and VRR-NUC domain containing proteins, in addition to a number of recognisable structural phage genes, such as phage integrase and capsid genes (Figure 4.5D). Other clusters, such as 84 and 73, located outside the main clusters did not show any obvious essential shared genes (structural proteins, replication genes…) in the viral sequences (Figure 4.5D).

Type VI-B mapped sequences (Figure 4.5B) were far more diverse and extensive than type VI-A both in terms of the number of clusters and the size of the largest clusters subjected to gene annotation. Compared with the mapped sequences corresponding to type VI-A systems, a larger proportion of mapped sequences were identified as plasmidic (Figure S4.4B). The only notable shared genes across plasmid dominated sequence clusters was the virD1 gene between nearby clusters 216 and 353 (Figure 4.5E) and virC1 gene between clusters 216 and 257. Both genes were involved in DNA processing and excision prior to plasmidic DNA conjugation. Several other type IV-secretion system genes were also observed across local clusters although none of them were conserved.

In contrast to type VI-B, type VI-D mapped spacers were mostly identified as phage encoded (Figure S4.4C). The type VI-D gene cluster network (Figure 4.5C) consisted of one central cluster of mapped sequences (clusters 52, 111 and 122) with homology to known dsDNAphages (Figure S4.4C). The remaining mapped sequence clusters (local clusters 114,128 and 148) did not display significant phage homology, which was consistent with the concurrent absence of any known phage gene components in each of these clusters (Figure 4.5F). These results suggest slightly different spacer target propensities between Type VI-subtypes. Type VI-B systems more commonly targeted plasmidic sequences while type VI-D systems were predominantly observed targeting phages. Type VI-A systems were observed targeting both types of mobile genetic elements though in more instances targeted phage-based sequences.

Despite these preferences, spacer mapped sequences to both these types of MGEs were at least weakly detected in all three type VI CRISPR-Cas subtypes. There was also a significant number of clusters detected across all three subtypes which were either mappings to host-encoded sequences, or targets to MGEs which Virsorter2 and PLASMe were unable to detect and classify. This demonstrates the significant diversity of MGEs and stratification of MGE gene diversity targeted by Type VI systems.

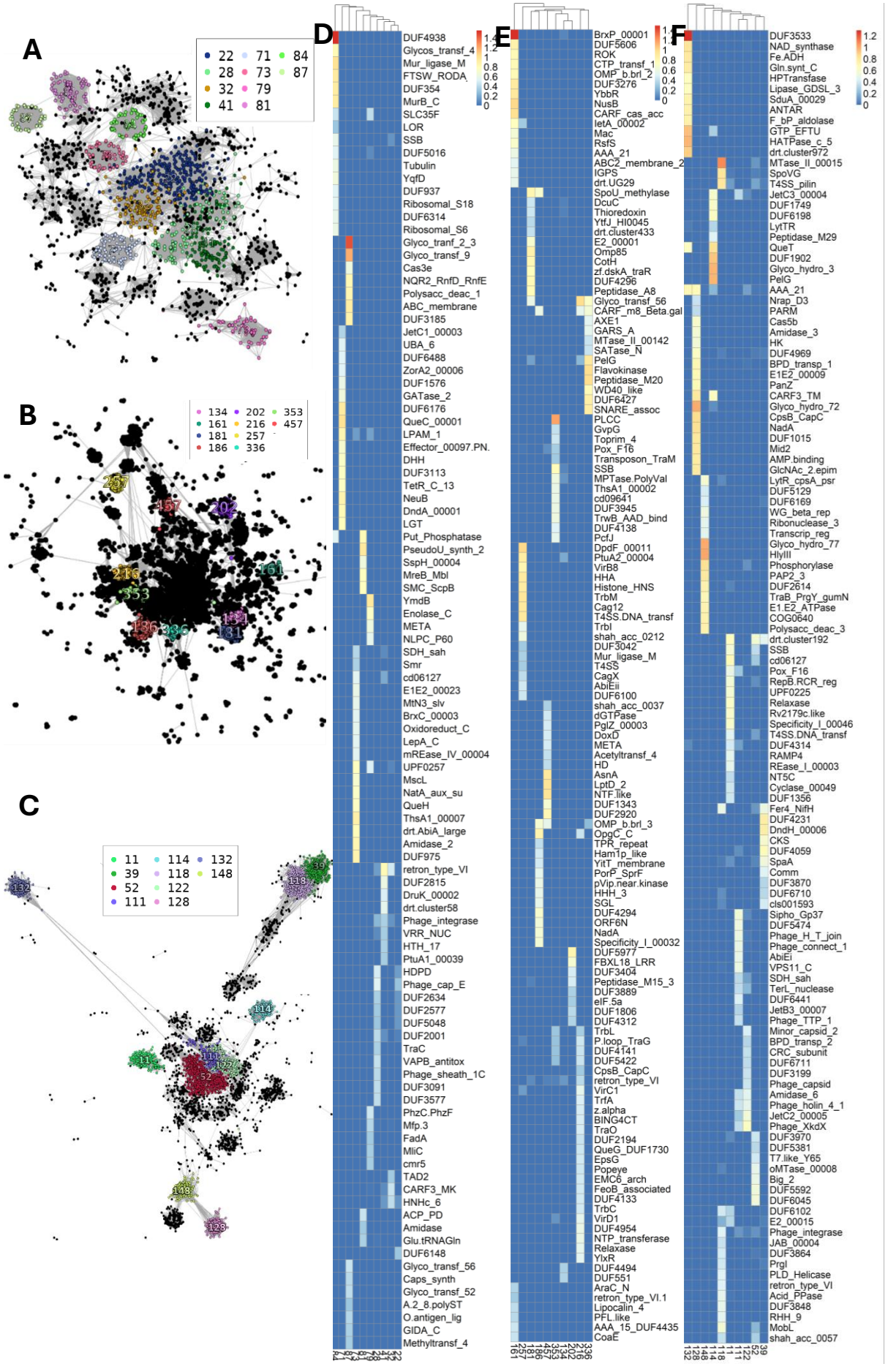

A
22 71 84
28 73 87
32 79
41 81
B
134 202 353
161 216 457
181 257
186 336
C
11 114 132
39 118 148
52 122
111 128
D
DUF4938
Glycos_transf_4
Mur_ligase_M
FTSW_RODA
DUF354
MurB_C
SLC35F
LOR
SSB
DUF5016
Tubulin
YqfD
DUF937
Ribosomal_S18
DUF6314
Ribosomal_S6
Glyco_tranf_2_3
Glyco_transf_9
Cas3e
NQR2_RnfD_RnfE
Polysacc_deac_1
ABC_membrane
DUF3185
JetC1_00003
UBA_6
DUF6488
ZorA2_00006
DUF1576
GATase_2
DUF6176
QueC_00001
LPAM_1
Effector_00097.PN.
DHH
DUF3113
TetR_C_13
NeuB
DndA_00001
LGT
Put_Phosphatase
PseudoU_synth_2
SspH_00004
MreB_Mbl
SMC_ScpB
YmdB
Enolase_C
META
NLPC_P60
SDH_sah
Smr
cd06127
E1E2_00023
MtN3_slv
BrxC_00003
Oxidoreduct_C
LepA_C
mREase_IV_00004
UPF0257
MscL
NatA_aux_su
QueH
ThsA1_00007
drt.AbiA_large
Amidase_2
DUF975
retron_type_VI
DUF2815
DruK_00002
drt.cluster58
Phage_integrase
VRR_NUC
HTH_17
PtuA1_00039
HDPD
Phage_cap_E
DUF2634
DUF2577
DUF5048
DUF2001
TraC
VAPB_antitox
Phage_sheath_1C
DUF3091
DUF3577
PhzC.PhzF
Mfp.3
FadA
MliC
cmr5
TAD2
CARF3_MK
HNHc_6
ACP_PD
Amidase
Glu.tRNAGln
DUF6148
Glyco_transf_56
Caps_synth
Glyco_transf_52
A.2_8.polyST
O.antigen_lig
GIDA_C
Methyltransf_4
E
BrxP_00001
DUF5606
ROK
CTP_transf_1
OMP_b.brl_2
DUF3276
YbbR
NusB
CARF_cas_acc
letA_00002
Mac
RsfS
AAA_21
ABC2_membrane_2
IGPS
drt.UG29
SpoU_methylase
DcuC
Thioredoxin
YtfJ_HI0045
drt.cluster433
E2_00001
Omp85
CotH
zf.dskA_traR
DUF4296
Peptidase_A8
Glyco_transf_56
CARF_m8_Beta.gal
AXE1
GARS_A
MTase_II_00142
SATase_N
PelG
Flavokinase
Peptidase_M20
WD40_like
DUF6427
SNARE_assoc
PLCC
GvpG
Toprim_4
Pox_F16
Transposon_TraM
SSB
MPTase.PolyVal
ThsA1_00002
cd09641
DUF3945
TrwB_AAD_bind
DUF4138
PcfJ
DpdF_00011
PtuA2_00004
VirB8
HHA
Histone_HNS
TrbM
Cag12
T4SS.DNA_transf
TrbI
shah_acc_0212
DUF3042
Mur_ligase_M
T4SS
CagX
AbiEii
DUF6100
shah_acc_0037
dGTPase
PglZ_00003
DoxD
META
Acetyltransf_4
HD
AsnA
LptD_2
NTF.like
DUF1343
DUF2920
OMP_b.brl_3
OpgC_C
TPR_repeat
Ham1p_like
YitT_membrane
PorP_SprF
pVip.near.kinase
HHH_3
SGL
DUF4294
ORF6N
NadA
Specificity_I_00032
DUF5977
FBXL18_LRR
DUF3404
Peptidase_M15_3
DUF3889
eIF.5a
DUF1806
DUF4312
TrbL
P.loop_TraG
DUF4141
DUF5422
CpsB_CapC
retron_type_VI
VirC1
TrfA
z.alpha
BING4CT
TraO
DUF2194
QueG_DUF1730
EpsG
Popeye
EMC6_arch
FeoB_associated
DUF4133
TrbC
VirD1
DUF4954
NTP_transferase
Relaxase
YlxR
DUF4494
DUF551
AraC_N
retron_type_VI.1
Lipocalin_4
PFL.like
AAA_15_DUF4435
CoaE
F
DUF3533
NAD_synthase
Fe.ADH
Gln.synt_C
HPTransfase
Lipase_GDSL_3
SduA_00029
ANTAR
F_bP_aldolase
GTP_EFTU
HATPase_c_5
drt.cluster972
MTase_II_00015
SpoVG
T4SS_pilin
JetC3_00004
DUF1749
DUF6198
LytTR
Peptidase_M29
QueT
DUF1902
Glyco_hydro_3
PelG
AAA_21
Nrap_D3
PARM
Cas5b
Amidase_3
HK
DUF4969
BPD_transp_1
E1E2_00009
PanZ
CARF3_TM
Glyco_hydro_72
CpsB_CapC
NadA
DUF1015
Mid2
AMP.binding
GlcNAc_2.epim
LytR_cpsA_psr
DUF5129
DUF6169
WG_beta_rep
Ribonuclease_3
Transcrip_reg
Glyco_hydro_77
HlyIII
Phosphorylase
PAP2_3
DUF2614
TraB_PrgY_gumN
E1.E2_ATPase
COG0640
Polysacc_deac_3
drt.cluster192
SSB
cd06127
Pox_F16
RepB.RCR_reg
UPF0225
Relaxase
Rv2179c.like
Specificity_I_00046
T4SS.DNA_transf
DUF4314
RAMP4
REase_I_00003
NT5C
Cyclase_00049
DUF1356
Fer4_NifH
DUF4231
DndH_00006
CKS
DUF4059
SpaA
Comm
DUF3870
DUF6710
cls001593
Sipho_Gp37
DUF5474
Phage_H_T_join
Phage_connect_1
AbiEi
VPS11_C
SDH_sah
TerL_nuclease
DUF6441
JetB3_00007
Phage_TTP_1
Minor_capsid_2
BPD_transp_2
CRC_subunit
DUF6711
DUF3199
Phage_capsid
Amidase_6
Phage_holin_4_1
JetC2_00005
Phage_XkdX
DUF3970
DUF5381
T7.like_Y65
oMTase_00008
Big_2
DUF5592
DUF6045
DUF6102
E2_00015
Phage_integrase
JAB_00004
DUF3864
Prgl
PLD_Helicase
retron_type_VI
Acid_PPase
DUF3848
RHH_9
MobL
shah_acc_0057

Figure 4.5: Gene cluster level network analysis and *compositional conservation of mapped sequences of each CRISPR-Cas subtype.* *Each edge in the network represents pairwise sequence similarity between genes encoded on separate contigs, as with host encoded contigs (figure 4). A network was generated by vConTACT2 for the spacer mapped sequences of: (A) Type VI-A, (B) Type VI-B, (C) Type VI-D subtypes. The gene composition between local clusters in the network was then annotated using heatmaps for (D) Type VI-A, (E) Type VI-B, (F) Type VI-D subtypes.*

4.2.6: Visualisation of bipartite interactions between host and mapped-target local clusters

After analysing the gene composition of local clusters of host-encoded type VI CRISPR-systems and their phage and plasmid based MGEs, I wished to visualise the pairwise interactions between the local clusters of these two groups. I hypothesised that certain local clusters of MGEs may specialise in targeting certain groups of host-encoded type VI subtypes. To address this question, I identified pairwise interactions between Type VI subtype clusters and their corresponding mapped sequences. Each viral cluster (VC) identified by vConTACT2, was converted into a single node in both host-encoded CRISPR-Cas subtype contigs, and corresponding mapped sequences. A bipartite host-MGE network was constructed encapsulating host-phage, host-plasmid and host-unknown interactions between clusters for each of the three Type VI CRISPR-Cas subtypes (Figure 4.6).

Most nodes predicted to be predominately phage-derived contigs (Figure 4.6) (clusters 22, 28, 32 and 41) were mapped from spacers from host cluster 7. Host cluster 6 appeared to encode spacers which targeted sequences that were not classified as host or phage derived (clusters 79, 81). There was some overlap, with a node from clusters 6 and 7 containing spacers mapping to both these types of sequences. Manual examination of one of the mapped contigs from cluster 79 unexpectedly revealed the presence of an IS605 transposon which was located downstream of the target site where Cas13a spacer targeting occurred. The CRISPR-arrays containing spacers targeting host-sequences were almost minimal (cluster 6) and contained only 1-3 spacers. Among the phages, 2 of 3 sequences illustrated were prophages (Figure 4.6A, clusters 22,28). It was not possible to discern whether these prophages were still active.

However, the fact that I was able to detect a large region of phage genes (>30kb) (cluster 22) suggests that most of the prophage was intact (80% completeness by genomAD[234]). The contrast in the gene architectures of sequences in each cluster targeted by type VI-A systems highlighted the diversity of potential MGEs these systems are capable of targeting.

Unlike Type VI-A spacers, most of the largest clusters of mapped sequences in type VI-B systems were mapped using spacers from phage encoded contigs (cluster 7) (Figure 4.5B, 4.5E, S4.3E). A large proportion of the targets to these phages were plasmidic (Figure 4.6B, S4.4D). Some of these MGEs which were targeted also had some unique distinguishing characteristics. Foremost among these was the detection of an ISCB encoded within the phage cassette (cluster 202). In a separate cluster, phage encoded tRNAs in addition to the complement of genes required for replication (cluster 134) were also detected. Additionally, several clusters (clusters 216,257) were predicted to be plasmidic. This demonstrated that spacers encoded on Type VI-B appeared to be more ubiquitous in targeting plasmids as well as phages.

Compared to Type VI-B and Type VI-A, the network developed for Type VI-D systems was host-encoded and featured a complex set of interactions within the central component (Figure 6C). This central component encapsulated most of the host-MGE interactions observed between both host and MGEs, indicating a relatively high level of tropism between clusters. In the central component clusters, the main mapped sequences were either predicted as phage, plasmid, or unassigned. Interestingly, the plasmid sequence identified (Figure 4.6C, cluster 52) was designated based on the presence of the VirD4 Type IV secretion gene, which was not detected by Virsorter or PLASMe. Given this observation, it is possible that a significant fraction of spacer targets which were unassigned may have also been in fact plasmidic in nature. This was supported by my earlier Pfam-DEFLOC annotation of these clusters which did weakly detect Type IV secretion ORFs in cluster 52, as well as clusters 11 and 118. This demonstrated the value of analysing the ORF compositions of contigs as opposed to relying purely on plasmid prediction programs automatic MGE assignments.

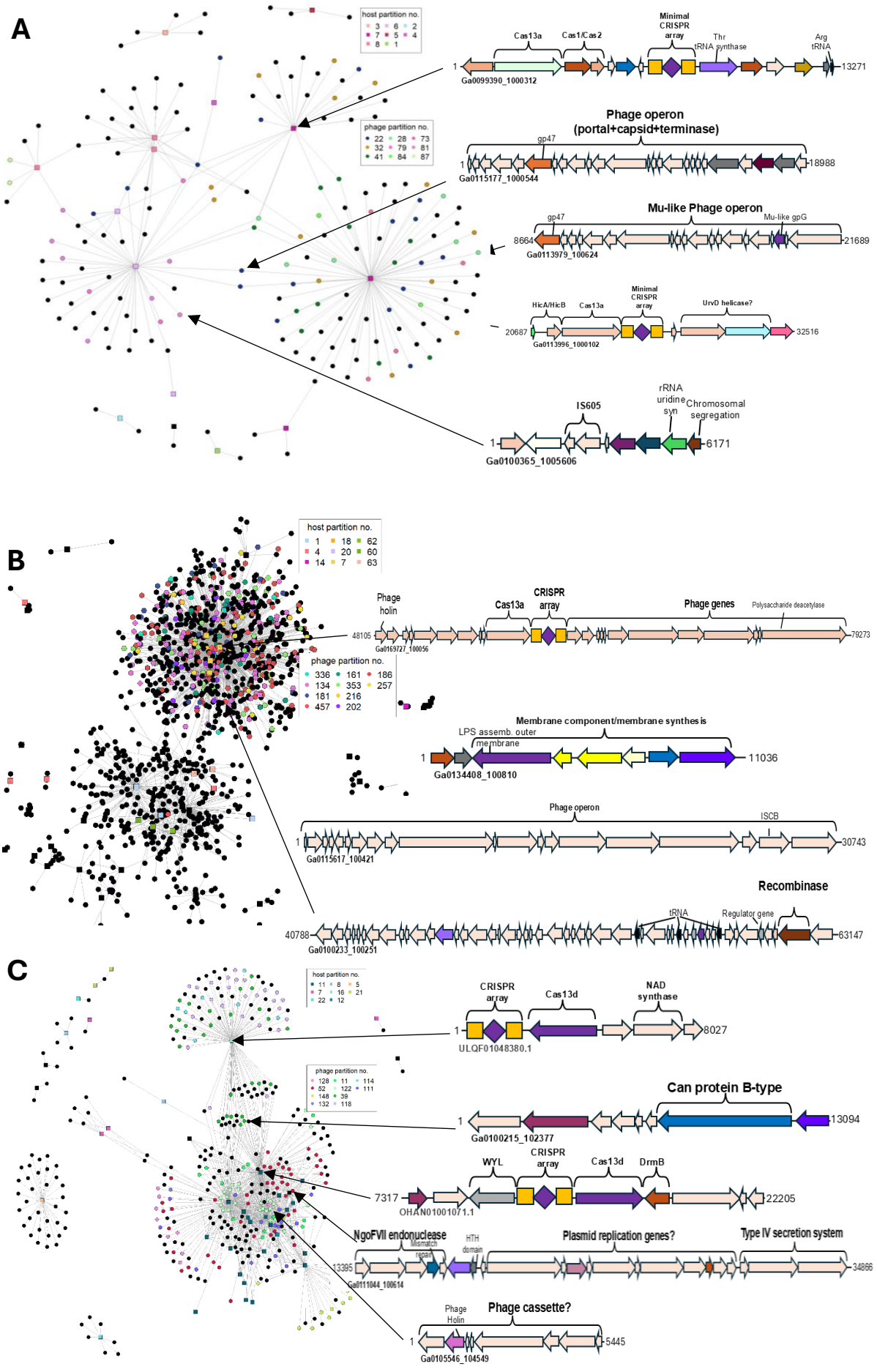

A
host partition no.
phage partition no.
Cas13a
Cas1/Cas2
Minimal CRISPR array
Thr tRNA synthase
Arg tRNA
Ga0099390_1000312
13271
Phage operon (portal+capsid+terminase)
gp47
Ga0115177_1000544
18988
Mu-like Phage operon
Mu-like gpG
8664
Ga0113979_100624
21689
HicA/HicB
Cas13a
Minimal CRISPR array
UvrD helicase?
20687
Ga0113996_1000102
32516
IS605
rRNA uridine syn
Chromosomal segregation
Ga0100365_1005606
6171
B
host partition no.
phage partition no.
Phage holin
CRISPR array
Cas13a
Phage genes
Polysaccharide deacetylase
48105
Ga0169727_100056
79273
Membrane component/membrane synthesis
LPS assemb. outer membrane
Ga0134408_100810
11036
Phage operon
ISCB
Ga0115617_100421
30743
Recombinase
tRNA
Regulator gene
40788
Ga0100233_100251
63147
C
host partition no.
phage partition no.
CRISPR array
Cas13d
NAD synthase
ULQF01048380.1
8027
Can protein B-type
Ga0100215_102377
13094
WYL
CRISPR array
Cas13d
DrmB
7317
OHAN01001071.1
22205
NgoFVII endonuclease
Mismatch repair
HTH domain
Plasmid replication genes?
Type IV secretion system
13395
Ga0111044_100614
34866
Phage Holin
Phage cassette?
Ga0105546_104549
5445

Figure 4.6: Bipartite interactions between CRISPR-Cas system subtype encoding contigs and their corresponding spacer mapped sequences. *Square nodes (□) represent Type VI-subtype encoded contigs. Circular nodes (○) represent mapped phage sequences. Only interactions between the top 10 largest host and mapped sequence partitioned clusters were included. Interaction networks were generated for the following subtypes and their mapped sequences: (A) Type VI-A, (B) Type VI-B, (C) Type VI-D. The gene structures of 3 samples from mapped-sequence clusters, and 2 CRISPR-array encoding contigs were drawn to visualize the gene-level differences between each cluster.*

Chapter 4.3: Discussion

4.3.1: Network based representations of type VI systems and spacer-mapped MGEs are effective at probing intra-subtype diversity

The inadequacy of current CRISPR-Cas taxonomic classification schemes to achieve consistency at the subtype level and below, is a known outstanding problem[90] that obscures an accurate understanding of the remaining diversity to be explored within orthologs of known systems. Previous attempts at more accurate tree of life from assembled sequence data found a substantial improvement when whole sequences were used instead of conserved genes, such as 16S ribosomal RNA[235]. Additionally, bipartite gene networks have previously been used effectively to differentiate and classify different families of viruses directly from assembled sequence data[151] . In this work, I appropriated these methodologies to probe the diversity of CRISPR systems with a specific focus on type VI CRISPR-Cas systems at the intra-subtype level. I used multigene-genome similarity networks generated by vConTACT2 to analyse the intra-subtype differences between type VI CRISPR-Cas systems. A comparison of this network to a traditional single-gene phylogeny made from the principal Cas13 ortholog (A-D) of each type VI CRISPR-Cas subtype investigated showed clearer segregation of local clusters in some subtypes compared to equivalent clades in a tree representation affirming the method's suitability for differentiating groupings of sub-subtype level

sequences. To annotate sequences of each CRISPR-Cas subtype within the network, I constructed a workflow designed to annotate local clusters within each vConTACT2 network. This workflow analysed the gene composition as well as the plasmid and phage propensities of each sequence. From this process, I discovered a large number of co-encoded antiphage defence related genes specific to each cluster. An analogous approach was performed to explore the MGE and virosphere of mapped sequences of each Type VI-subtype, which revealed the conserved genes present both within and across each local cluster, a subset of which were also homologous to antiphage defence genes. Spacer mapped sequences of local clusters in each subtype were extremely diverse, even in samples with strong taxa biases resulting from low sample diversity. Although not highly prevalent among the mapped phage targets, several genes of special interest were detected in the mapped sequences of type VI subtypes. Foremost among these were IS605 and ISCBs orthologs in the mapped phage sequences of type VI-A and VI-B subtypes. These proteins may interact with Type VI systems via host-phage interactions in an as yet undetermined manner.

4.3.2: Probing the intra-subtype diversity of type VI CRISPR-Cas systems showed few conserved genes between local clusters

Compared to traditional phylogenetics, network analysis using vConTACT2 was effective at visualising local-cluster specific islands of co-encoded defence genes, but less effective at identifying shared genes between clusters. The detection of different orthologs of antiphage defence genes between orthologs was consistent with previous observations that the composition and number of antiphage defences varies significantly both within and between closely related species[23,86]. The inability of vConTACT2 to establish edges between distant clusters may have partly been a consequence of the default cutoff utilised by vConTACT2 to identify sequence similarity between genes on contigs being higher than would be utilised when performing sequence-searches using a traditional phylogeny. This issue may be remedied by relaxing this threshold. Nevertheless, analysing the gene composition of each local cluster revealed very few shared genes between local clusters delineated by cluster partitioning. In all CRISPR-Cas type VI subtypes investigated, only the main effector

protein used for subtype identification was conserved across all local clusters in the network.

Among those genes conserved in more than one cluster within each subtype, HicA-HicB, Csx27/Csx28 and WYL were the most notable in type VI-A, VI-B and VI-D systems respectively. Of these, only WYL proteins were conserved in a majority of local clusters in the respective subtype in which it occurred (Type VI-D). It is unknown why WYL was so disproportionally conserved with these systems, although the relatively narrow host-range among which Type VI-D systems were found (predominantly *Ruminoccoccus*) means that this could not be ruled out as a genus or species-specific effect. Given that WYL-domain containing proteins have been previously shown to function as both activators transcriptional repressors [88,214,220,221] it is possible that other clades within type VI systems use an alternate protein for gene regulation. This protein may be trans encoded. Alternatively, it is possible that the activity enhancement observed in Type VI-D by the expression of WYL is not a conserved behaviour of WYL proteins encoded in other type VI subtypes, which may be repressed by their respective WYL proteins instead. Within Type VI-D systems, a differentiation of WYL domain protein functionality is supported by prior phylogenetic analysis showing that some clades of Type VI CRISPR WYL proteins contain conserved single or multiple fused ωHTH or RHH domains, while others consist of unknown protein domain fusions[88]. These have been shown to have different binding preferences for dsDNA and ssRNA binding[236]. This supports the prediction that the activity of these proteins differs between orthologous systems.

Acquisition proteins were detected in more than one local cluster in each subtype but were not a conserved feature. It has been shown that acquisition genes may be repurposed from co-encoded CRISPR-Cas systems to facilitate spacer adaptation in type VI systems[140,237]. While the presence of these systems encoded in *trans* outside the bounds of contigs can't be excluded, co-encoded CRISPR-Cas systems from other types were not detected in most local clusters. This suggests co-option of other CRISPR systems by type VI subtypes for adaptation is not universal.

4.3.3: The mapped sequence landscape of type VI-systems reveals a significant number of shared genes between local clusters

To date the nature of the mapped targets of type VI-CRISPR Cas systems has not been comprehensively surveyed. There was strong evidence among all CRISPR-subtypes for a preference for phage and plasmic targets. Interestingly, weak edges between clusters existed between the mapped sequences of each subtype. This indicated that at least some homologous genes were shared between clusters, which was verified by an analysis of the gene content of different phage clusters which showed greater intra-cluster overlap compared with equivalent networks of host-encoded subtypes. While the gene composition of individual clusters was too divergent to suggest any evolutionary relatedness between clusters, these linkages may indicate some level of gene exchange. Some of the genes conserved across more than one cluster included nucleases such as *HNH* and *TerL*, which have both previously been shown to be involved in the packaging of phage DNA into capsids prior to export [238-241]. A significant number of detected genes in certain clusters predicted to be circular contained homologs of Type IV secretion or other conjugal transfer systems, which was further evidence that a significant subset of spacers conferred immunity against plasmids in addition to phages. Overall, the target preferences of Type VI spacers bore a strong similarity to existing type I-V subtypes which was striking given the different substrate specificity and mechanism compared with DNA-targeting effectors, as well as the dramatically lower overall conservation of the acquisition genes within each subtype. A snapshot of representatives from each CRISPR-Cas subtype clusters, and clusters of their corresponding mapped sequences revealed several encoded features of special interest. One of the main local clusters of Cas13b-orthologs was phage-encoded itself and possessed spacers which targeted phages and especially plasmids, defying the naïve assumption that phage co-opted CRISPR-Cas systems instead target host genome loci. Although no previous study systematically analysed the spacer target preferences of phage encoded type VI-B CRISPR systems, these target preferences were similar to previous spacer mapping investigations concerning plasmid encoded CRISPR-Cas subtypes, which found that a majority of plasmid encoded CRISPR-Cas systems target mainly phages and other plasmids, as opposed to chromosomal loci[142,144,242]. This has been shown via modelling to confer protection against both the

plasmid and host-cell under most parameters tested [243]. Based on this, it could be postulated that a set of conditions under which accessory defence genes encoded in phages instead of plasmids provide similar benefits to the host-cell.

4.3.4: Limitations of utilising network-based interrogation of intra-subtype diversity in type VI systems and their corresponding mapped sequences

Despite its utility in analysing the joint host subtype – mapped MGE landscape, there were several shortcomings to my approach which limit the applicability of these findings. Differentiated clades of sequences from each subtype reflected the underlying substructural diversity but using a lower threshold when generating the networks may have been more sensitive in establishing additional edges between distantly related clusters. Oversampling of certain clusters may have overestimated some of the conservation scores used to compute the relative abundance of ORFs in each cluster. While diversity indexes were able to reveal the extent to which these sequences were sampled from different environments, these scores likely do not reflect the real abundance of these genes in the environments from which samples were taken. Because the identification of each ORF was performed by remote sequence similarity to Pfam or DEFLOC HMM profiles, false positive rates were high in instances where remote homology scores were relatively low (e-value > 10^{-10}). This could confuse the correct annotation and assignment of unknown ORFs to Pfam/DEFLOC profiles but did not alter the observation that detectable similarity to Pfam/DEFLOC derived domains existed for these genes.

4.3.5: Summary of findings

Overall, my findings from this chapter demonstrated the value of analysing partitioned local clusters within networks as a means to explore intra-subtype evolutionary diversity. When investigating type VI systems using this technique, several notable intra-subtype features were revealed, such as phage encoded type VI-B systems, and host-MGE interactions between type VI systems and mapped targets comprising a mixture of phage and plasmids. These findings established the intra-subtype gene cluster diversity

different of type VI CRISPR-Cas systems, as well as the repertoire of MGEs targeted by their respective spacers.

To link these differences in both aggregate and intra-subtype diversity to the functional differences in acquisition and interference across CRISPR-Cas subtypes, the next chapter harnesses the large-scale spacer mapping undertaken in this chapter to perform an analysis of spacer acquisition biases in a range of CRISPR-Cas subtypes. This includes some subtypes in which a bias has not previously been reported. This enabled the prediction of novel acquisition biases in some systems, providing an additional layer of functional information regarding acquisition and interference at the subtype level which can't be inferred by gene-genome network analysis alone.

Chapter 5: Characterisation of spacer mapping patterns across CRISPR-Cas systems

5.1: Background

5.1.1 General mechanisms for priming effects during CRISPR-spacer acquisition

Since the first mapping of spacers to phage genomes identified the *raison d'etre* for CRISPR-immunity[63,198,244], the exact mechanism by which many spacers are acquired by CRISPR-arrays remains a key unknown among many CRISPR-Cas subtypes. Spacer acquisition is often not random but is predisposed toward certain sites as a result of "priming" processes or indirect acquisition biases. Primed spacer acquisition is distinguished from an acquisition bias by the presence of a priming protospacer (PPS), which directs the subsequent acquisition of more spacers by effector proteins during interference, which are produced as by-product of guided-RNA degradation of a target site. This phenomenon is a key feature of type I CRISPR-Cas immunity, where interference by the CASCADE complexes and degradation by Cas3 produces ssDNA fragments which serve as pre-spacers for further acquisition[127,131,165,166,174,245]. It has also been demonstrated that a bias in spacer acquisition can occur indirectly through the production of ssDNA fragments during RecBCD mediated repair of double stranded DNA breaks (DSBs) by homologous recombination. These breaks can be produced naturally [179] or by nuclease cleavage, as has been observed in Type II-A *S. pyogenes* Cas9 [246]. However, the existence of acquisition biases in CRISPR-Cas subtypes outside of type I and II CRISPR-Cas systems has been less comprehensively surveyed. The highly conserved nature and the essentiality of DNA repair across all prokaryotes, suggests that acquisition biases may be observed in organisms encoding other CRISPR-Cas systems.

5.1.2 Computational surveying of spacer acquisition biases profiles reveals the extent of their conservation across CRISPR-Cas subtypes

Although several works have experimentally characterised spacer acquisition bias in *E. coli* by type I, II, and V CRISPR systems[127,166,170,173,174,179,181,246,247], complementary computational characterisation of these biases across distantly related taxa encoding the same CRISPR-Cas system is required to generalise these findings to all clades. Furthermore, no acquisition biases have been observed to date in RNA-targeting systems (type III, VI) apart from requiring complementarity to the template (transcribed) strand [248]. A recent *in silico* method called "spacer distribution analysis" was developed[63,198,244] to estimate biases in the acquisition of secondary spacers as a result of an ancestral priming spacer [175]. This method relies on the detection of two spacers from the same CRISPR array and assumes that spacer integration predominantly occurs at the leader-repeat junction[64,74,249] . Therefore, the oldest integrated spacers are hypothesised to be further away from the leader-repeat junction, while newer spacers are encoded closer to the junction. The oldest spacer with a mapped phage target is termed the *priming protospacer* (PPS). Spacers can map in two-by-two possible directions with respect to the PPS, mapped sites being either upstream or downstream, and in either the sense or antisense direction (Figure 5.1). This generates a fingerprint, which allows the discernment of different priming activities at both the clade and subtype level.

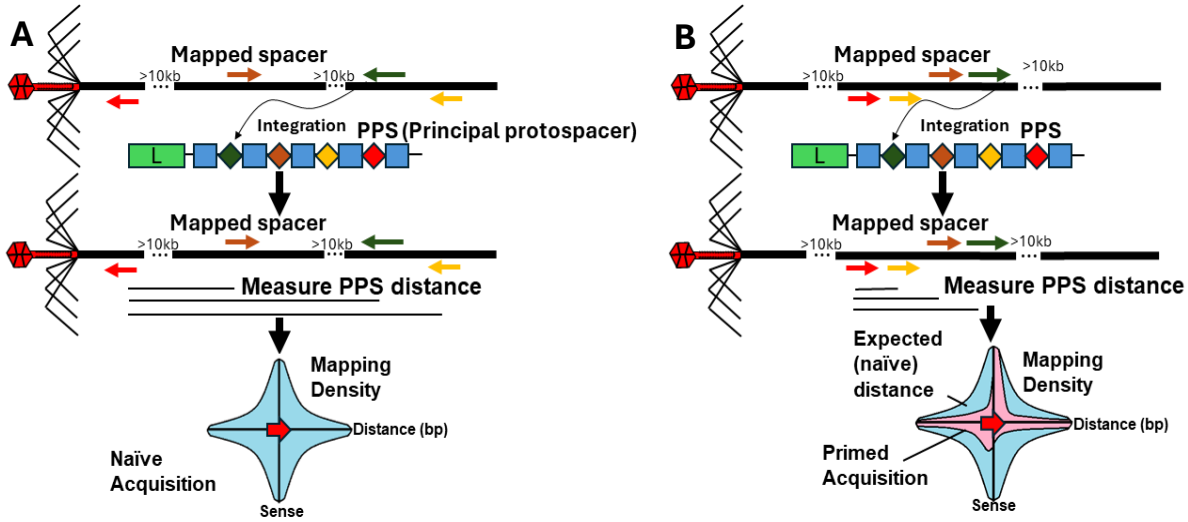


Figure 5.1: ***Visualising naïve vs. primed adaptation using spacer distribution analysis.*** *In cases where two or more spacers map to the same contig, the spacer-match which is encoded furthest from the leader sequence (L) in the host CRISPR-array is assigned as the Priming Protospacer (PPS). Pairwise distances are then computed between the PPS and other mapped spacers. In naïve acquisition, (A) distribution of distances with respect to the PPS is a function of contig size while in primed acquisition, (B) spacers are preferentially acquired at short distances from the PPS.*

There are two limitations in scaling-up the spacer distribution analysis technique in its current form. Foremost among these is the requirement for at least two mapped spacers to the same phage, which was shown to occur in less than 10% of mapped sequences [175]. This meant that when spacer distribution analysis was applied to common CRISPR-Cas subtypes such as Type-IIA, the statistical power was too low to detect a significant priming effect. A follow up study experimentally confirmed the presence of priming in type II-A system [246]. Increasing the sensitivity of this technique among rarer CRISPR-Cas subtypes requires detecting a greater number of mapped spacer pairs, especially to the same phage.

5.1.3: Potential improvements to existing bioinformatic estimations of spacer acquisition biases

One means of enhancing the sensitivity of the approach is to increase the database size used for mapping spacers. As a small-scale proof-of-concept, the source dataset of phage genomes and DNA fragments utilised for mapping was the IMG-VR_2018 [250]. At

approximately 750,000 contigs, this represented only a small fraction of the abundance of mapped phage sequences present as uncultivated sequences in common environments. Although subsequent versions of the IMG-VR database has dramatically increased in size by more than one order of magnitude since 2018, the number of phage sequences (approximately 15 million [251]) remains dwarfed by the presence of phage sequence fragments in metagenome samples, most of which are too small to make a positive phage determination using the IMG phage annotation pipeline [251] or other pipelines. I hypothesised that expanding the database size to increase the number of mapped phage sites would increase the sensitivity of the approach.

Previous work has established that escape mutations to phages develop over extremely short timescales [252]. Exact matches to two or more sites on the same phage are expected to be rare because of the low probability of sequencing the phage before an escape mutation has occurred in either the PPS or the spacers acquired from priming. These phage-host matches may thus not be representative of normal primed spacer acquisition events in response to phage infections. Searching for partial spacer matches whose target sites have evolved escape mutations, which I denote as "degraded spacer matches," would increase the sample size and determine if degraded spacers have the same mapped spacer distributions as whole matches.

I employed both an expanded metagenome-based search for spacer mapped phages, as well as mapping of degraded spacer target sites to increase the sensitivity of the approach and its statistical power. Using the phage and plasmid discovery pipeline established in Chapter 4, I expanded the phage/plasmid database to perform spacer distribution analysis with a higher degree of resolution compared with using GenBank and VR/IMG sequences alone. In conjunction with developing a kmer-based search approach to apply spacer distribution analysis on degraded target sites, I attempted to address whether primed spacer acquisition is a ubiquitous phenomenon throughout different CRISPR-Cas subtypes. I found evidence that spacer acquisition biases appear to be a conserved phenomenon across and within many CRISPR-Cas subtypes. However, although using degraded spacer target sites increased the sensitivity of spacer acquisition bias detection, the absence of strand directionality revealed these sites to be less effective at determining the strand direction over which these biases

were occurring. Overall, I demonstrated that primed spacer acquisition may occur in more subtypes than currently known and have improved the sensitivity of detecting spacer acquisition biases, enabling detection when the number of complete target site matches is too low.

Chapter 5.2: Results

5.2.1: Workflow to map and measure the acquisition bias of CRISPR-Cas subtypes

To detect spacer acquisition biases at the aggregate CRISPR-Cas subtype I analysed the distances between the target sites of spacers mapped to the core (10TB) block of aggregated assembled metagenomic and NGS data used throughout all 3 Chapters of my investigation. A set of 20kb windows encoding the corresponding CRISPR-Cas system operon for 13 different CRISPR-Cas subtypes were extracted via sequence similarity searches by the presence of a conserved signature gene, which was also used for subtype identification in Chapter 4 (see table S4.1). The operons from each subtype were then examined for overlap to other subtypes to ensure the effectors in each operon were classified to a single subtype. This ensured that the CRISPR-Cas subtypes used for spacer mapping and acquisition bias analysis were only from just one subtype. Contigs returning ambiguous classifications when attempting to subtype the interference module were excluded from the analysis. The spacers from the corresponding CRISPR-arrays in each subtype were then extracted and used for spacer mapping. Mapping was performed using the spacers to identify target sequences within the 10TB data block (Figure 5.2A). For each spacer sequence, 7bp handles from the direct repeats upstream and downstream of each spacer were concatenated to each spacer to eliminate spacers mapped to their own CRISPR arrays. These sequences were concurrently used as queries to search the 10TB. Cases where both spacer sequences and these concatenated DR-spacer handles mapped to the same target sites with the same identity and coverage thresholds were excluded, as these were hits to CRISPR-arrays. My prior investigation of mapped target sequences (Chapter 4.3) provided a genomic context for the subtypes analysed in this work, labelling different groupings of the mapped targets as plasmid, ssDNAphage, dsDNAphage or uncharacterised.

Mapped target sequences were then filtered to retain only matches with more than two spacers matches to the same target sequence. Pairwise distances between the PPS and each of the other mapped spacers were then calculated. Because of the redundant nature of the assemblies deposited in Genbank and JGI repositories from which this data was derived, extensive deduplication of PPS-spacer pairs was required to prevent artefacts. Two general classes of PPS-spacer redundancy were observed as either two identical spacers mapping to the same target sequence at the same position, or a spacer mapping to two identical target sequences (Figure 5.2B). In contrast to previous implementations of PPS-spacer pair deduplication[175], I employed a single unified approach to deduplicate both classes simultaneously. First, the calculated distances for each PPS-spacer pair were pooled together. Each distance was then used to form a non-redundant set of distances for each mapped phage (within 5bp). This ensured that the distribution of PPS-spacer distances were due to biases in acquisition rather than artefacts from sequence redundancy.

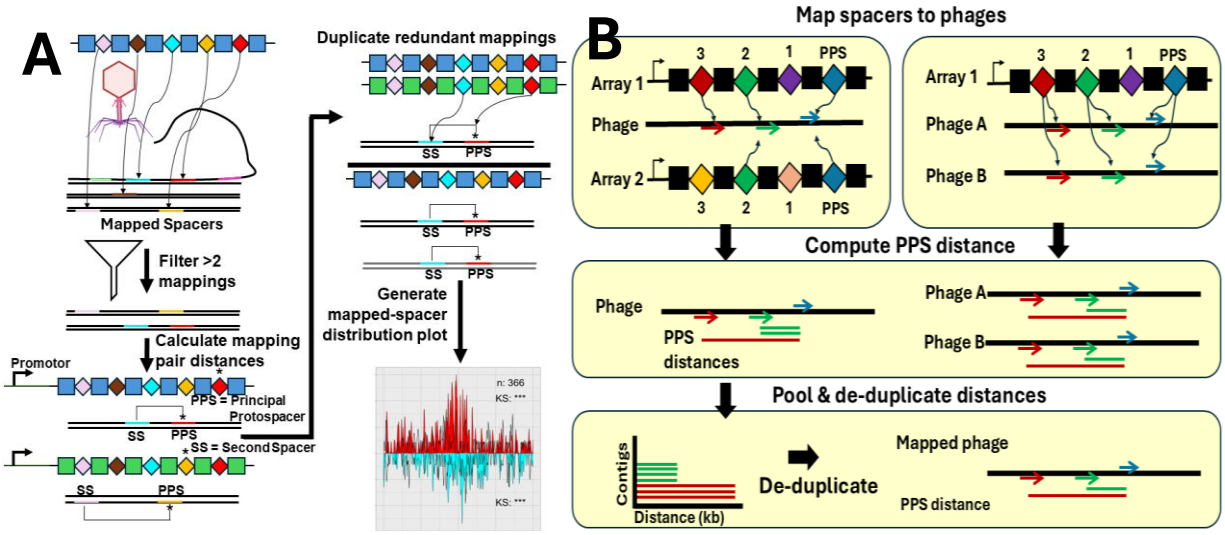
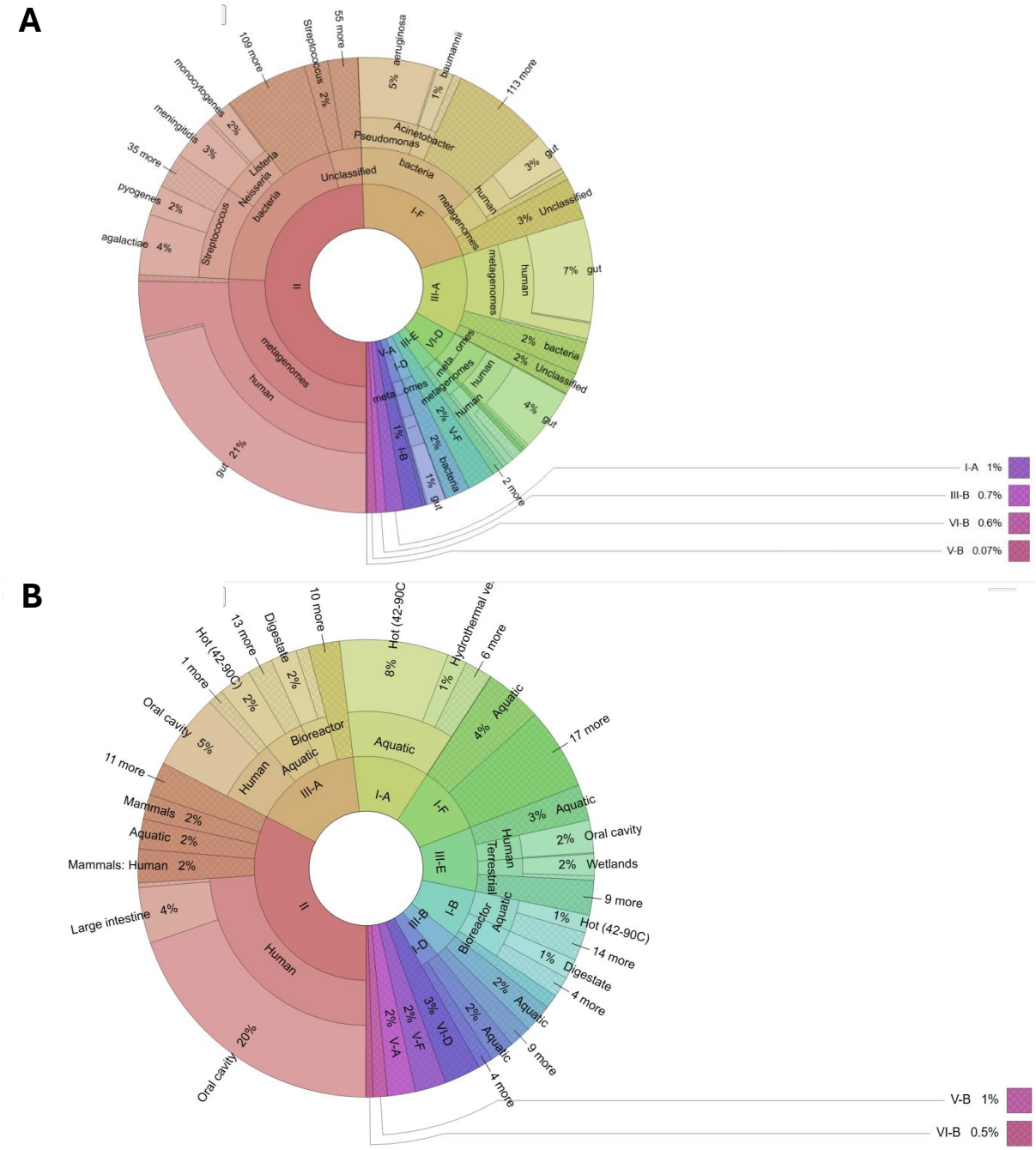


Figure 5.2: Computation of spacer acquisition biases from mapped spacers to metagenome sequences. *(A) Workflow used to generate the input data for computing the aggregated distances of mapped spacers from the priming protospacers, in cases of two or more mapped sites per target sequence. Mapped spacers to their own arrays were then filtered and distances scores to the PPS calculated. (B) Schema for deduplicating both host-encoded spacers and mapped phage target sites. There were two types of instances of redundant spacer matches: those stemming from duplicate CRISPR-arrays and duplicate target sequences. The computed distances for both were*

then filtered such that only a single distance was allowed for each pooled grouping of arrays and target sequences.

5.2.2: Census of CRISPR-Cas subtype host input data

To assess whether host-derived spacers were from a small subset of organisms, or broadly representative of multiple clades, the taxonomy of the sequences encoding the 13 CRISPR-Cas subtypes with which spacer mapping was performed were analysed. For all subtypes investigated (Figure 5.3), the subset of host sequences encoding CRISPR arrays possessing two or more spacers with matches to the MGE were derived from a diverse array of environments and species. There was a bias towards uncultivated (metagenomic) human gut microbiota from Genbank deposited sequences. This may have been because NCBI contigs were longer on average, less fragmented and more likely to contain two or more mapped sites compared with JGI sequencing data. This effect was only observed in type V-A, type V-F1, and type I-B (Figure 5.3A). In type III-A systems an enrichment of two or more mapped sites in JGI data (Figure 5.3A) was observed compared to the original proportion of JGI data forming the subtype was observed (Figure S5.1). The taxonomic representation of sequences in these subtypes defined the scope within which the subsequent findings from spacer distribution analysis were applicable.



<u>Figure 5.3:</u> Census of CRISPR-Cas subtype host genomes after the elimination of self-target spacers and duplicate mapped sites/genomes. *Uncultivated sequence data from (A) NCBI and (B) JGI was hierarchically classified by subtype, location and sample. NCBI Sequences derived from cultivated microbes were labelled by taxa.*

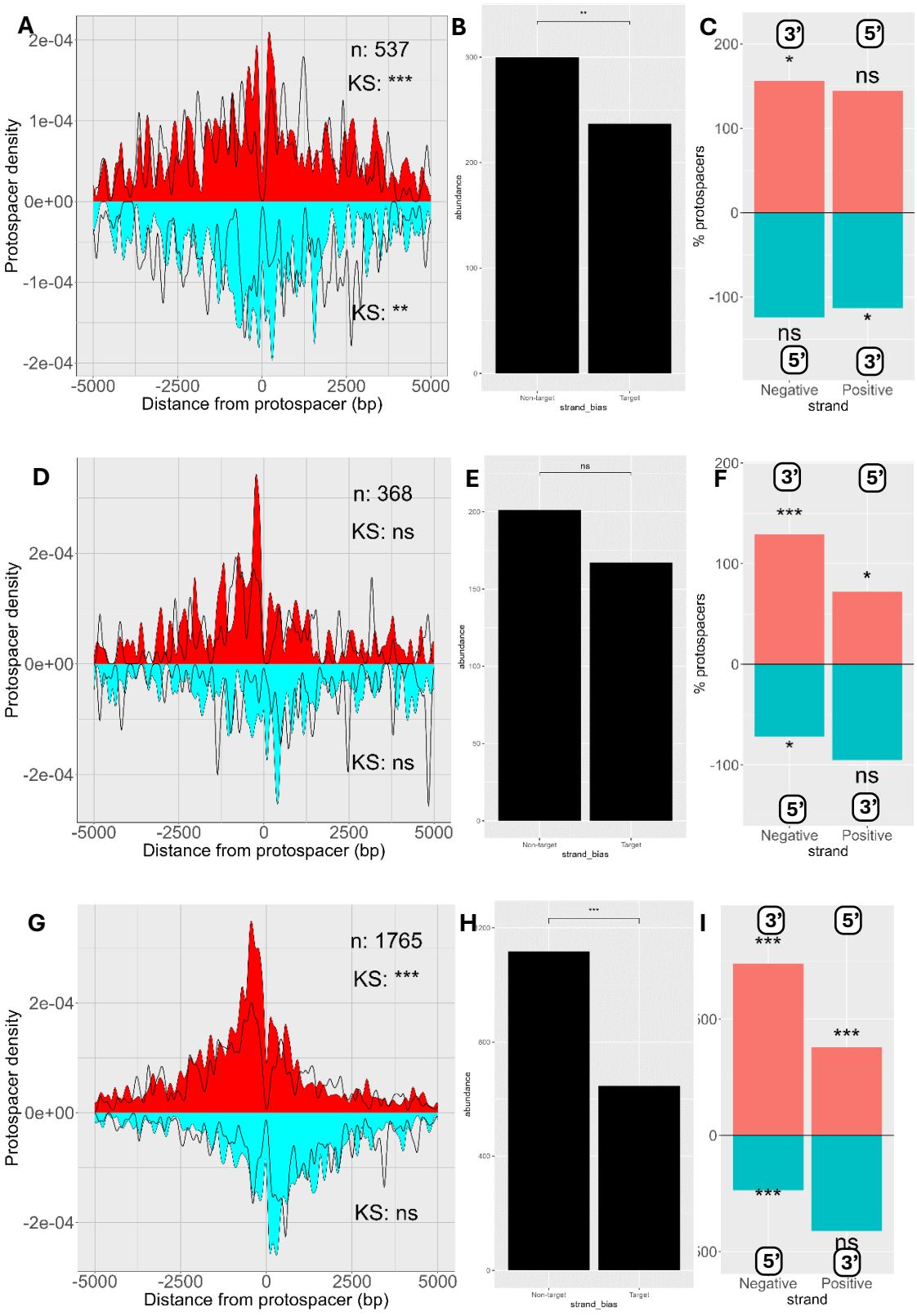
5.2.3: Detection of spacer acquisition biases in type I, II, III, V and VI CRISPR-Cas subtypes

To assess whether spacer acquisition biases were widespread across different CRISPR-Cas subtypes, I expanded the range of subtypes in which spacer distribution analysis has been performed. An acquisition bias was measured by whether the PPS-spacer distances for each subtype, scaled for the length of the contig, differed significantly from a [0,1] uniform distribution. The probability of two spacers mapping at any given contig at random was purely a function of the size of the contig. When scaled by the length of the contig, the probability of finding any distance between two spacers was constant (uniform). An observed depression in these values compared to a uniform distribution thus indicated that the distances between the two spacers was smaller than would be expected by chance. I next reasoned that the population of expected distances of two spacers drawn at random on a contig of arbitrary size formed a lower triangular distribution, can be compared to a uniform distribution after transformation of the scaled spacers to a cumulative distribution. To increase the sensitivity of this approach compared with a past investigation[175], I used a larger metagenome sequence dataset to increase the number of measured PPS-spacer distances.

Type I systems:

I initially investigated spacer acquisition biases in type I systems. Spacer acquisition biases in these systems are well established in type I-B and I-F systems[175]. Applying my own approach to these subtypes thus functioned as a positive control and affirmed that my methodology was effective and consistent with these previously known results as well as characterising the biases in type I-A where the spacer distribution was not previously known. Given that primed spacer acquisition is well characterised in type I-B, I-E, and I-F systems, I sought to determine whether evidence for primed spacer acquisition also existed in rarer type I systems. Spacer distribution analysis was performed for type I-A, I-D, I-B and I-F systems (Figure 5.4). Several subtypes, such as Type I-C and Type I-E were not characterised as some homologs of the effector proteins from the retrieved set of contigs were found to overlap. Type I-U and other subtypes known to encode promoters within the direct repeats of their CRISPR-arrays violated the underlying assumptions by which an acquisition bias was measured and were also

not tested. A clear acquisition bias was observed in type I-A systems (Figure 5.4A). Acquisition was more common on the non-target compared to the target strand (Figure 5.4B). However, a weak albeit still significant directional bias was observed in the left-non-target and right target quadrants. In type I-D systems no acquisition bias was detected. However, this may have been due to a lack of statistical power ($1-\beta < 0.8$) (Table S5.1). Compared with type I-B and type I-F systems (Figure 5.4G,5.4J) the acquisition bias observed in type I-A systems was had a different topology from either subtype. Acquisition in type I-B systems was strongly biased toward acquisition in the left non-target quadrant (Figure 5.4G) while acquisition in type I-F was biased towards the right-target and left-non-target directions (Figure 5.4L). This directionality was notably unique among all systems tested, which may reflect the unique primed spacer acquisition mechanism of type I-F systems [127]. Taken together my results show that significant divergences exist in biased spacer acquisition in type I systems, which may reflect different mechanisms of spacer acquisition.


A
n: 537
KS: ***
KS: **
Protospacer density
Distance from protospacer (bp)
B
abundance
Non-target
Target
strand_bias
C
% protospacers
Negative
Positive
strand
D
n: 368
KS: ns
KS: ns
E
F
G
n: 1765
KS: ***
KS: ns
H
I

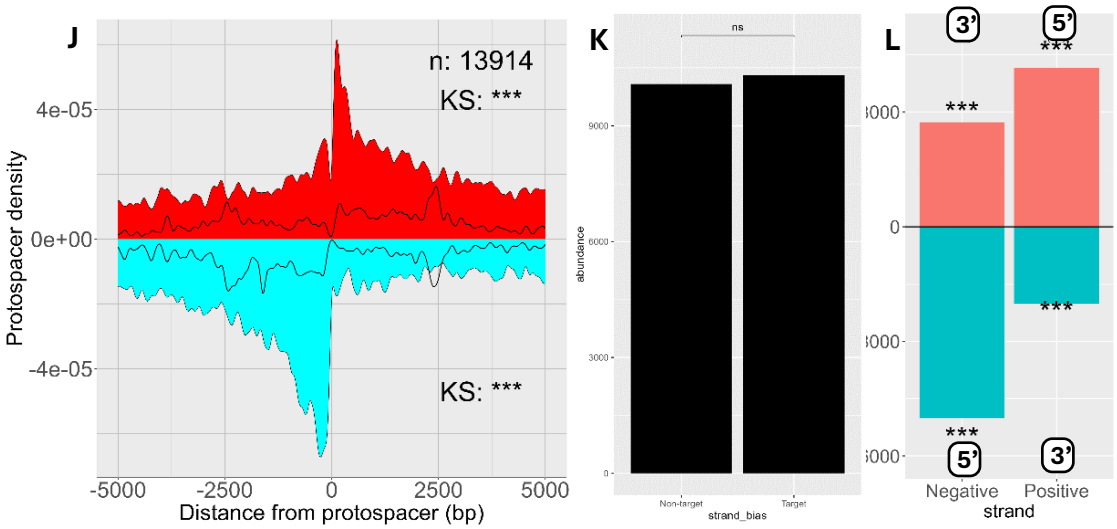
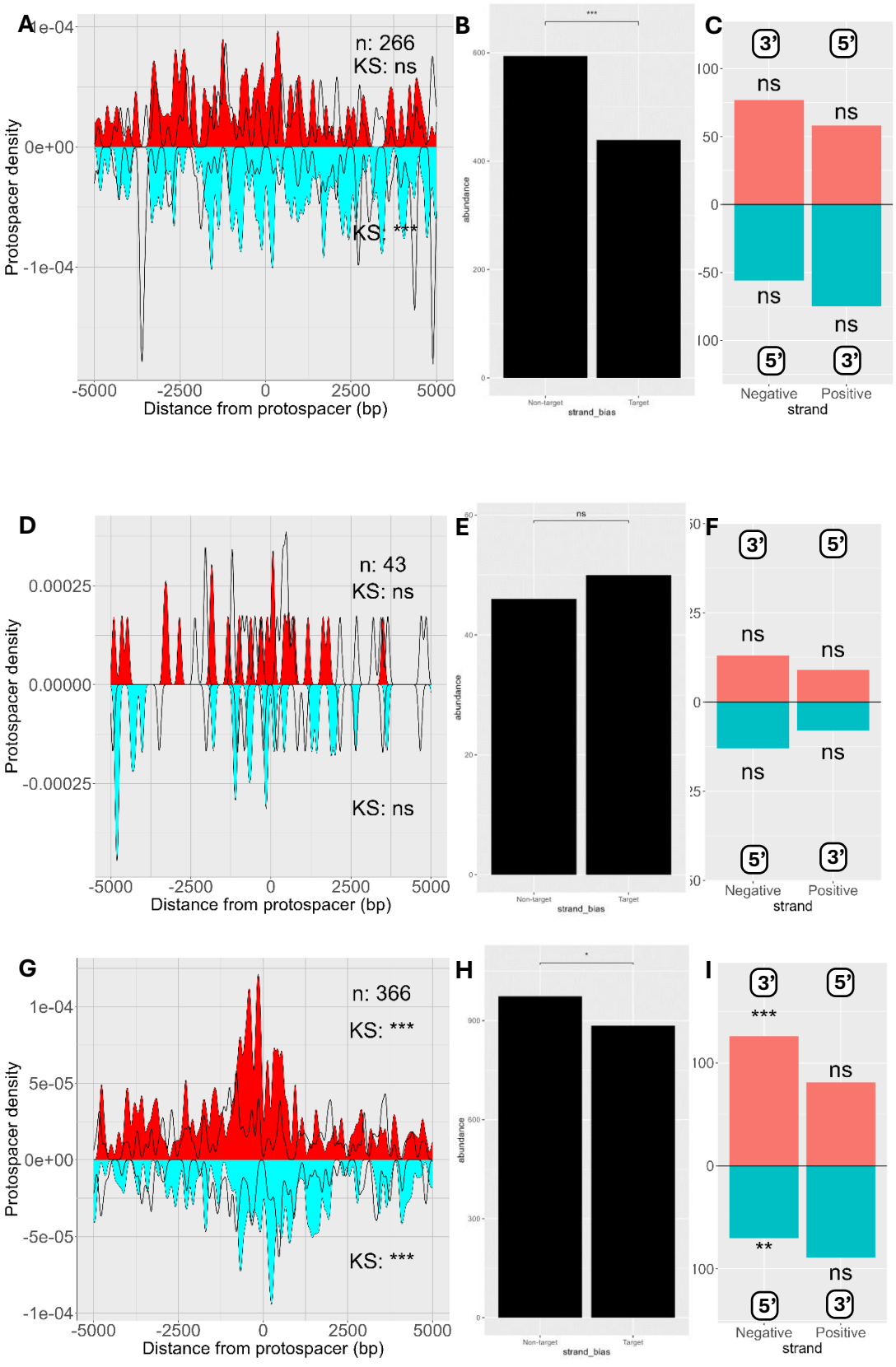


Figure 5.4: Expected vs. mapped spacer distances in Type I CRISPR-Cas over a 5kb interval. *The spacer distribution diagrams (left side) were divided into 4 quadrants, based on whether the mapped spacer was upstream or downstream of the PPS (left-right) or on the same strand identical to the gRNA (non-target strand) or the complementary strand (target strand). Protospacer densities from the PPS at the origin were estimated using a Kernel Density Estimation (KDE). Negative density values (cyan) mean that mapping occurred on the non-target strand compared to the target strand (red), which is defined as the sense direction of a bound gRNA+DNA heteroduplex encoding a spacer. A bias toward the target and non-target strand (center), as well as each quadrant (right side) was determined by estimating the binomial probability of a spacer falling within one quadrant compared to the mean of all quadrants. This enabled the observation of the direction of biased spacer acquisition with respect to the PPS. Plots were generated for each Type I system investigated: (A) Type I-A, (D) Type I-D, (G) Type I-B, (J) Type I-F. The significance of any observed strand bias in the abundance of mapping densities was estimated for each subtype using binomial probability (B,E,H,K). This technique was further extended to measure strand biases in each quadrant (C, F, I, L). Note: sense direction was assigned based on the target strand.*

Class 2 systems:

To further characterise spacer acquisition biases, I expanded my investigation to Class 2 CRISPR-Cas subtypes. Previous works have demonstrated that Cas9-induced cleavage and natural DSB formation breaks induce acquisition biases from RecBCD-based resection of Double strand breaks [175,179,181,246]. This implies that some type V-

systems, which produce DSBs, might also produce acquisition biases. To assess this postulate, I conducted spacer distribution analysis on Type V-A, V-B, V-F1 and II-A systems (Figure 5.5). An acquisition bias was observed on the non-target strand of type V-A systems (Figure 5.5A, 5.5B). In contrast, no acquisition bias on the non-target strand was observed in type V-B systems due to a lack of statistical power (1-beta < 0.8) (Figure 5.5D, 5.5E). In type V-A and V-B systems, no quad-based preference for acquisition was observed (Figure 5.5C, 5.5F). Interestingly, a strand directional bias was observed in Type V-F1 systems which was absent from V-A and V-B (Figure 5.5G-I). This suggests a possible difference in the underlying mechanism of biased acquisition. This was also different from type II systems (Figure 5.5J-L), which displayed a very strong acquisition bias, but no directionality. These trends were also observed with a wider observation window up to 50kb (Figure S5.2). This raises the possibility that spacer acquisition in type V-F1 systems functions differently from other Class 2 systems which display a significant, but non-directional bias in the acquisition of new spacers.


A
B
C
D
E
F
G
H
I
n: 266
KS: ns
KS: ***
n: 43
KS: ns
KS: ns
n: 366
KS: ***
KS: ***
Protospacer density
Distance from protospacer (bp)
abundance
strand_bias
Non-target
Target
3'
5'
ns
***
**
Negative
Positive
strand

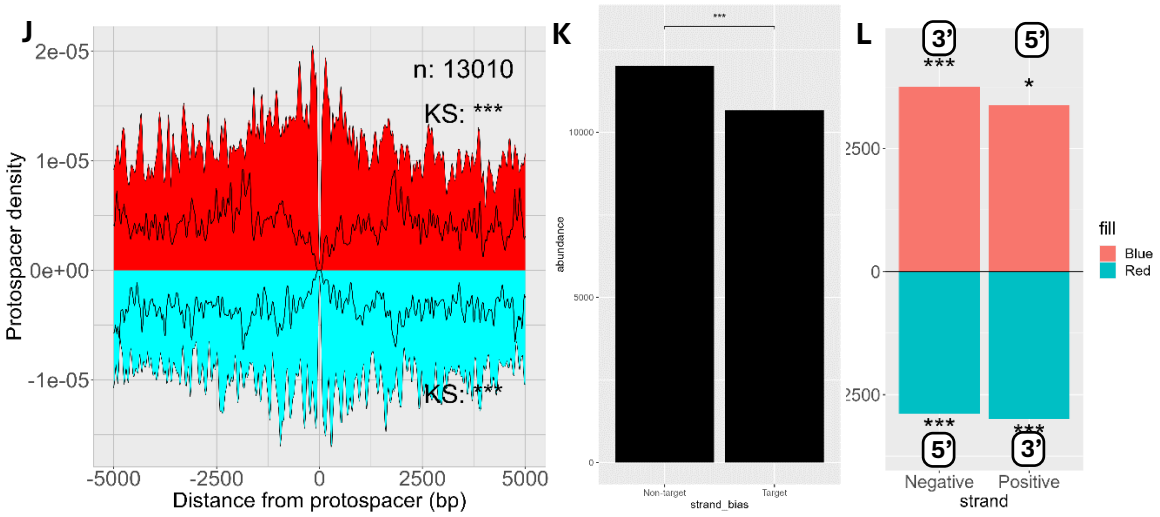


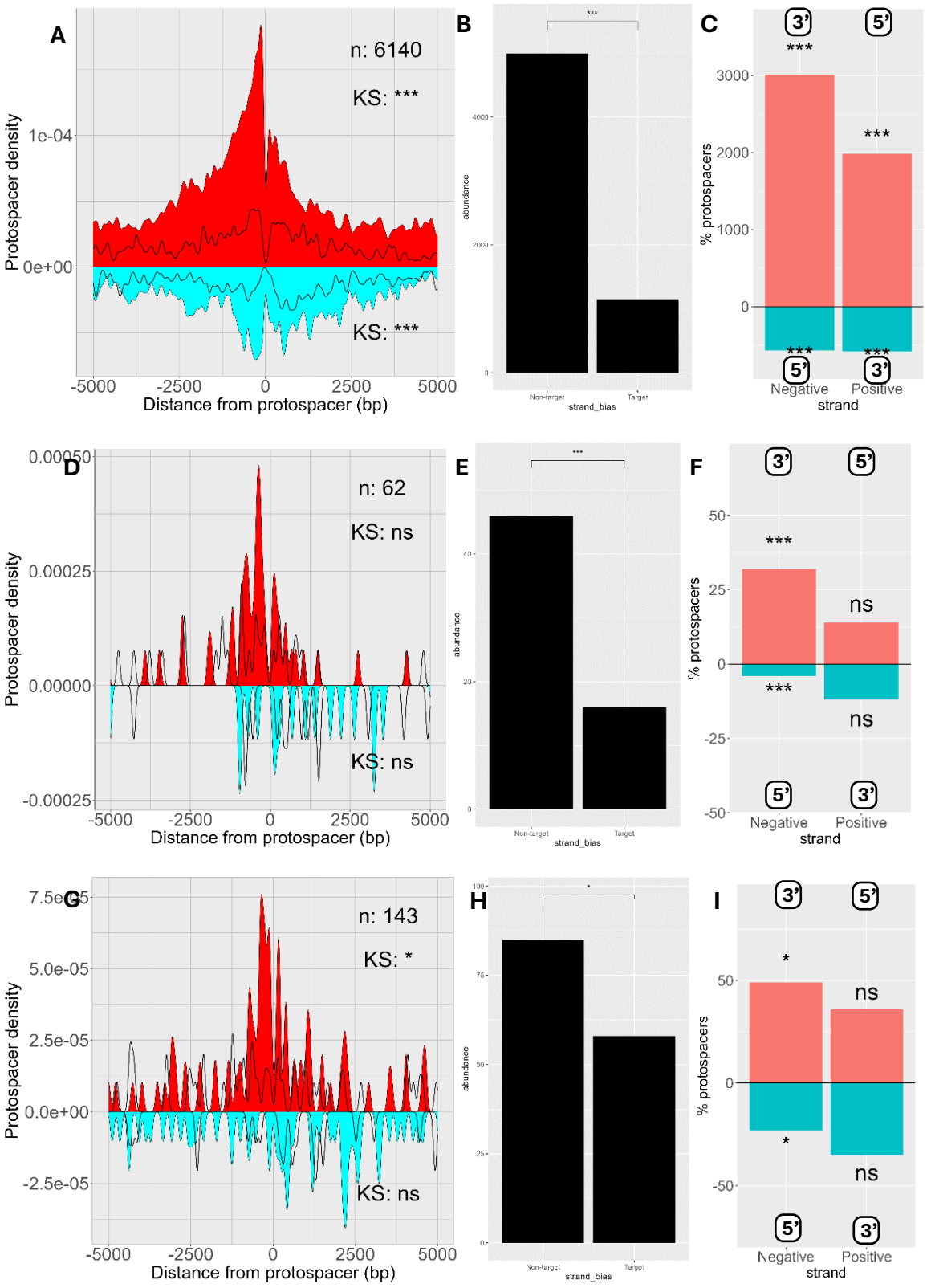
Figure 5.5: Expected vs. mapped spacer distances in Type II and V CRISPR-Cas subtypes over a 5kb interval. *Plots were generated using the same procedure as per Figure 5.4, for each class 2 DNA targeting CRISPR-Cas system for whom an acquisition bias was investigated: (A-C) Type V-A, (D-F) Type V-B, (G-I) Type V-F1, (J-L) Type II. In each subtype, a spacer distribution diagram (left side), binomial probability estimation of target/non-target strand bias (centre) and quadrant specific strand bias (right side) is shown.*

RNA targeting systems:

After uncovering spacer acquisition biases in CRISPR-Cas subtypes which predominantly target DNA, I hypothesised that similar biases might exist in subtypes which primarily target RNA. To test this, I applied the same technique to measure the distance between mapped spacers and the presumptive PPS. Importantly however, because both type III and VI systems target RNA, and the data was only the equivalent DNA coding sequence, this method could not take into account disparities in the fraction and quantity of a given ORF which was transcribed. However, the directionality of adaptation with respect to the PPS could still be measured accurately. I observed a stark contrast between these types in terms of the spacer acquisition pattern in type VI systems compared to type III (Figure 5.6). Type III-A systems displayed a strong acquisition bias in the left-non-target strand < 500bp from the PPS (Figure 5.6A-C). This was the same sense direction as the transcribed RNA strand. This was consistent in type III-A regardless of whether the main acquisition protein included a reverse transcriptase (Figure S5.3), suggesting that RT and RNA itself may be dispensable for

the observation of this pattern. To test whether an acquisition bias was still observed when both spacers were mapped to the same gene, I calculated a spacer mapping distribution for each RNA-targeting subtype wherein I took the subset of mapped-spacer pairs which only match to a single gene (Figure S5.4). This revealed a significant acquisition bias for type III-A systems, but not for other type III subtypes (Figure 5.6D-I) due to a lack of statistical power. This meant that the acquisition bias observed in type III-A RNA-targeting subtypes may have been a consequence of a preference for strongly transcribed regions, such as protein-coding genes, rather than any specific priming effects. However due the lack of statistical power, it was not possible to conclude whether similar topologies occurred in other CRISPR RNA targeting systems (Table S5.1).

Type VI systems in contrast did not display significant acquisition biases. Although a bias was detected on the target strand of type VI-B and VI-D systems (Figure 5.6J-L,5.6M-O) this reflected a lack of statistical power (Table S5.1). Spacer distribution analysis was not performed on type VI-A systems, as the number of instances where two or more spacers mapped to the same contig was too low. However, it is notable trend that most spacers detected were located on the non-target strand (Figure 5.6K,5.6N). As with type III systems, this is consistent with acquisition being performed from transcribed RNA as well as DNA, although unlike in type III systems, no directional bias exists. This may indicate a difference in the underlying means by which spacers are acquired between these two types.


A
n: 6140
KS: ***
KS: ***
Protospacer density
1e-04
0e+00
-5000
-2500
0
2500
5000
Distance from protospacer (bp)
B
***
abundance
4000
2000
0
Non-target
Target
strand_bias
C
3'
5'
***
***
***
***
% protospacers
3000
2000
1000
0
5'
3'
Negative
Positive
strand
D
n: 62
KS: ns
KS: ns
Protospacer density
0.00050
0.00025
0.00000
-0.00025
-5000
-2500
0
2500
5000
Distance from protospacer (bp)
E
***
abundance
40
20
0
Non-target
Target
strand_bias
F
3'
5'
***
***
ns
ns
% protospacers
50
25
0
-25
-50
5'
3'
Negative
Positive
strand
G
n: 143
KS: *
KS: ns
Protospacer density
7.5e-05
5.0e-05
2.5e-05
0.0e+00
-2.5e-05
-5000
-2500
0
2500
5000
Distance from protospacer (bp)
H
*
abundance
100
75
50
25
0
Non-target
Target
strand_bias
I
3'
5'
*
*
ns
ns
% protospacers
50
0
-50
5'
3'
Negative
Positive
strand

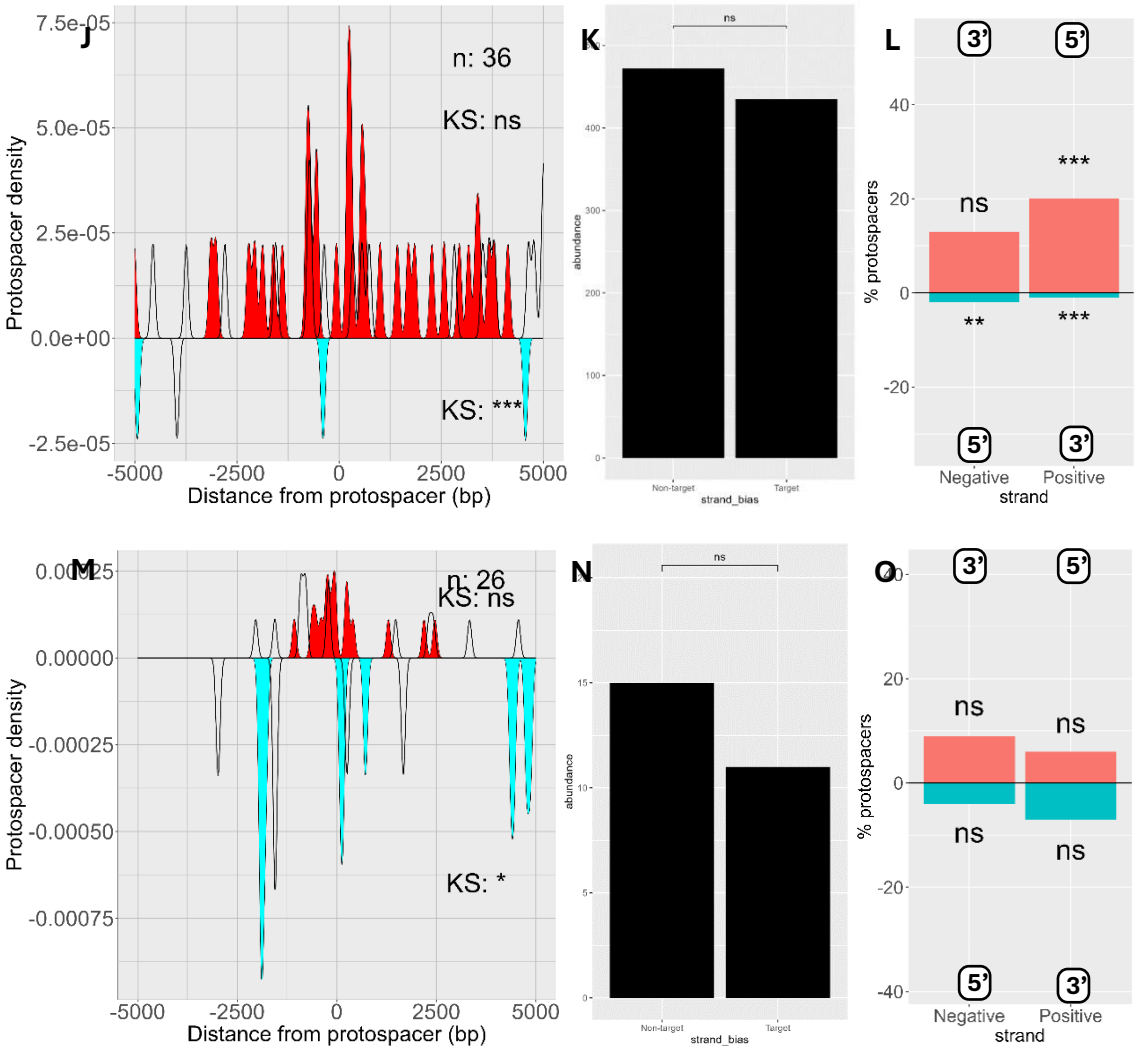


Figure 5.6: Expected vs. mapped spacer distances in RNA targeting CRISPR-Cas subtypes over a 5kb interval. *Plots were generated for each RNA targeting CRISPR-Cas system for whom an acquisition bias was investigated: (A-C) Type III-A, (D-F) Type III-B, (G-I) Type III-E, (J-L) Type VI-B, (M-O) Type VI-D. The procedure and layout utilised is the same as per Figure 5.4 & Figure 5.5.*

5.2.4: Modifying the spacer distribution analysis workflow to include partial matches

One shortcoming of spacer distribution analysis in its original form [175] was the requirement for relatively complete (> 90% similarity) spacer matches. I hypothesised that, provided one spacer match to a target genome exists, mappings with lower similarity, corresponding to sequences of phage genomes which have developed

escape mutations, can also be detected in fragmented genomes. To detect partial spacer mappings, a modified version of the original workflow which mapped spacers and measured spacer distributions was developed (Figure 5.7). Spacers from CRISPR-arrays containing at least one complete match to a mapped sequence were first divided into kmers of 10-14bp in size, then queried against the same mapped sequence as the whole spacer match. These kmer matches were then combined and subject to the same filtering and sequence de-duplication regime as applied to whole matches (Figure 5.2A). Comparing this spacer distribution with the analogous distribution containing complete matches revealed whether these additional mapped sites could enhance the sensitivity and context of the observed spacer acquisition biases in the CRISPR-Cas subtypes investigated.

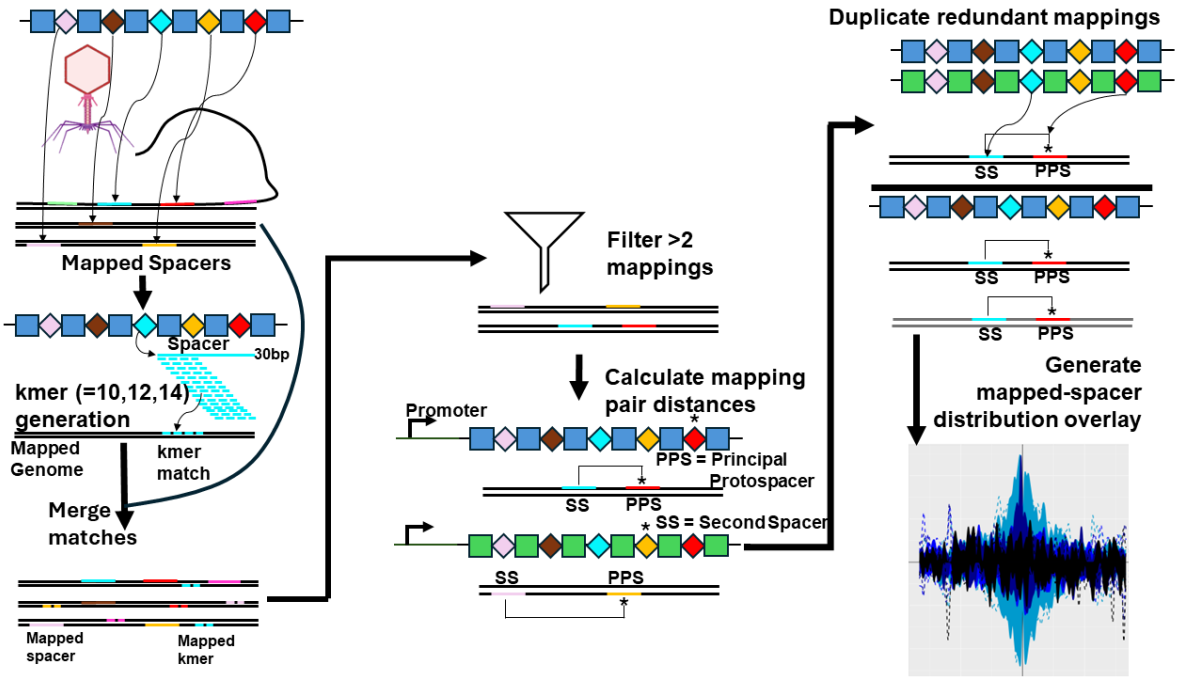


Figure 5.7: ***Outline of workflow used to perform spacer distribution analysis using partial spacer matches.*** *Spacers were used to produce sets of 10,12,14 letter kmers. These kmers were then used to search contigs containing at least one previous complete spacer match. Spacer distribution analysis was then performed as with complete spacer matches (Figures 5.4-5.6).*

a5.2.5 Partial spacer matches are enriched with complete matches to the same loci

Unlike with whole matches, the smaller number of nucleotides in each kmer used to search for matches in the mapped phage genomes incurs a higher probability of false-positive matches. However, these searches were undertaken in the context of a single relatively complete match on each contig, which was queried for partial matches. To determine whether the presence of complete matching spacers was associated with a high enrichment of partial spacer matches in mapped phage contigs tested due to the acquisition of multiple phage-escape mutations from historical spacer targeting, I took the mapped sequences (excluding mapping to their own arrays) where at least one spacer mapped to a target sequence, and then mapped the remaining spacers in the corresponding CRISPR-array to the same contig. I then took the same set of mapped contigs and performed a Monte-Carlo simulation whereby the remaining spacers in the CRISPR-array were mapped to a different contig, which was chosen at random from the mapped contig set. Target sites of the first complete spacer match were masked to prevent matches between any duplicate spacer-contig pairings. To determine whether this difference was significant, contig randomisation was repeated 100 times for each subtype and a p-value calculated at a $p<0.01$ threshold based on whether the number of partial matches to the same contig exceeded the number of partial matches to a randomised match. This showed whether the complete match was associated with an increased number of partial matches. As a negative control for this comparison, a set of contigs of the same size were generated with random nucleotide strings to determine the background level of partial matches by spacers.

Mapping of spacers using kmers derived from the kmerisation of each unmapped CRISPR spacer in a same array as at least one mapped spacer was initially performed using kmers of different sizes (Figure S5.5). However, it was found that kmers smaller than 14 were prone to high numbers of matches, which was reflected in the detection of higher levels of mapping of kmers of size 10, compared with 12, as a result of artifacts caused by high levels of non-specific mappings. For this reason, simulations to assess the number of partial spacer matches focused on a kmer size of 14 only.

For all subtypes investigated, the number of partial matches to contigs containing at least one whole match (Figure 5.8) significantly exceeded the number of partial matches to a different contig selected at random from the same mapped contig set in six CRISPR-Cas subtypes investigated. Although multiple simulations were not performed in the simulated random contig control. The lower number of kmer matches in this category compared with the matching using spacer-derived kmers against a reshuffled contig, as well as significantly lower numbers of matches across all six subtypes, suggests that the significance level is well below the 0.01 threshold observed for the comparison between the mappings of the reshuffled contigs and the mappings to the original contigs for which a complete spacer match was found. The difference in the number of mapped spacer targets between subtypes was partially related to the number of CRISPR-array encoding spacers which were used for each subtype. This result suggested that the additional partial matches were increasing as a consequence of a complete matching protospacer. However, this did not in and of itself prove the existence of an acquisition bias as the proximity between the complete and partial spacer matches was not calculated. To determine whether a distance-based spacer acquisition bias existed when partial spacer matches were detected in addition to complete matches, the spacer distribution calculation used above (Figure 5.4-5.6) was performed using partial spacer matches in addition to complete matches.

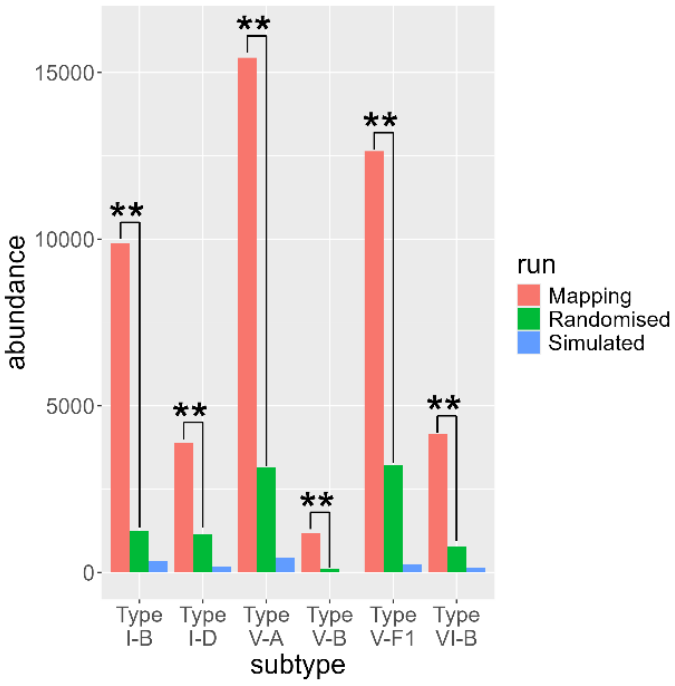


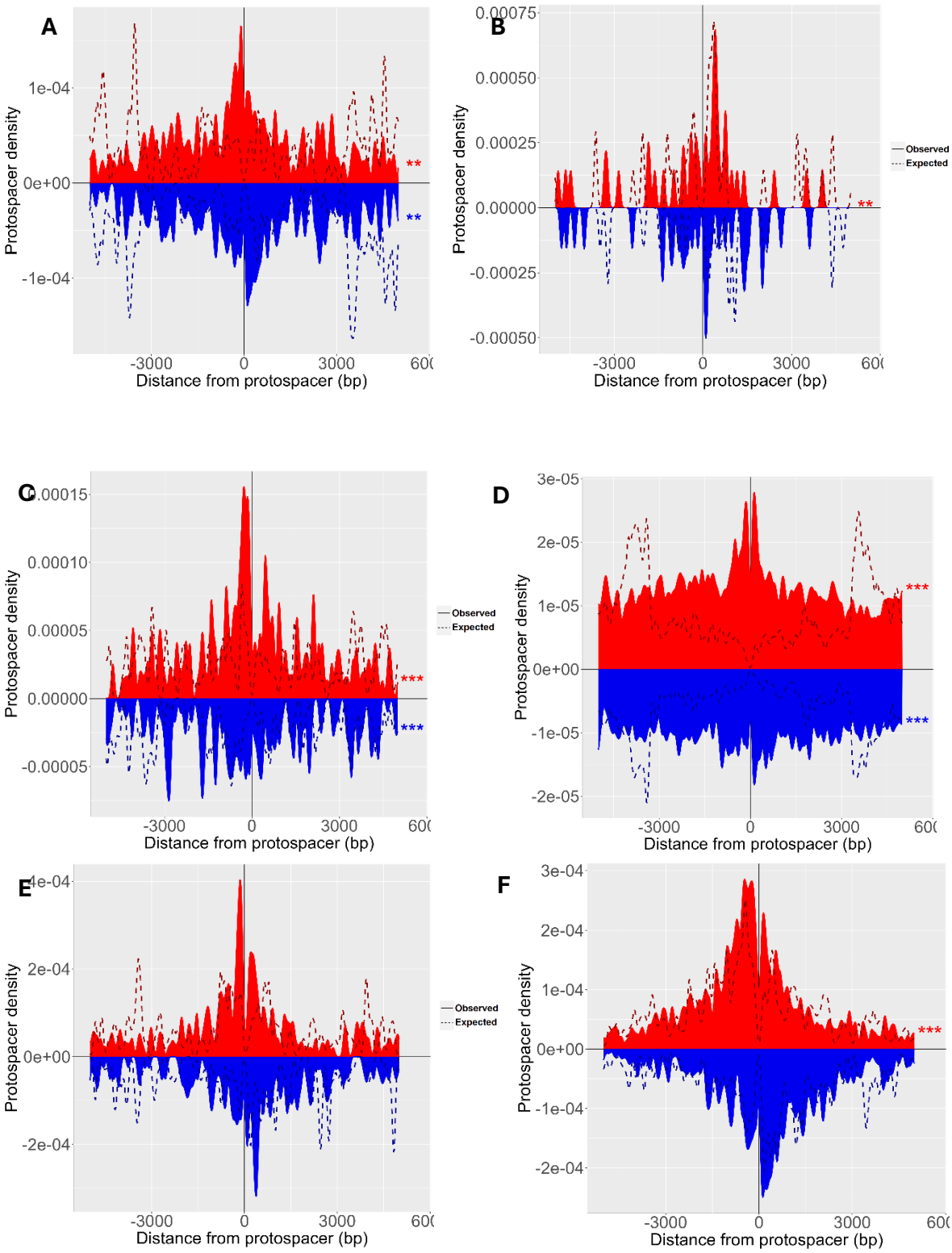
Figure 5.8: Comparison of spacer detection frequencies between test, randomized and synthetic controls in representative CRISPR-Cas subtypes. *Randomised spacer-contig pairings were simulated 100 times. A kmer size of 14 was selected as comparisons between Mapped, Randomised and Simulated lanes were inconsistent with smaller kmer sizes due to the high number of false positive matches as a consequence of the higher background probability of a match (see Figure S5.5). A p-value was assigned based on the number of times the value of the mapped spacer contig exceeded the randomized spacer contig. The simulated lane acted as a negative control and represented the number of successful mappings against a kmer generated as a random sequence of nucleotides of the same length as the original mapped contig. CRISPR-Cas subtypes where target contig randomisation and simulation was performed: A) Type V-A, B) Type V-B, C) Type V-F1, D) Type VI-B, E) type-IB, (F) type I-D.*

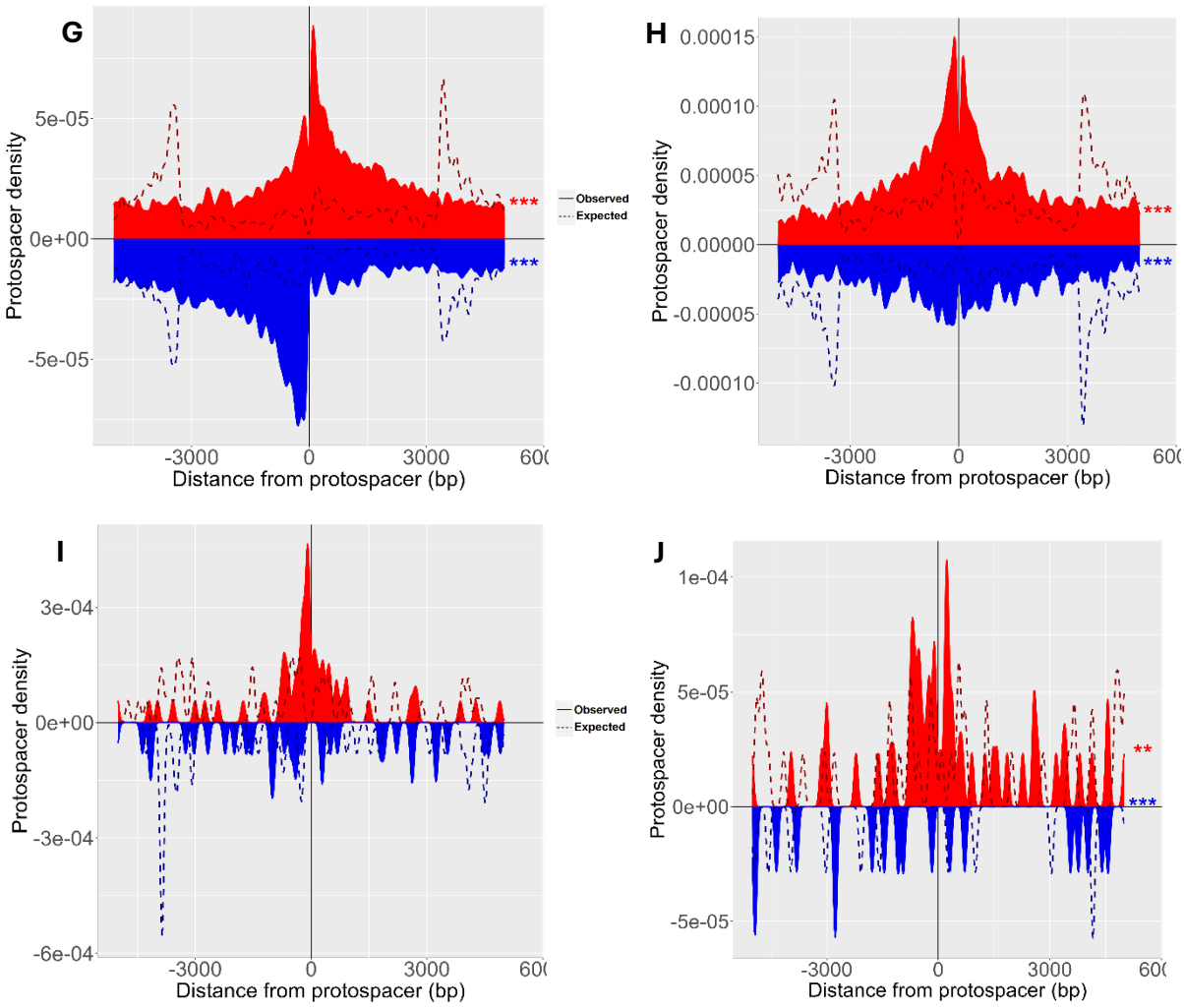
5.2.6: Detection of spacer-acquisition biases using partial spacer matches

To apply the improve sensitivity of target site spacer mapping using partial kmer-based matches, I performed spacer distribution analysis on the union of complete spacer matches and kmer matches of a particular spacer size in a subset of the CRISPR-Cas subtypes investigated. Partial matches less than 50bp from each other were excluded from analysis. This, in conjunction with the deduplication of kmer matches at the same position meant that only unique partial matches added to the sample size. Kmer matching was performed on 20kb phage contig windows upon which at least one complete spacer match occurred. Using sliced phage contigs in this manner introduced a small artefact into the data which inflated the mapping density of the expected distances. This reduced the sensitivity of kmer mapping and the statistical significance of the results, but did not otherwise alter the results.

The increase in sample size produced a modest increase in the sensitivity of the spacer distribution analysis technique. In Type V systems (Figure 5.9A-5.9C) an acquisition bias was observed on both the top and bottom strands in subtypes V-A and V-B. This contrasted with complete matches where a bias was only observed in type V-A on the non-target strand (Figure 5.5B) and type V-F1 on both strands (Figure 5.5I). This bias was more accentuated near the origin compared with complete spacer matches. In Type I

and II systems (Figure 5.9D-5.9G) the same result was observed as whole matches (Figure 5.4, 5.5). The additional sample size of kmer matches produced a significant result in type VI-B systems on the transcribed non-target strand (Figure 5.9J). This was consistent with the expected pattern of acquisition if VI-B systems acquired spacers from RNA as opposed to DNA. In type III-A and III-B systems, an acquisition bias on the non-target strand was observed (Figure 5.9H-I). Interestingly, this bias was not detected in Type III-B systems when whole matches were used alone. One notable difference between partial and whole matches in type III systems was the presence of a higher peak in the upper right quadrant, suggesting less of a bias upstream or downstream of the PPS than observed with complete spacer matches only (see figure 5.6A-F). These results demonstrated that spacer distribution analysis, using partial spacer matches in concert with complete matches, results in a modest increase in sensitivity but also the appearance of several protospacer density peaks which were not visible when complete matches were utilised for the distribution analysis alone.


A
B
C
D
E
F
Protospacer density
Distance from protospacer (bp)
Observed
Expected
**
***



<u>Figure 5.9:</u> Spacer distribution analysis of the union of whole and partial subtype matches to a subset of CRISPR-Cas subtypes investigated. *The spacer distributions were generated using the same approach as for whole spacer matches (Figure 4). Spacer distribution analysis using partial spacer matches was performed on following subtypes: (A) Type V-A, (B) Type V-B, (C) Type V-F1, (D) Type II, (E) Type I-D, (F) type I-B, (G) type I-F, (H) type III-A, (I) Type III-B, (J) Type VI-B.*

5.2.7: Biases in strand directionality using partial spacer matches

I reasoned that a bias in directionality would be observed for partial spacer matches which might be different from complete matches in the same CRISPR-Cas subtype. To compare the directionality of partial spacer matches to complete matches, a bias in the number of PPS-spacer mapping pairs in each quadrant was computed using binomial probability. For DNA targeting class II subtypes, no significant bias in directionality was observed in type V-A, V-B and V-F systems (Figure 5.10A-C). In type II systems a bias in

directionality was observed on the non-target strand (Figure S5.6A) in the 3' direction of the PPS (Figure 5.10D). Apart from a much clearer bias in this direction in type I-B systems, the directionality of type I CRISPR-Cas subtypes using partial spacer matches (Figures 5.10E-G) was broadly concordant with the directional bias observed in complete matches (Figure 5.4). In RNA targeting subtypes, a slight significant directional bias was observed in the 3' direction from the PPS on the non-target strand of VI-B systems (Figure 5.10E). This was less significant compared to the directional bias observed in whole matches (Figure 5.6L). This effect was also observed in type III-B systems in the form of a less significant directional bias in spacer acquisition toward the target strand 3' from the PPS (Figure 5.10J) compared to complete spacer matches (Figure 5.6F).

Figure 5.10: Analysis of strand directional biases in spacer acquisition using partial spacer matches. *Significant differences in the amount of acquisition in each quadrant were determined using binomial probability compared to the mean number of spacers in each quadrant (As per figure 5.4-5.6). This was only performed for partial matches where Kmer size = 14. This was performed for: (A) Type V-A, (B) Type V-B, (C) Type V-F1, (D) Type II, (E) type I-B, (F) Type I-D , (G) type I-F, (H) type VI-B, (I) type III-A, (J) Type III-B,.*

Chapter 5.3: Discussion

Although previous investigations have characterised different mechanisms of primed spacer acquisition, the widespread nature of priming in many rare and RNA targeting CRISPR-Cas subtypes remains undetermined[127,131,170,173,175,245,246,253]. I applied a previously developed technique to bioinformatically estimate spacer acquisition biases with an enhanced sample size by performing spacer-mapping on metagenomes at the terabyte scale, to type III, VI and V subtypes. In these subtypes, the degree to which priming is a widespread phenomenon has not to date been determined, unlike other subtypes. I then further extended the technique by including partial matches which may represent sites of escape mutation acquisition in response to site-specific spacer targeting, which has been shown to occur rapidly on short timescales [252].

I found evidence for acquisition biases in the majority of subtypes tested including RNA -targeting III-A systems as well as replicating findings for more conventional DNA targeting subtypes such as type V-F1. In line with previous findings[127,131,174,175] , the directionality of biased acquisition was observed to be subtype-specific. Comparing this spacer distributions between subtypes revealed a similar fingerprint for DNA-targeting class II systems and RNA targeting type III and type I systems. These fingerprints are effectively a representation of multivariate factors which influence the acquisition process. This suggests underlying differences in the mechanisms of acquisition between these types, which is consistent with prior observations that pre-spacer substrates can be generated from a priming protospacer in three ways: resection during DNA repair[179,181,254] , acquisition directly from transcribed RNA[148,183,255] or from the production of DNA fragments via processive strand degradation mediated by an interference complex[68,170,171,256].

5.3.1: Acquisition biases are a conserved feature of most, yet not all type I systems

While priming was found to occur in most subtypes, there were several notable cases where the acquisition bias presented differently or was absent compared with priming from other members of the same type. The acquisition bias observed for type I-F systems was unique in appearing to occur with the opposite directionality (5’ of the PPS) compared to other type I systems. This has been previously observed

computationally [175]. However, Cas3, which performs the unwinding and presumptive production of pre-spacer fragments in other subtypes such as type I-E, I-C and I-B [166,170,174,177,254,256-259], has been shown to possess 3' to 5' unwinding activity in type I-F systems[127,173], in common with other subtypes. This suggests that an alternative undiscovered pathway for biased acquisition exists in type I-F systems. In contrast, the acquisition bias observed for type I-B had similar directionality to previously observed biases in type I-C and I-E, which have been previously observed both bioinformatically and experimentally. [170,245,256] No strand bias was found for type I-D systems due to the lack of statistical power,[260] Interestingly, an acquisition bias for type I-A systems was also observed, which has not been reported before. Unlike other type I subtype acquisition biases, no strong strand directional preference was observed, although there was a slight but significant preference for spacers acquired 3' of the protospacer. This was significant, given that the requirements for type I-A acquisition have been experimentally characterised and features mandatory integration from the leader-repeat junction [261] which was a key-assumption required for the spacer distribution analysis to work correctly. The lack of strand directionality and relatively low effect size of the biased spacer distances bears some resemblance to the biases observed in DNA targeting class II systems, which may suggest a similar passive cause, such as the repair of strand breaks or ablation caused by the interference complex. Contradicting this postulate, was the absence of a similar effect being observed bioinformatically and experimentally in other type I systems which facilitate interference in a similar manner[127,170,173,175,245,256].

5.3.2: Acquisition biases are observed in class 2 systems, but are distinct from type I systems

Acquisition biases in DNA targeting class 2 CRISPR-Cas systems (subtypes II and V) were generally characterised by a mild preference for spacers from the non-target strand as well as a relative lack of 5' or 3' strand directional preference. Previous work has shown that acquisition biases arise indirectly as a consequence of errors in DNA damage repair, due to the ability of the RecBCD complex and it's orthologs to produce ssDNA fragments during resection prior to HDR or MMEJ directed repair. These fragments are ideal substrates for the Cas1-Cas2 acquisition proteins [7,9,179,181]. This is

the most likely source of the acquisition biases I observed in type V-A, V-B and II systems. However, the acquisition bias I observed in type V-F systems featured a distinct preference for the target strand 3' of the PPS, which has not been previously reported. Given that these systems appear to be plasmid and phage encoded (Chapter 4: Figure S4.3F) it is plausible that these systems function in an adaptive manner analogous to host-encoded systems but with a unique adaptation pathway to acquire new spacers.

5.3.3: Evidence of Acquisition biases in RNA targeting systems

Unlike the subtypes in which spacer distribution analysis was applied, type III and type VI have been shown to target RNA either co-transcriptionally, or in the host cell cytoplasm [116-118].Among RNA targeting systems, an acquisition bias was detected only in type III-A systems, This bias was still observed in type III-A when PPS-spacer distances were confined to the range of single genes, suggesting that this effect is not a consequence of the differential expression of nearby encoded genes or an overestimation of the expected spacer distribution based on the contig size . Comparing a subset of type III-A systems with and without Cas1-RT also suggests that this effect was independent of RT, although it is possible that an RT which facilitates acquisition could be distantly encoded from CRISPR loci, and thus not detected. Previous observations of spacer-acquisition from RNA in type VI-B[114,148,183] suggested a bias toward the non-target strand, which was rationalised as direct acquisition of RNA. The acquisition bias observed in my results could reflect this, but the possibility of DNA-based priming protospacers couldn't be excluded.

5.3.4: Performing spacer distribution analysis using partial spacer matches increases sensitivity but reduces strand bias directionality

Using partial spacer matches increased the number hits observed in some subtypes but changed the topology of the spacer distributions observed. In class 2 DNA targeting systems, additional peaks were observed on the target and non-target strand in a symmetrical manner, which were not present when complete spacer matches were

used. An unexplored aspect of spacer mapping is how much time has elapsed between when an acquisition event occurred for a given CRISPR locus in a given subtype, and when this locus was physically sequenced. Given that phage escape mutations have been shown to occur rapidly between host and phage in solution [252], the detection of complete spacer matches reflects either the timepoint immediately post-acquisition, escape mutations in the PAM only, or target sites which themselves are under counter selection pressures which prevent them from escaping. Previous work suggests that escape mutations do not remain confined to the PAM[262-264]. Therefore, in the absence of an internal regulatory role, the differences in complete and partial spacer mappings likely reflect differences in recent target site preference compared to older target sites. The stronger peaks present at short distances from the priming protospacer in some subtypes such as type V-A and II systems may thus represent a significant number of older target sites against which phage and plasmidic MGEs have evolved escape mutations.

5.3.5: Limitations inherent in the bioinformatic estimation of spacer acquisition biases

Although my findings predict priming-like effects in several subtypes where this effect has not previously been reported, the findings come with several caveats in addition to those already mentioned. The priming protospacer (PPS) was presumed, rather than identified for each CRISPR-array, as the oldest spacer in the array for which mapping to a target sequence could be identified. This was effective at estimating acquisition biases in the aggregate but not systematically quantifying them. Additionally, this approach could only identify biases in acquisition as a function of the presumptive protospacer distance, it could not ascertain the cause of these biases directly. The difference between spacer distributions when partial matches were pooled together compared with whole matches alone, may suggest a different timepoint in the post-acquisition timeframe between from these spacers were acquired and when counter selection pressures led to escape mutations or fixation. Only an experimental characterisation *in vivo* of spacer acquisition dynamics could distinguish this however, which was outside the scope and resources of this investigation. This, coupled with a

more rigorous inference methodology to identify the PPS from CRISPR-arrays would build on the effectiveness of this method to enable more quantitative predictions of widespread acquisition biases across any given clade of CRISPR-array encoding sequences of interest.

5.3.6: Summary of findings

Overall, my results have computationally predicted spacer acquisition biases using spacer distribution analysis in a number of additional CRISPR-Cas subtypes where this phenomenon was not known to occur. Extending the technique to use partial spacer matches appeared to increase the sensitivity of spacer distribution analysis, yet also changed the distribution of primed spacers in certain subtypes. This suggests that the apparent priming effects generated by CRISPR-Cas subtypes are a diverse and complex phenomenon, many aspects of which still remain to be characterised.

Chapter 6: General discussion

Many significant biotechnological advances in the past ten years have been made by exploring and exploiting the highly ubiquitous and diverse nature of anti-phage defence systems, with a specific focus on CRISPR-Cas systems. The majority of these systems have been discovered through the use of guilt-by-association search strategies[22,57,89]. These leverage Hidden-Markov model (HMM) profiles of known proteins to recognise large volumes of assembled genome sequence data, sourced in large-part from metagenomic sequencing of environmental samples[22,91]. Previous works have focused their approaches on the discovery of new CRISPR effectors or accessory proteins using a single gene approach in a dominantly "CRISPR effector centric" manner[57,87,100,102,110]. A key missing gap in the field is to evaluate the complex relationship between the CRISPR effectors, their accessory proteins and the phages in terms of their global diversity. This ultimate goal aims to elucidate the biology of CRISPR defence at contig/taxa rather than single gene level.

6.1: Key findings

In this study I employed a three-pronged approach to survey different aspects of CRISPR-Cas systems diversity. These comprised a guilt by association-based data mine of known and unknown CRISPR-Cas systems and associated genes from assembled genome sequence data; a network-based interrogation of intra-subtype and spacer-mapped taxonomic cluster diversity and a computational analysis of spacer acquisition biases as a consequence of spacer mapping by priming protospacers. From each of these approaches, I uncovered novel co-associated genes in Type VI CRISPR-Cas systems, explored the diversity in gene composition at the level of local clusters in subtypes of type VI systems as well as the diversity of the corresponding mobile genetic elements (i.e. phages/plasmids) which predate them, and uncovered evidence of primed spacer acquisition in CRISPR-Cas subtypes where this phenomenon was not previously known to occur. Each of these findings is further delineated in the following subsections:

6.1.1: The landscape of CRISPR-associated genes is filled with MGEs or Antiphage defence genes.

The results from my computational investigation of 10 TB of assemblies using a guilt-by-association based methodology revealed extensive numbers of genes encoded in proximity to various CRISPR-Cas subtypes with detectable homology to Mobile Genetic Element (MGE) or anti-phage defence genes. Although many of these genes may not perform a role in CRISPR-immunity directly, a significant fraction may play related roles conferring protection against phages, forming part of larger defence islands which coordinate a more systematic response to phage infection. This suggests that metrics such as CRISPRicity, which were designed to screen for the core-CRISPR-Cas module genes involved in interference, acquisition and processing[90,91], may be less effective when screening for genes which perform accessory roles or are involved in cross-talk with other anti-phage defence systems. As a case in point of this, I discovered two additional modules derived from anti-phage systems: The *HicAB* toxin-antitoxin system[222], and *DrmB* from the DISARM system[32], which appear to co-occur with Type VI-A and Type VI-D CRISPR-Cas subtypes respectively. This demonstrates that additional genes with potential functional importance in the facilitation of CRISPR-immunity may yet still be discovered despite the already extensive use of this approach to date to uncover novel CRISPR-Cas subtypes and functions.

6.1.2: Network based taxonomic classification reveals extensive segregation at the level of local clusters

Using vConTACT2[152] to analyse the intra-subtype diversity of type VI CRISPR-Cas systems revealed a high degree of cluster segregation, with relatively few shared genes between clusters. Interestingly, only the effector C*as13* gene was completely conserved in each subtype. This included the acquisition genes *Cas1* and *Cas2* where many local clusters cases were not observed co-encoded either with Cas13 or another nearby subtype. This contrasted with the conservation of anti-phage defence genes within local clusters, with was consistent with previous findings showing that antiphage defence modules tend to be frequently exchanged even at the level of individual taxa[23] which the local cluster groupings often approximated. This illustrated the effectiveness of network-based approaches in segregating the intra-subtype diversity of type VI systems

into well-defined groups and highlighted the differences in gene co-association in intra-subtype clusters compared to the entire type VI CRISPR-Cas subtype.

6.1.3: Regions upstream and downstream of target sites of mapped type VI spacers display large numbers of semi-conserved genes.

Unlike network-based representations of Type VI-systems, the spacer-mapped targets of each type VI CRISPR-Cas subtype formed a single-component network of semi-conserved shared genes. The targets of most CRISPR-spacers were predominantly either phage or plasmidic in nature. Interestingly, this was true even for a distinct cluster of Type VI-B systems encoding contigs which were predicted to be prophage encoded, suggesting that these systems participate in inter-viral competition with other phages. Most of the genes shared between clusters were essential genes involved in phage replication. These included genes with nuclease domains such as HNH and TerL which fulfill roles in the packaging of phage genomes[238-240]. Although not a conserved feature between clusters, several phage encoded TnpBs and IscBs were also detected. The role of these genes in phage survival and replication remained undetermined however. These findings illustrated the high degree of gene flow between mobile genetic elements and serves as the first effort to annotate and differentiate taxonomic clusters of spacer-mapped phage and plasmidic mobile genetic elements which predate Type VI-CRISPR Cas system encoding host-cells.

6.1.4: Primed spacer acquisition is a widespread feature in diverse CRISPR-Cas systems

By performing spacer distribution analysis across diverse CRISPR-Cas subtypes, I identified uncovered evidence of spacer acquisition biases in type I-A and a type V-F systems. I also discovered evidence for a priming effect in type III-A although this evidence was confounded somewhat by the ability of some type III-A systems to acquire spacers directly from RNA. The pattern of spacer acquisition differs dramatically between subtypes. Some subtypes, such as type I-F systems, appear to acquire spacers in the 5’ direction bidirectionally (with respect to the non-target strand), while type I-B systems appear to acquire spacers in the 3’ direction. In type II and V systems an acquisition bias is generated in a much less strand specific manner. This likely

reflects different modes by which the underlying acquisition bias is generated. It is possible that some spacer mapping distributions are formed by a combination of these modes, but the dominant form is what defines the shape of the observed spacer mapping distribution. Nevertheless, this does not detract from my overall conclusion that primed acquisition is conserved phenomenon present in a wide range of CRISPR-Cas Subtypes.

6.1.5: The sensitivity of spacer distribution analysis can be enhanced by partial spacer matches yet sometimes changes the underlying distribution

I attempted to expand the effectiveness of spacer distribution analysis by additionally mapping partial spacer matches in addition to near-complete spacer matches in cases where at least one match was already found to a mapped sequence. Although this did result in an increase in the number of spacers, particularly in type V systems, it also resulted in alterations in the overall topology of the spacer distribution in some subtypes. One possible explanation for this, is that the observation of complete and partial spacer matches occurs at different time points post-acquisition of the mapped spacer sequence in question. Despite this, my approach represented an expansion and improvement in sensitivity upon the original method used to measure acquisition biases.

6.2: Significance of the work

Although many guilt-by-association based studies have been performed in the past to uncover new CRISPR-Cas systems, this work is distinguished from these investigations by its emphasis on surveying and profiling the output from this process. I clearly establish that a significant fraction of the co-encoded genes in proximity to CRISPR-arrays are not false positives but genes with related functions as Mobile Genetic Elements involved in cassette exchange or anti-phage defence. Additionally, I demonstrate through the discovery of the HicAB genes co-encoded with Type VI-A systems and DrmB co-encoded with Type VI-D, that the pool of genes left to discover with potential accessory roles in CRISPR immunity may not be as depleted as the large

number of previous guilt-by-association-based data mines imply[21,22,81,84,87,88,90,99,100,102,110,123].

In using vConTACT2 to construct gene networks as opposed to trees for taxonomic representation, I attempted to jointly annotate intra-cluster level differences between host-encoded type VI systems and the mapped sequences of each subtype's CRISPR spacer repertoire. The evolution of these two groupings is inextricably linked through red-queen competition[252,265]. My investigation was the first to investigate both segregated and shared genes in the context of this reality in type VI CRISPR-Cas subtypes. This uncovered a number of shared genes among the mapped sequences, and affirmed the identity of the mapped sequences as phage or plasmid even when the spacers used for mapping were derived from prophage encoded Type VI-B arrays.

Finally, I uncovered evidence that primed spacer acquisition may also occur in type I-A systems as well as V-F systems. I also extended the bioinformatic technique used to measure acquisition biases by including partial spacer matches. This was effective in increasing the sensitivity of the technique in certain subtypes, enabling this technique to be used more effectively when estimating acquisition biases in rare CRISPR-Cas subtypes with few mapped sequences.

6.3: Limitations and future directions

The nature of this work being confined to computational surveys, the findings in this study are presented fundamentally as predictions. As such, the exact functions and activities of associated or shared genes noted in this study may only be elucidated by wet-lab based experimental characterization. This reality also applies to the acquisition biases predicted for type I-A and type V-F systems.

There was additionally significant over-representation of assembled sequence data used in this investigation from certain environments. Metagenome sequencing of human gut microbiota and sewerage derived sample is performed at an industrial scale in contrast to other samples. More recent sequencing investigations performed in the period since this investigation was first undertaken have significantly diversified the

sampling base, and consequently reduced this bias[102]. The HMM based tools used for annotation, while still effective, have also begun to be supplanted by direct structure-based search methods[98], as a result of dramatic improvements stemming from the development of Alphafold2 for protein fold prediction[97]. A future investigation continuing on from this work presented across all 3 chapters would incorporate these advances into its design.

6.4: Final statement

The three-pronged survey employed by this study revealed numerous pockets of unexplored CRISPR-Cas system diversity in terms of both associated genes and acquisition biases which suggest priming-like effects in additional CRISPR-Cas subtypes. Despite the extension use of guilt-by-association in uncovering new CRISPR-associated genes[21,22,57,79-89,90], the detection of co-encoded HicAB and DrmB genes in certain type VI systems, along with a substantial fraction of co-encoded detected genes with no known role in CRISPR immunity, suggests a remaining pool of genes with potential key ancillary roles in facilitating acquired immunity, interference or synergistic anti-phage defence, which have yet to be discovered. This finding remained valid when my investigation was extended toward an analysis of the intra subtype diversity of CRISPR-Cas systems and the target sites of spacers from type VI systems. Furthermore, existing computational approaches to assess priming effects, which serve as a fingerprint of CRISPR-Cas mediated interference[175], were further refined and extended to new subtypes which again revealed subtypes where the mechanism of spacer acquisition and interference appears to contain unique features which have not yet been characterised. These findings demonstrate the value of several inter-related yet distinct methods for uncovering additional novel activities within CRISPR-Cas subtypes using *in silico* techniques, providing a stronger basis for further characterization of these systems.

Code Availability:

The code and external data used for this project can be found in a public repository via the following url:

https://github.com/u5581638/Tripartite-CRISPR-diversity

Appendix:

Supplementary data for Chapter 3

Using a computational pipeline to survey gene diversity associated with CRISPR-Cas systems

Pipeline stage	Remaining contig/protein sequences
Total number of contigs	*136.22 billion contigs [6.99 billion (NCBI), 129.24 billion(JGI)]*
Size of data	*9.8TB [1.67TB (NCBI), 8.11TB (JGI)]*
Number of CRISPR windows extracted	*3.76 million*
Number of predicted ORFs	*12 million*
Number of clusters	*72398*
Number of clusters after filtering	*31066*
Clusters <10kb from array after filtering	*15468*

Table S3.1: ***Breakdown of the origin, type and amounts of assembled metagenome sequence data processed for putative CRISPR-Cas systems***

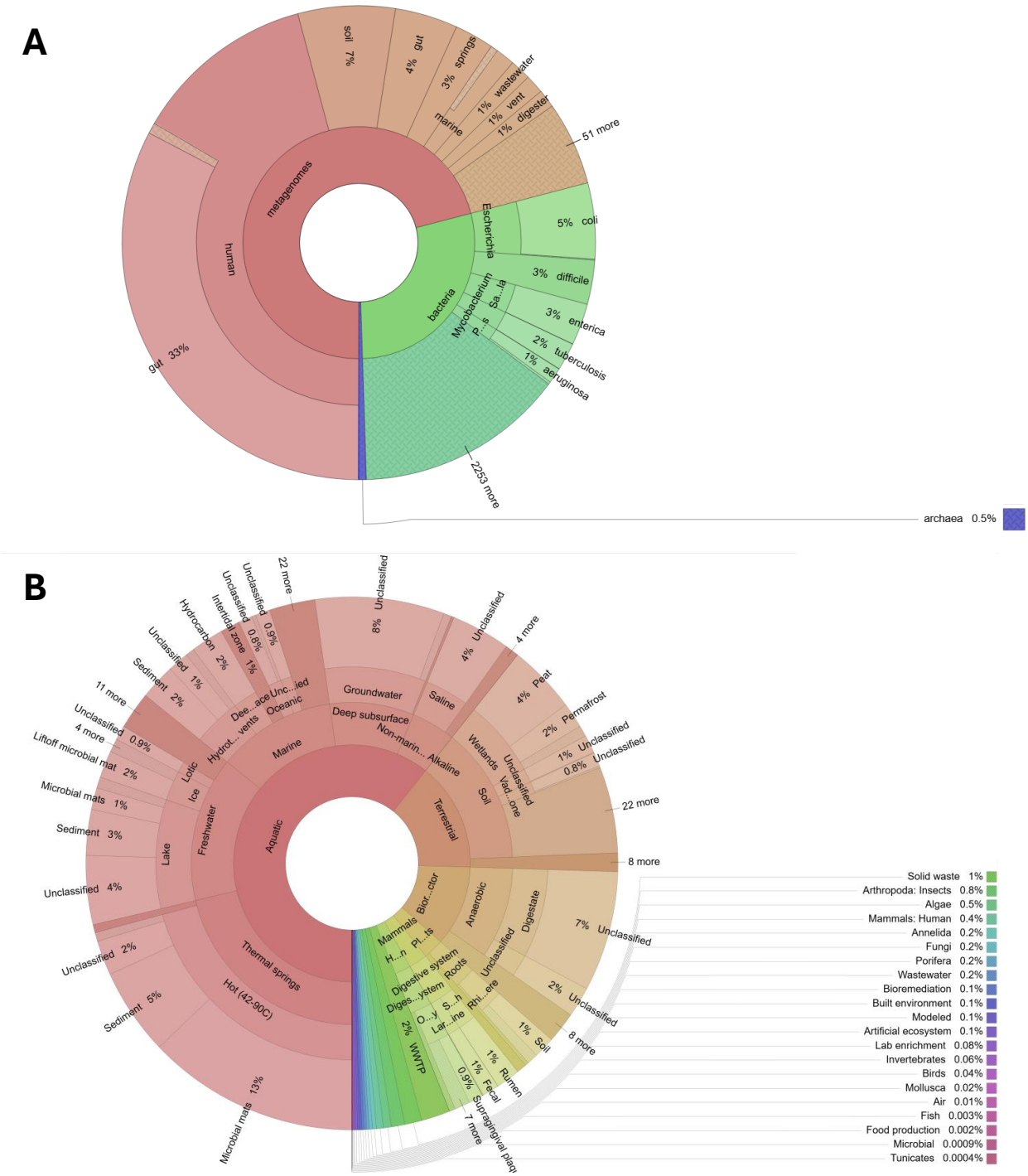


Figure S3.1: ***Taxonomic and environment composition of extracted CRISPR-array containing windows sourced by the (A) NCBI-GenBank and (B) JGI repositories.***

Statistic	Co-occurrence	Distance	Abundance
Variance	0.95276	0.066772	1.036626
F-test	14.26884 **(Co-occurrence/distance)**	1.088025 **(Abundance/Co-occurrence)**	15.52485 **(Abundance/Distance)**
P-value	<0.00001	<0.00001	<0.00001
$Log_{10}(p)$	-11009.1	-10.7839	-11586.1

Table S3.2: ***Summary statistics computed for each dimension used to screen CRISPR-associated proteins***

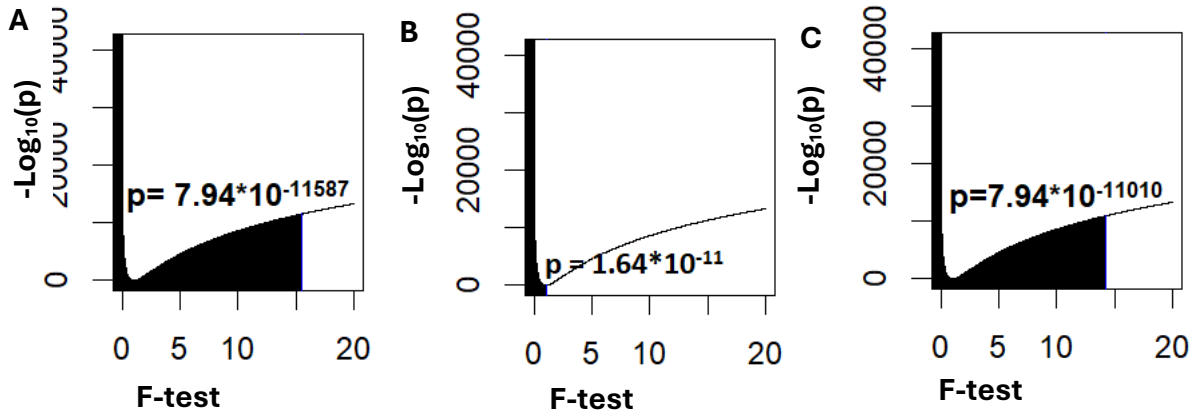


Figure S3.2: F-distributions comparing the variances of each dimension used to screen CRISPR-Associated proteins. *A) Abundance/Distance, B) Abundance/Co-occurrence, C) Co-occurrence/Distance.*

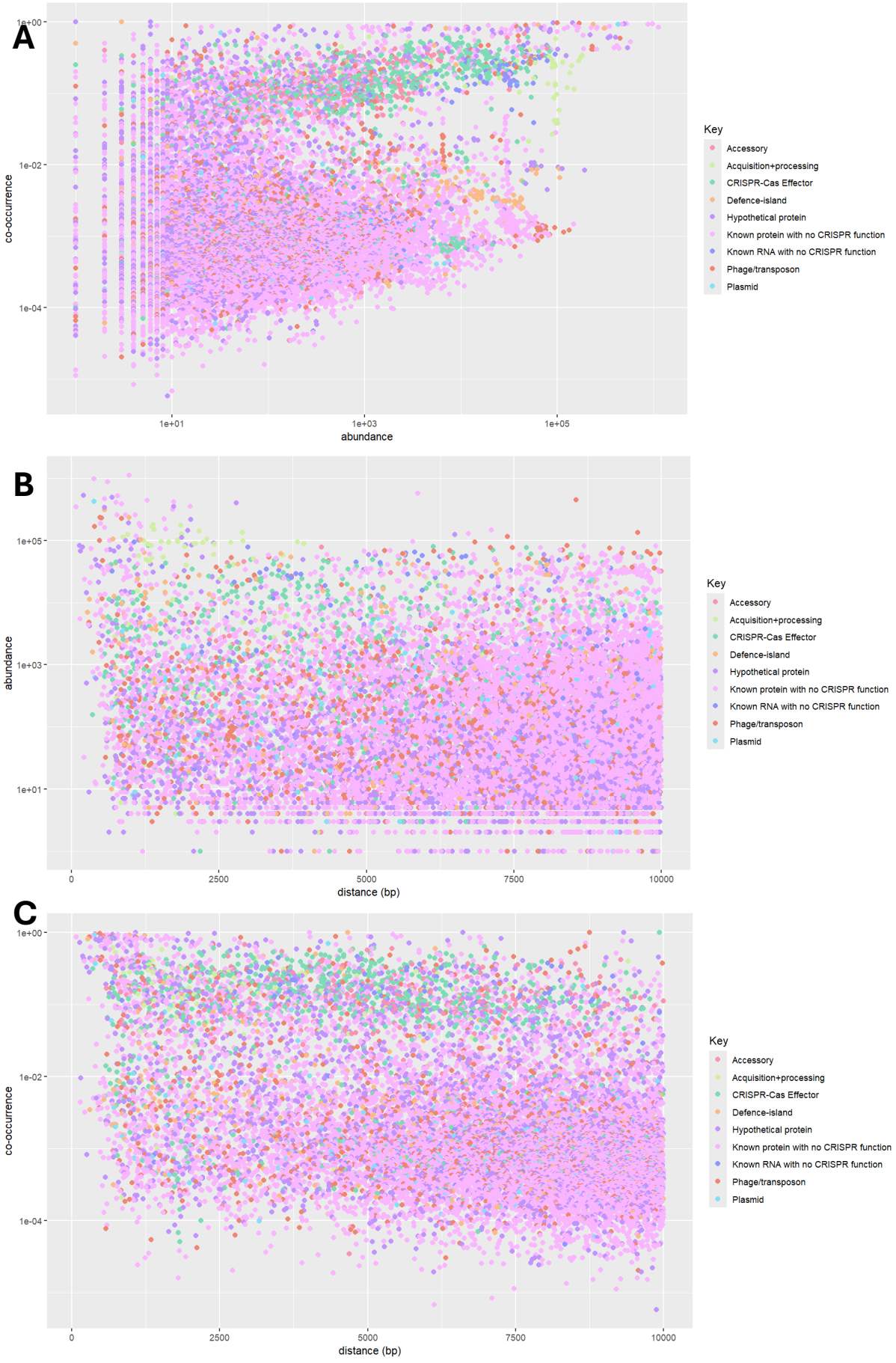

A
co-occurrence
1e+00
1e-02
1e-04
1e+01
1e+03
1e+05
abundance
Key
Accessory
Acquisition+processing
CRISPR-Cas Effector
Defence-island
Hypothetical protein
Known protein with no CRISPR function
Known RNA with no CRISPR function
Phage/transposon
Plasmid
B
abundance
1e+05
1e+03
1e+01
0
2500
5000
7500
10000
distance (bp)
Key
Accessory
Acquisition+processing
CRISPR-Cas Effector
Defence-island
Hypothetical protein
Known protein with no CRISPR function
Known RNA with no CRISPR function
Phage/transposon
Plasmid
C
co-occurrence
1e+00
1e-02
1e-04
0
2500
5000
7500
10000
distance (bp)
Key
Accessory
Acquisition+processing
CRISPR-Cas Effector
Defence-island
Hypothetical protein
Known protein with no CRISPR function
Known RNA with no CRISPR function
Phage/transposon
Plasmid

Figure S3.3: ***Distribution of candidate CRISPR-associated proteins by classified identity.*** *Each point was classified into one of 9 different categories. These were plotted onto three separate axes:* (A) *Relations between co-occurrence and distance from the CRISPR array, (B) abundance within 20kb of a CRISPR-array and distance and (C) co-occurrence and distance. Each point represents one putative CRISPR-associated protein family.*

Name	Description	Examples
CRISPR-Cas effector	Known CRISPR-Cas effectors	cas9,cas12a, cse1,csy3, csm6,cas10
Acquisition+ processing	Core genes required for spacer acquisition and downstream pre-crRNA processing	Cas1,cas2, cas4,csn2, cas6
Accessory	Known accessory genes associated with CRISPR-Cas systems	WYL,csx27, csx,28,csx20
Defence-island	Genes involved in anti-phage defence outside CRISPR immunity – usually co-encoded in large defence-island gene superstructures	ParB,ietA, JetC,AbiE
Phage/Transposon	Genes of unknown function with homology to phage/transposons	Transposase, Integrase, Phage Terminase
Plasmid	Genes of unknown function with homology to conserved genes involved in plasmid replication or conjugation	Conjugal transfer genes, plasmid replicase genes.
Known RNA with no known CRISPR-function	Genes of unknown function associated with CRISPR arrays which encode RNA molecules	DUF2800, Retrons (some overlap with Defence islands)
Known protein with no known CRISPR-function	Genes of unknown function which are protein encoding and associated with CRISPR-arrays	Glycotransferase/ metabolism genes.
Hypothetical proteins	Genes of unknown function with no annotated homology to any sequences in the database	N/A

Table S3.3: ***Categories used to assign and quantify putative CRISPR-associated proteins by domain homology to pfamA_35.0 profiles.***

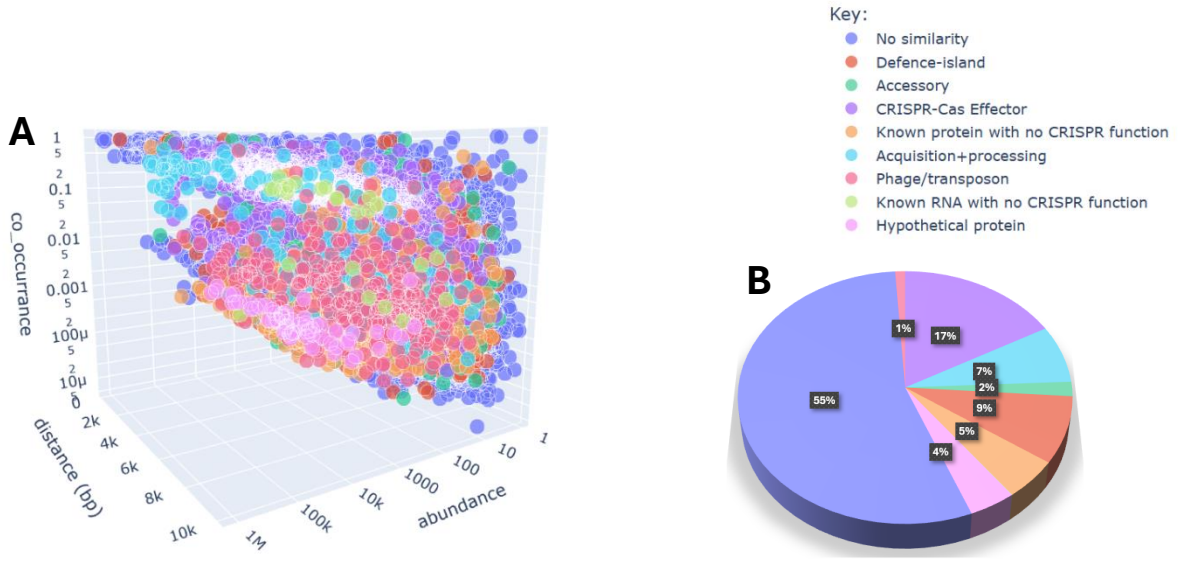
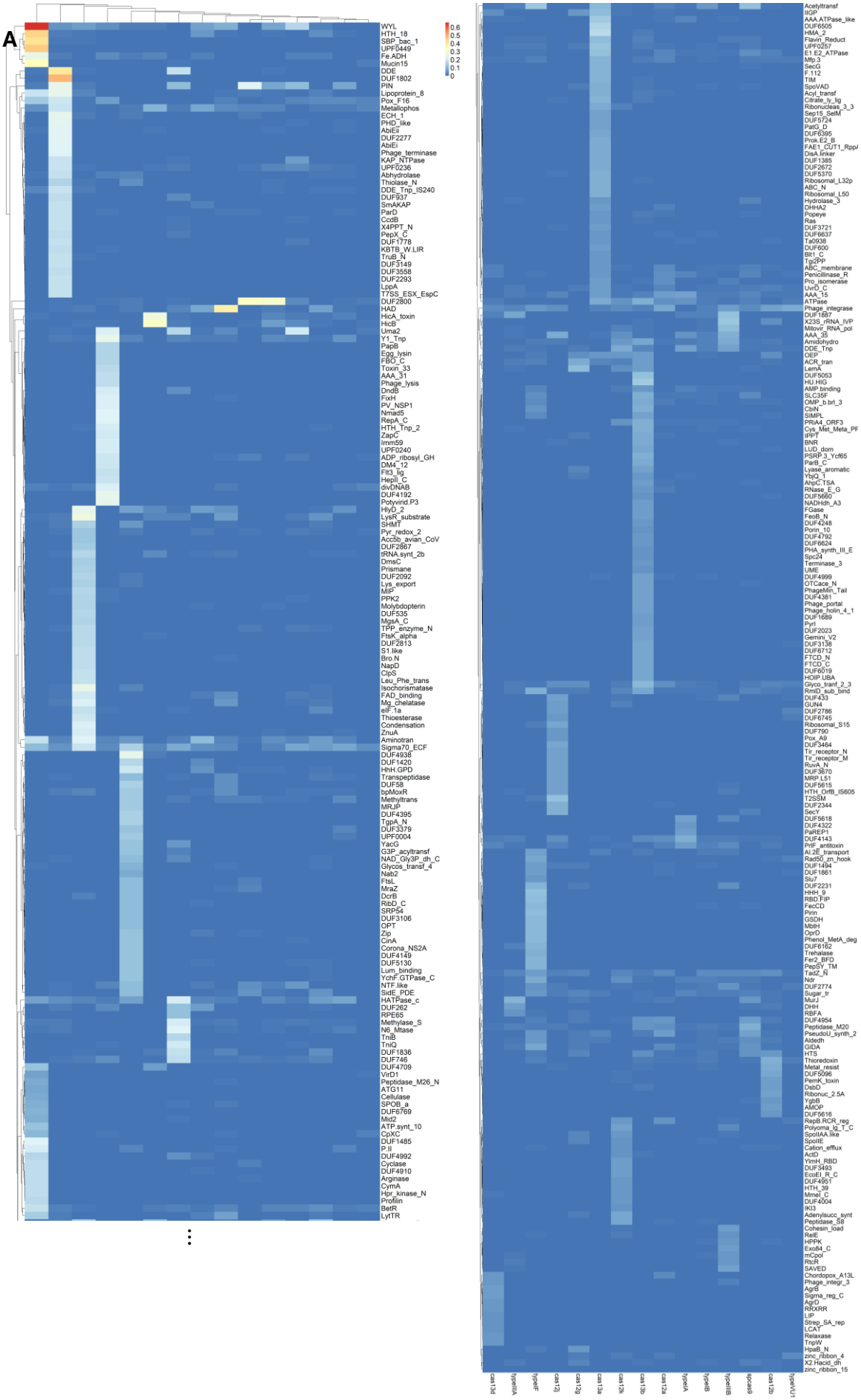


Figure S3.4: Survey of CRISPR-associated proteins using the DEFLOC database for annotation. (A)*Each CRISPR-associated protein was represented in (distance, CRISPRicity, abundance) space and labelled by category, which was derived from protein sequence annotation. (B) the proportions of CRISPR-associated proteins were quantified by total abundance. Abundances <1% were absent from the chart.*

Subtype	Type V-A	Type V-B	Type V-G	Type V-H	Type V-I	Type V-J	Type V-K	Type VI-A	Type VI-B
no. contigs	4167	644	74	7	22	164	315	185	952
Subtype	Type VI-D	Type I-A	Type I-B	Type I-F	Type III-A	Type III-B	Type V-U1	Type II-A	
no. contigs	723	599	1375	8098	13653	359	523	18918	

Table S3.4: ***Number of contigs used for heatmap generation (Figures 4, S6) by CRISPR-Cas subtype***


A
0.6
0.5
0.4
0.3
0.2
0.1
0
WYL
HTH_18
SBP_bac_1
UPF0449
Fe.ADH
Mucin15
DDE
DUF1802
PIN
Lipoprotein_8
Pox_F16
Metallophos
ECH_1
PHD_like
AbiEii
DUF2277
AbiEi
Phage_terminase
KAP_NTPase
UPF0236
Abhydrolase
Thiolase_N
DDE_Tnp_IS240
DUF937
SmAKAP
ParD
CcdB
X4PPT_N
PepX_C
DUF1778
KBTB_W.LIR
TruB_N
DUF3149
DUF3558
DUF2293
LppA
T7SS_ESX_EspC
DUF2800
HAD
HicA_toxin
HicB
Uma2
Y1_Tnp
PaiB
Egg_lysin
FBD_C
Toxin_33
AAA_31
Phage_lysis
DndB
FixH
PV_NSP1
Nmad5
RepA_C
HTH_Tnp_2
ZapC
Imm59
UPF0240
ADP_ribosyl_GH
DM4_12
Flt3_lig
HepII_C
divDNAB
DUF4192
Potyvirid.P3
HlyD_2
LysR_substrate
SHMT
Pyr_redox_2
Acc5b_avian_CoV
DUF2867
tRNA.synt_2b
DmsC
Prismane
DUF2092
Lys_export
MIP
PPK2
Molybdopterin
DUF535
MgsA_C
TPP_enzyme_N
FtsK_alpha
DUF2813
S1.like
Bro.N
NapD
ClpS
Leu_Phe_trans
Isochorismatase
FAD_binding
Mg_chelatase
eIF-1a
Thioesterase
Condensation
ZnuA
Aminotran
Sigma70_ECF
DUF4938
DUF1420
HhH.GPD
Transpeptidase
DUF58
bpMovR
Methyltrans
MRJP
DUF4395
TgpA_N
DUF3379
UPF0004
YacG
G3P_acyltransf
NAD_Gly3P_dh_C
Glycos_transf_4
Nab2
FtsL
MraZ
DcrB
RibD_C
SRP54
DUF3106
OPT
Zip
CinA
Corona_NS2A
DUF4149
DUF5130
Lum_binding
YchF.GTPase_C
NTF.like
SidE_PDE
HATPase_c
DUF262
RPE65
Methylase_S
N6_Mtase
TniB
TniQ
DUF1836
DUF746
DUF4709
VirD1
Peptidase_M26_N
ATG11
Cellulase
SPO8_a
DUF6769
Mid2
ATP.synt_10
CpXC
DUF1485
P.II
DUF4992
Cyclase
DUF4910
Arginase
CymA
Hpr_kinase_N
Profilin
BetR
LytTR
Acetyltransf
IIGP
AAA.ATPase_like
DUF6505
HMA_2
Flavin_Reduct
UPF0257
E1.E2_ATPase
Mfp_3
SecG
F.112
TIM
SpoVAD
Acyl_transf
Citrate_ly_lig
Ribonucleas_3_3
Sep15_SelM
DUF5724
PatG_D
DUF6395
Prok.E2_B
FAE1_CUT1_RppA
DisA.linker
DUF1385
DUF2672
DUF5370
Ribosomal_L32p
ABC_N
Ribosomal_L50
Hydrolase_3
DHHA2
Popeye
Ras
DUF3721
DUF6637
Tat098
DUF600
Bit1_C
Tgt2PP
ABC_membrane
Penicillinase_R
Pro_isomerase
UvrD_C
AAA_15
ATPase
Phage_integrase
DUF1887
X23S_rRNA_IVP
Mitovir_RNA_pol
AAA_35
Amidohydro
DDE_Tnp
OEP
ACR_tran
LemA
DUF5053
HU.HIG
AMP.binding
SLC35F
OMP_b.brl_3
ClpN
SIMPL
PRiA4_ORF3
Cys_Met_Meta_PP
dPPT
BNR
LUD_dom
PSRP.3_Ycf65
ParB_C
Lyase_aromatic
YbjQ_1
AhpC.TSA
RNase_E_G
DUF5660
NADHdh_A3
FGase
FeoB_N
DUF4248
Porin_10
DUF4792
DUF6624
PHA_synth_III_E
Spc24
Terminase_3
UME
DUF4999
OTCace_N
PhageMin_Tail
DUF4381
Phage_portal
Phage_holin_4_1
DUF1689
Pyrl
DUF2023
Gemini_V2
DUF3138
DUF6712
FTCD_N
FTCD_C
DUF6019
HOIP.UBA
Glyco_tranf_2_3
RmlD_sub_bind
DUF433
GUN4
DUF2786
DUF6745
Ribosomal_S15
DUF790
Pox_A9
DUF3464
Tir_receptor_N
Tir_receptor_M
RuvA_N
DUF3670
MRP.L51
DUF5615
HTH_Orf8_IS605
T2SSM
DUF2344
SecY
DUF5618
DUF4322
PaREP1
DUF4143
PrlF_antitoxin
AI.2E_transport
Rad50_zn_hook
DUF1494
DUF1861
Slu7
DUF2231
HHH_9
RBD.FIP
FecCD
Pirin
GSDH
MbtH
OprD
Phenol_MetA_deg
DUF6162
Trehalase
Fer2_BFD
PepSY_TM
TadZ_N
Ndr
DUF2774
Sugar_tr
MurJ
DHH
RBFA
DUF4954
Peptidase_M20
PseudoU_synth_2
Aldedh
GIDA
HTS
Thioredoxin
Metal_resist
DUF5096
PemK_toxin
DsbD
Ribonuc_2.5A
YgbB
AlkD
DUF5616
RepB.RCR_reg
Polyoma_lg_T_C
SpoIIAA.like
SpoIIE
Cation_efflux
ActD
YlmH_RBD
DUF3493
EcoEI_R_C
DUF4951
HTH_39
Mmel_C
DUF4004
IK03
Adenylsucc_synt
Peptidase_S8
Cohesin_load
RelE
HPPK
Exo84_C
mCpol
RtcR
SAVED
Chordopox_A13L
Phage_integr_3
AgrB
Sigma_reg_C
AgrD
RRXRR
LIP
Strep_SA_rep
LCAT
Relaxase
TnpW
HpaB_N
zinc_ribbon_4
X2.Hacid_dh
zinc_ribbon_15
cas13d
typeIIIA
typeIF
cas12g
cas13a
cas12k
cas13b
cas12a
typeIA
typeIB
typeIIIB
spoIIIE
cas12b
typeIV.U1

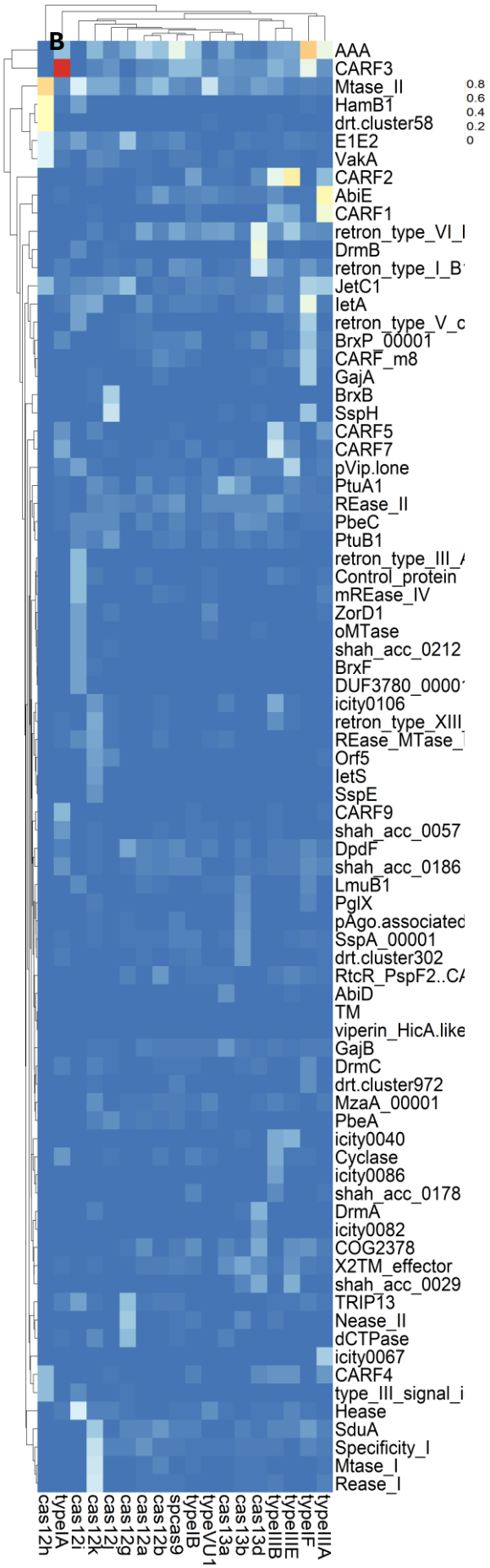
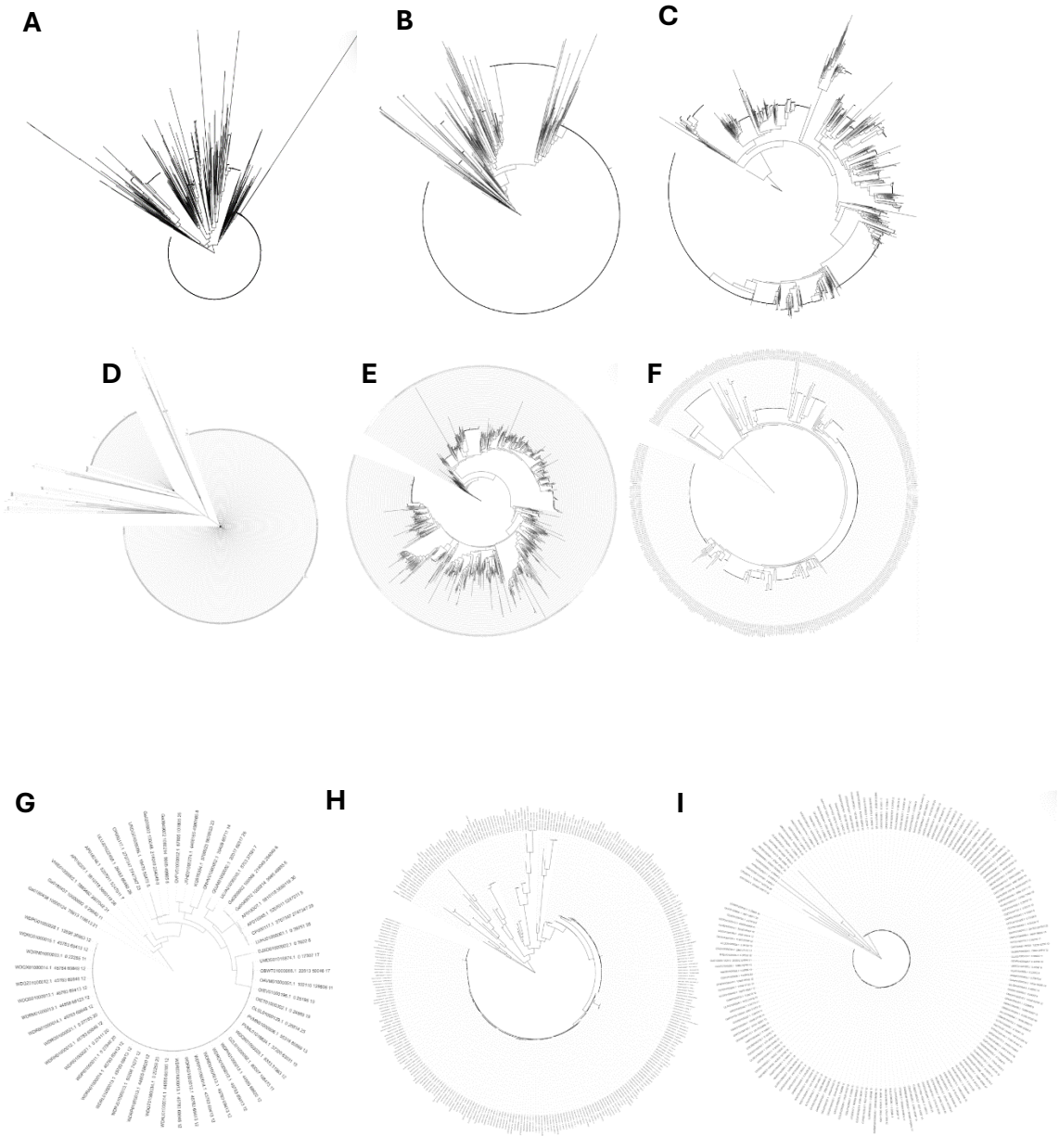
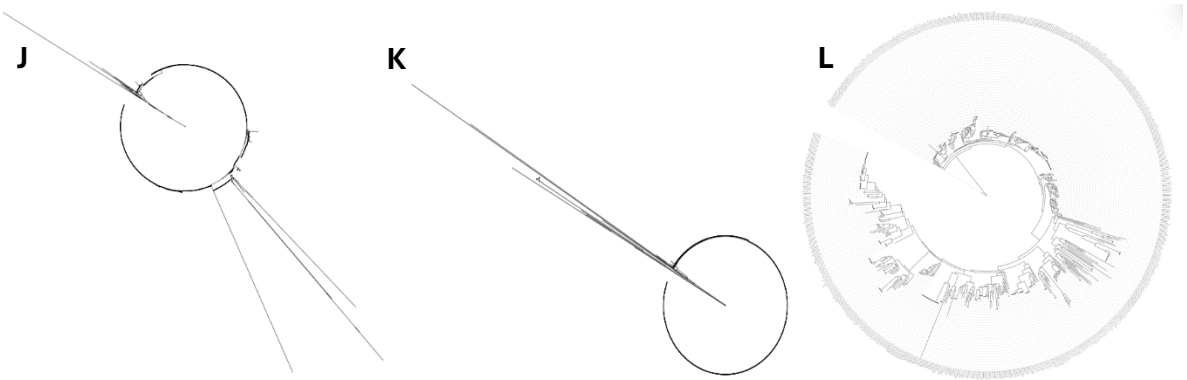
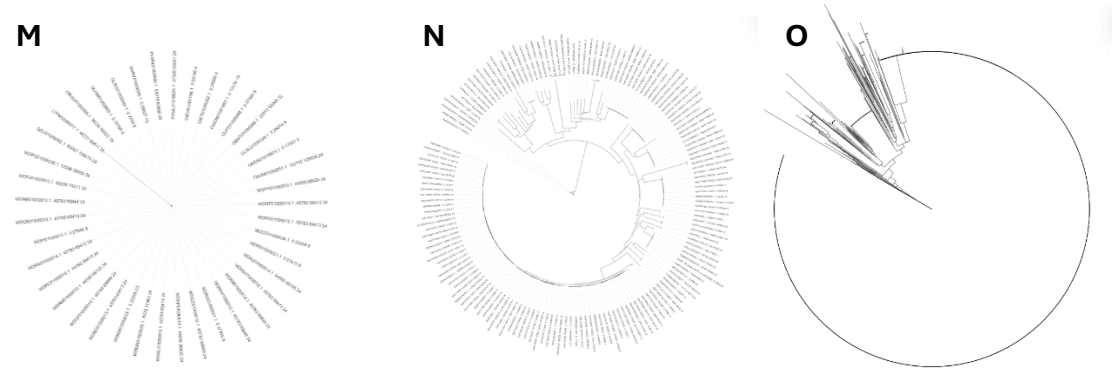
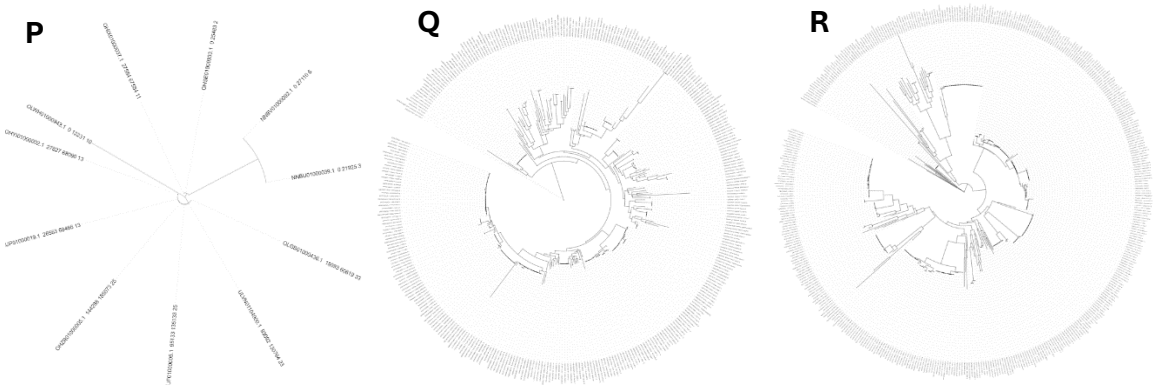


***Figure S3.5:* Conservation score calculation for all CRISPR-Cas subtypes.** *Entries matching core CRISPR-Cas system proteins were excluded. Annotation was performed using A) Pfam* (top page)*, B) DEFLOC*

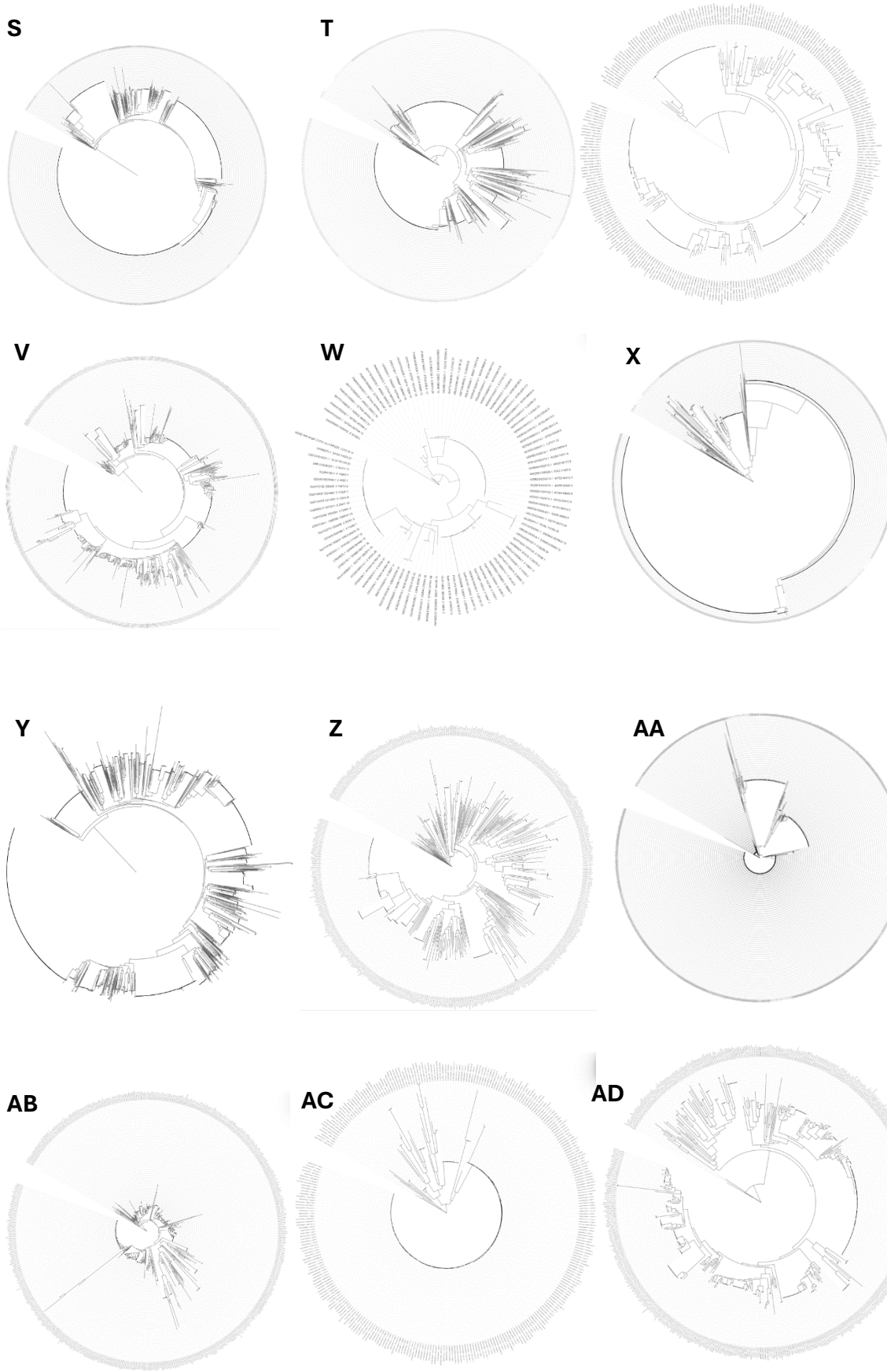

A
B
C
D
E
F
G
H
I





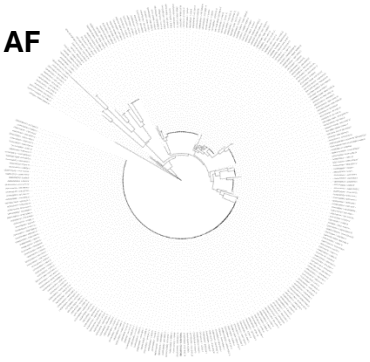


U


S
T
V
W
X
Y
Z
AA
AB
AC
AD

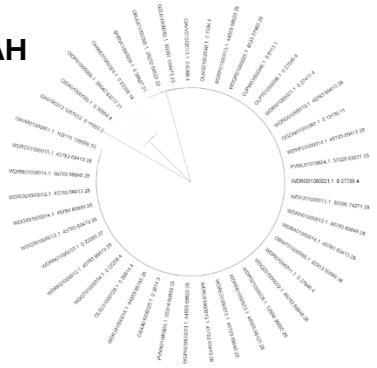
AE

AF

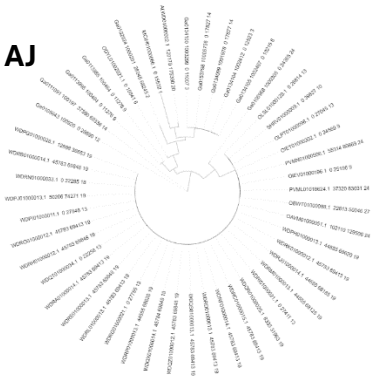


AG

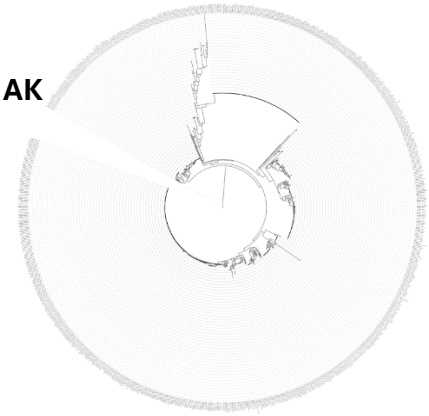
AH



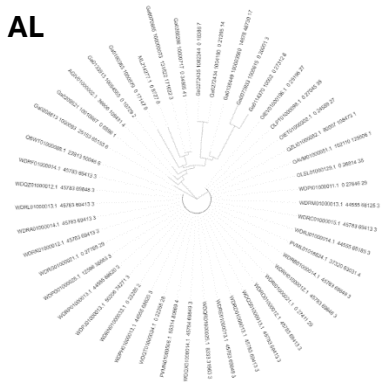
AJ



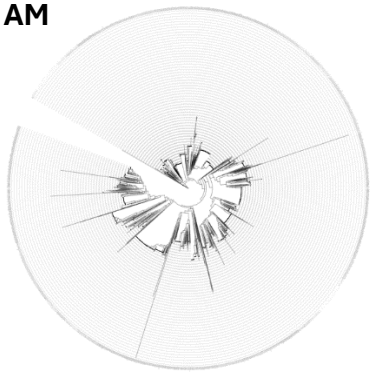
AK



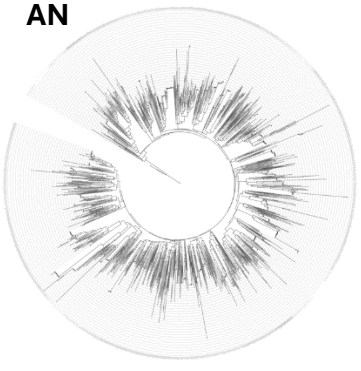
AL



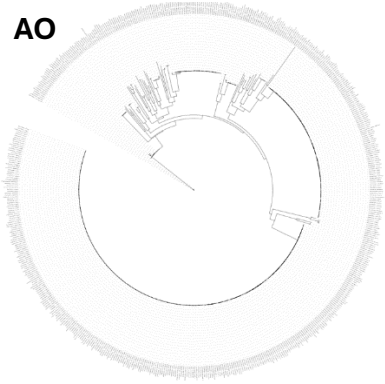
AM



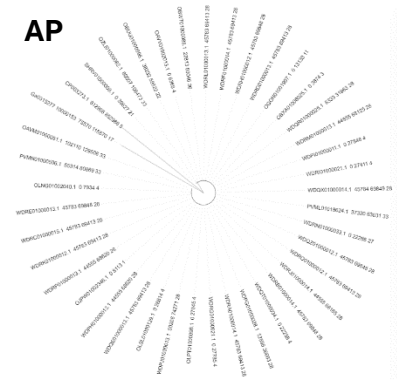
AN



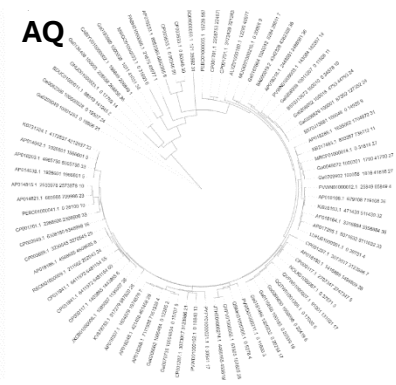
AO



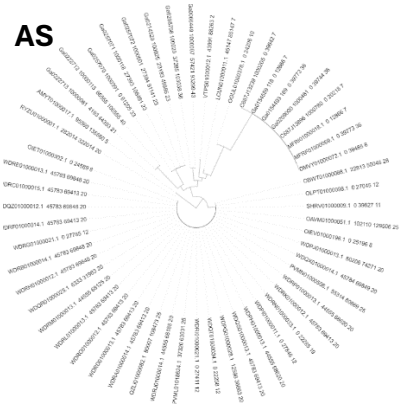
AP

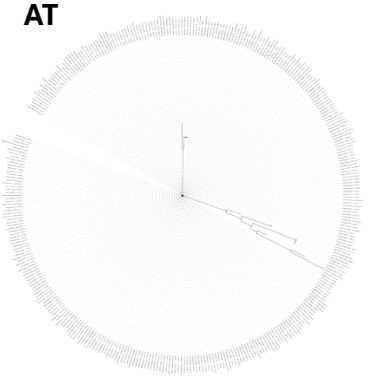


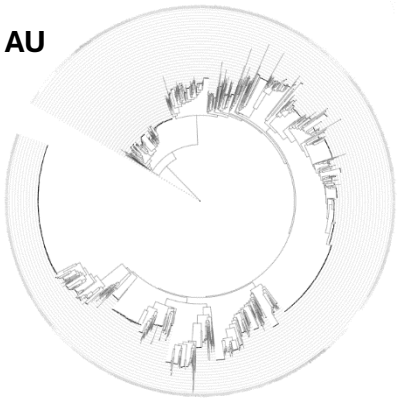
AQ

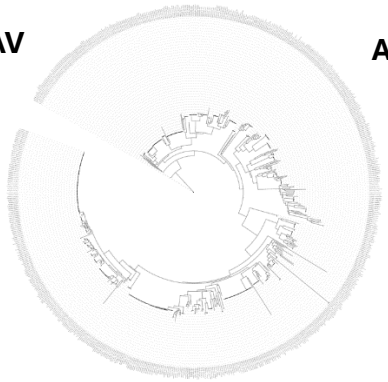


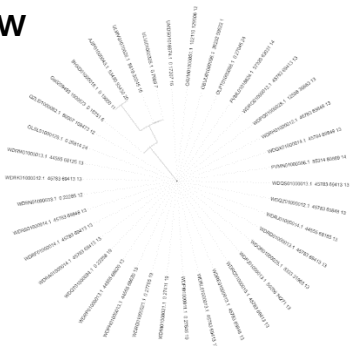
AR

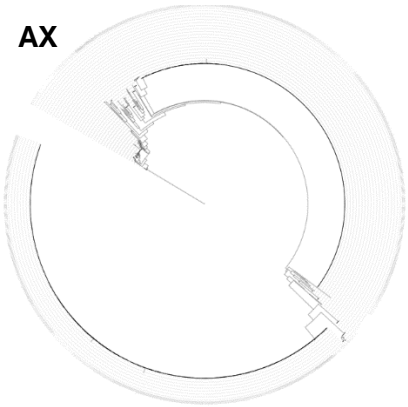

AS

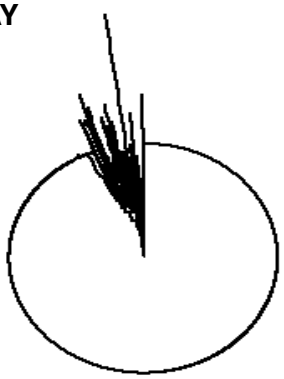

AT


AU


AV


AW


AX


AY

AZ

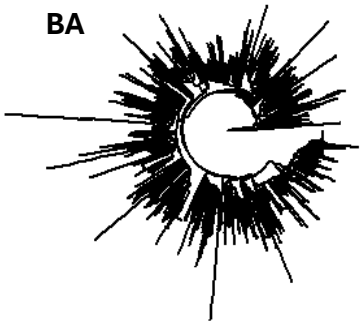

BA

Figure S3.6: ***CRISPR-associated protein candidate phylogenies.*** *Each tree was generated via iqtree2[192] and rendered using the circular format in Interactive Tree of Life (ITOL) web server [266]. The designation as to whether each protein family corresponding to each tree was thought to be a distinct group of related homologs with a large enough effective sample size is given in table S3.5. Some extremely abundant proteins were omitted as a tree was too large to be generated. These proteins are likely to be false positives regardless as a consequence of their high general abundance. (A) AbiEi, (B) AbiEii, (C) Aminotran, (D) Condensation, (E) DDE, (F) divDNAB, (G) DndB, (H) DrmB, (I) DUF1485, (J) DUF1802, (K) DUF2277, (L) DUF2800, (M) ECH_1, (N) FAD_binding, (O) FixH, (P) HATPase_c, (Q) HicA_toxin, (R) HicB, (S) Hlyd_2, (T) HTH_18, (U) HTH_Tnp_2, (V) ietA, (W) Imm59, (X) Isochromatase, (Y) LysR_substrate, (Z) Methylase_S, (AA) Mg_Chelatase, (AB) Mtase_I, (AC) Mucin15, (AD) N6_Methylase, (AE) nmad5 ,(AF) P-II, (AG) PIN, (AH) Potyvirid-P3., (AJ) PV_NSP1, (AK) Rease_I, (AL) RepA_C, (AM) Retron_type_I_B1_atpase_toprim_cluster22_2, (AN) retron_type_VI_hth_cluster2, (AO) SBP_bac_1, (AP) SspH, (AQ) tniQ, (AR) Uma2, (AS) UPF0240, (AT) UPF0449, (AU) WYL, (AV) Y1_tnp, (AW) ZapC, (AX) ZnuA, (AY) PHD-like (rendered by ggtree as ITOL could not render branch nodes), (AZ) Phage Terminase (99% duplicate sequences) (rendered by ggtree as ITOL could not render branch nodes), (BA) CARF3—CARF(rendered by ggtree as ITOL could not render branch nodes).*

Pfam		**Padloc**	
Protein	**Unified clade**	**Protein**	**Unified clade**
AbiEi	FALSE	AAA_21	TRUE
AbiEii	TRUE	AbiEi	FALSE
Aminotran	TRUE	AbiEii	TRUE
Condensation	FALSE	CARF3..CARF_MK_CARF_01001287	TRUE
DDE	TRUE	DrmB	TRUE
DUF1485	FALSE	letA	TRUE
DUF1802	FALSE	Mtase_I	TRUE
DUF2277	FALSE	Rease_I	TRUE
DUF2800	TRUE	SspH_00003	FALSE
DUF4192	FALSE	retron_type_I_B1_atpase_toprim_cluster22_2	TRUE
DUF4938	TRUE	retron_type_VI_hth_cluster2_1	TRUE
DndB	FALSE		
ECH_1	TRUE		

FAD_binding	FALSE		
Fe.ADH	FALSE		
FixH	FALSE		
HATPase_c	FALSE		
HTH_18	TRUE		
HTH_Tnp_2	TRUE		
HicA_toxin	TRUE		
HicB	TRUE		
HlyD_2	TRUE		
Imm59	TRUE		
Isochorismatase	TRUE		
Lipoprotein_8	FALSE		
LysR_substrate	TRUE		
Methylase_S	TRUE		
Mg_chelatase	FALSE		
Mucin15	FALSE		
N6_Mtase	TRUE		
Nmad5	FALSE		
P.II	TRUE		
PHD_like	FALSE		
PIN	TRUE		
PV_NSP1	FALSE		
Phage_terminase	FALSE		
Potyvirid.P3	FALSE		
RepA_C	FALSE		
SBP_bac_1	TRUE		
TniQ	TRUE		
UPF0240	FALSE		
UPF0449	FALSE		
Uma2	TRUE		
WYL	TRUE		
Y1_Tnp	TRUE		
ZapC	FALSE		
ZnuA	FALSE		
divDNAB	TRUE		

Table S3.5: ***Designation of whether each putative CRISPR-asscociated protein corresponded to a single family of orthologs with an effective (non-redundant sample size > 12).***

Supplementary data for Chapter 4

Network based characterization of intra-subtype diversity of host-MGE interactions at the gene cluster level.

Table S4.1: BLAST query characteristics:

Subtype	Protein name	Query ID	Database	Sample Origin	Species
Type V-A	Cas12a	sp|U2UMQ6|CS12A_ACISB	Protein Data Bank	Cultured	Acidoaminococcus (strain BV3L6)
Type V-B	Cas12b	sp_T0D7A2_CS12B_ALIAG	Protein Data Bank	Cultured	Alicyclobacillus acidoterrestris (strain ATCC 49025 / DSM 3922 / CIP 106132 / NCIMB 13137 / GD3B)
Type V-F	NA (TnpB-like)	WP_000109044.1	Protein-nr	Cultured	*Bacillus Thuringiensis*
Type VI-A	Cas13a	NA (pipeline derived) ,WP_021746774.1 (99% identity)	10TB metagenome block (chapter 3)	Metagenome	*Leptotrichia Wadei*
Type VI-B	Cas13b	WP_044065294.1	Protein-nr	Cultured	*Prevotella sp.*
Type VI-D	Cas13d	WP_075424065.1	Protein-nr	Metagenome	*Ruminococcus sp. XPD3002*
Type I-A	Cas7a	tr_A0A0E3GVE3_A0A0E3GVE3_SACSO	Uniprot	Cultured	*Saccharolobus solfataricus*
Type I-B	Cas8b	tr|A0A143MKY0|A0A143MKY0_9BACI	Uniprot	Cultured	*Geobacillus subterraneus*
Type I-D	Cas10d	tr|A0A8J7A0A1|A0A8J7A0A1_9CYAN	Uniprot	Cultured	*Fortiea sp.*
Type I-F	Csy1	tr|A0A396TZS3|A0A396TZS3_9GAMM	Uniprot	Cultured	*Colwellia sp.*
Type III-A	Cas10	sp|A0A0A7HFE1|CAS10_STRTR	Uniprot	Cultured	*Streptococcus thermophilus*
Type III-B	Cmr2	tr|A0A8J7APQ0|A0A8J7APQ0_9CYAN	Uniprot	Cultured	*Romeria aff. gracilis LEGE 07310*

Type II	Cas9	sp|Q99ZW2|CAS9_STRP1	Uniprot	Cultured	*Streptococcus pyogenes serotype M1*

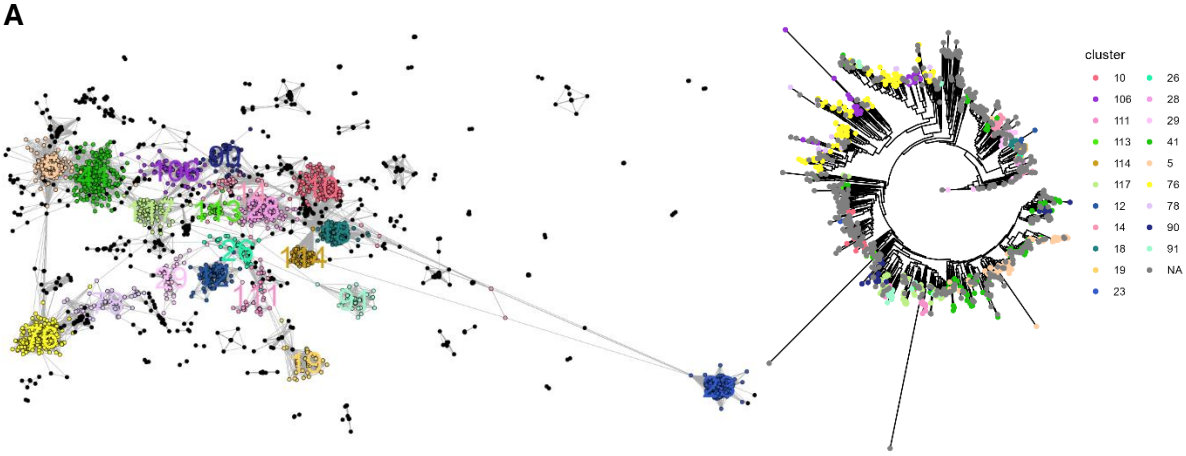
Table S4.2: Parameters used to generate and visualize host and mapped sequence vCONTACT2 networks.

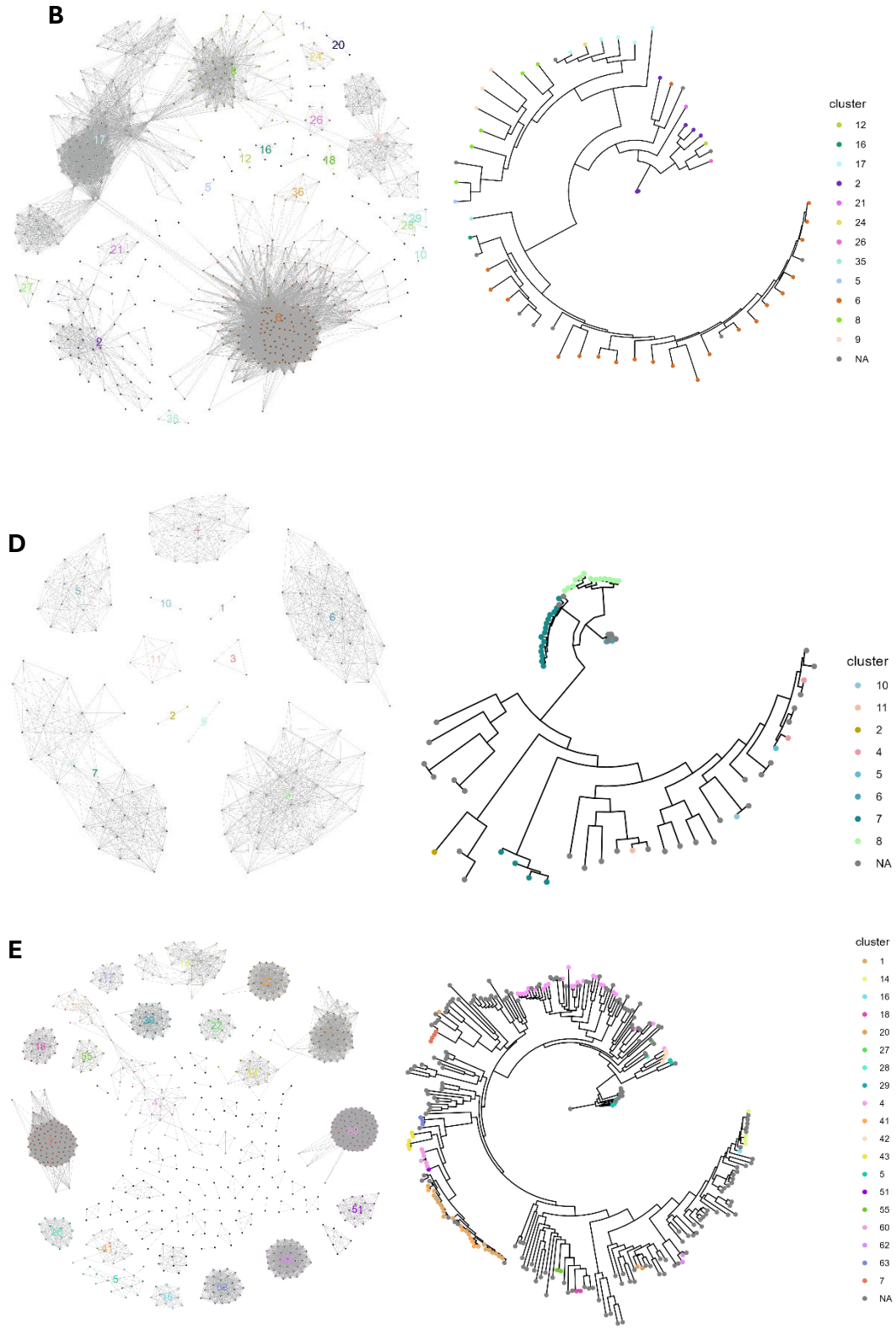
Subtype	Set.edge.attribute value	Leiden resolution parameter
Type V-A	0.05	0.01
Type V-B	0.01	0.01
Type V-F1	0.01	0.01
Type VI-A	0.1	0.01
Type VI-B	0.1	0.01
Type VI-D	0.01	0.01
Type II	0.0001	0.01
Type I-A	0.01	0.01
Type I-B	0.01	0.01
Type I-D	0.01	0.01
Type I-F	0.01	0.01
Type III-A	0.001	0.01
Type III-B	0.01	0.01

Table S4.3: Host-Phage visualization parameters:

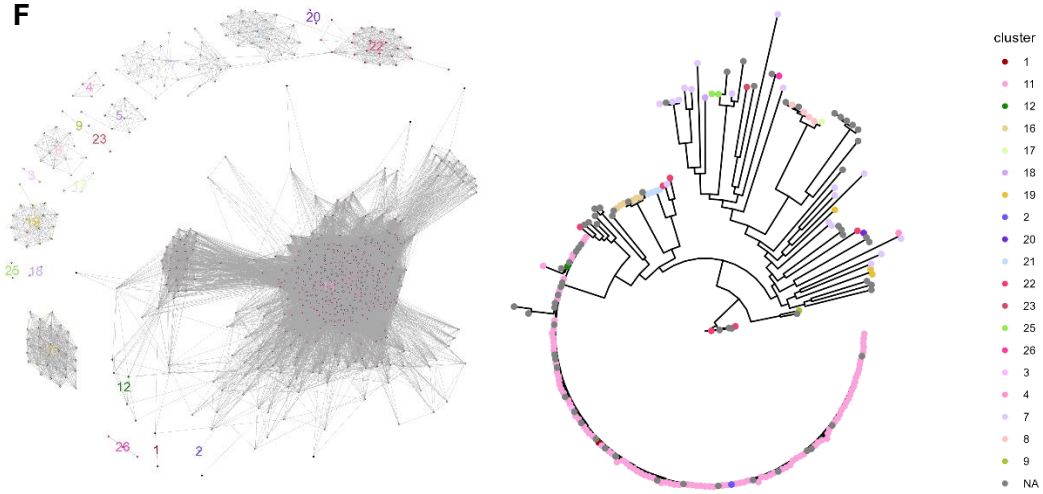
Subtype	Set.edge.attribute value
Type V-A	0.001
Type V-B	1
Type V-F1	0.01
Type VI-A	1
Type VI-B	0.005
Type VI-D	1
Type I-A	0.1
Type I-B	0.0001
Type I-D	0.5
Type I-F	0.01
Type III-A	0.0005
Type III-B	1

A

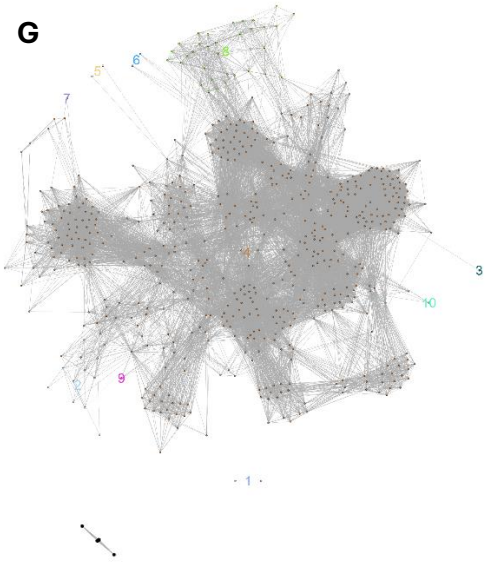

cluster
10
106
111
113
114
117
12
14
18
19
23
26
28
29
41
5
76
78
90
91
NA

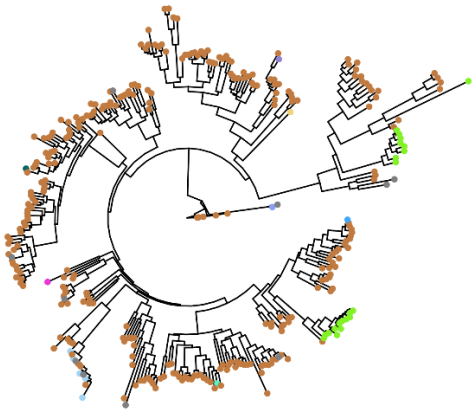

B
cluster
12
16
17
2
21
24
26
35
5
6
8
9
NA
D
cluster
10
11
2
4
5
6
7
8
NA
E
cluster
1
14
16
18
20
27
28
29
4
41
42
43
5
51
55
60
62
63
7
NA

F

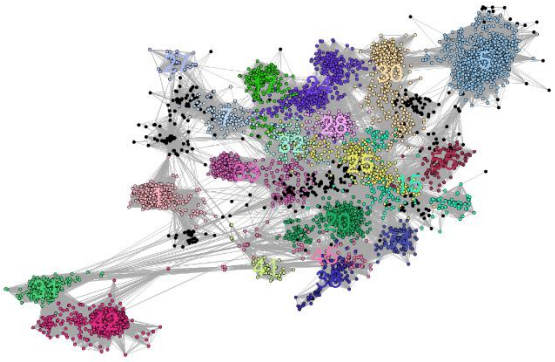

20
22
4
5
9
23
25
18
12
26
1
2
cluster
1
11
12
16
17
18
19
2
20
21
22
23
25
26
3
4
7
8
9
NA

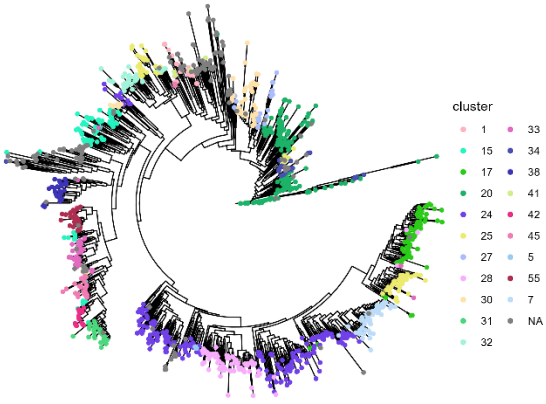
G



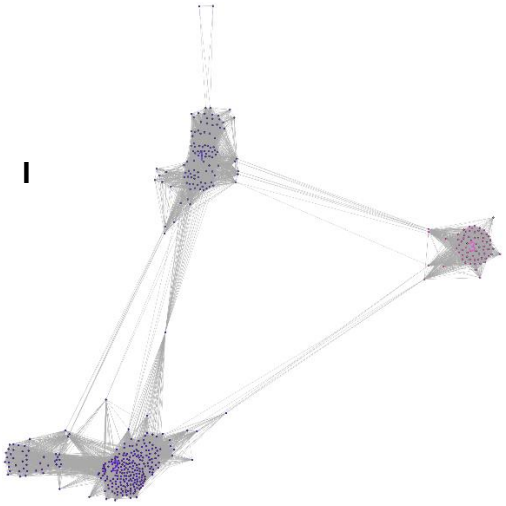

cluster
1
10
2
3
4
5
6
7
8
9
NA

H




cluster
1
15
17
20
24
25
27
28
30
31
32
33
34
38
41
42
45
5
55
7
NA

I




cluster
1
2
3
NA

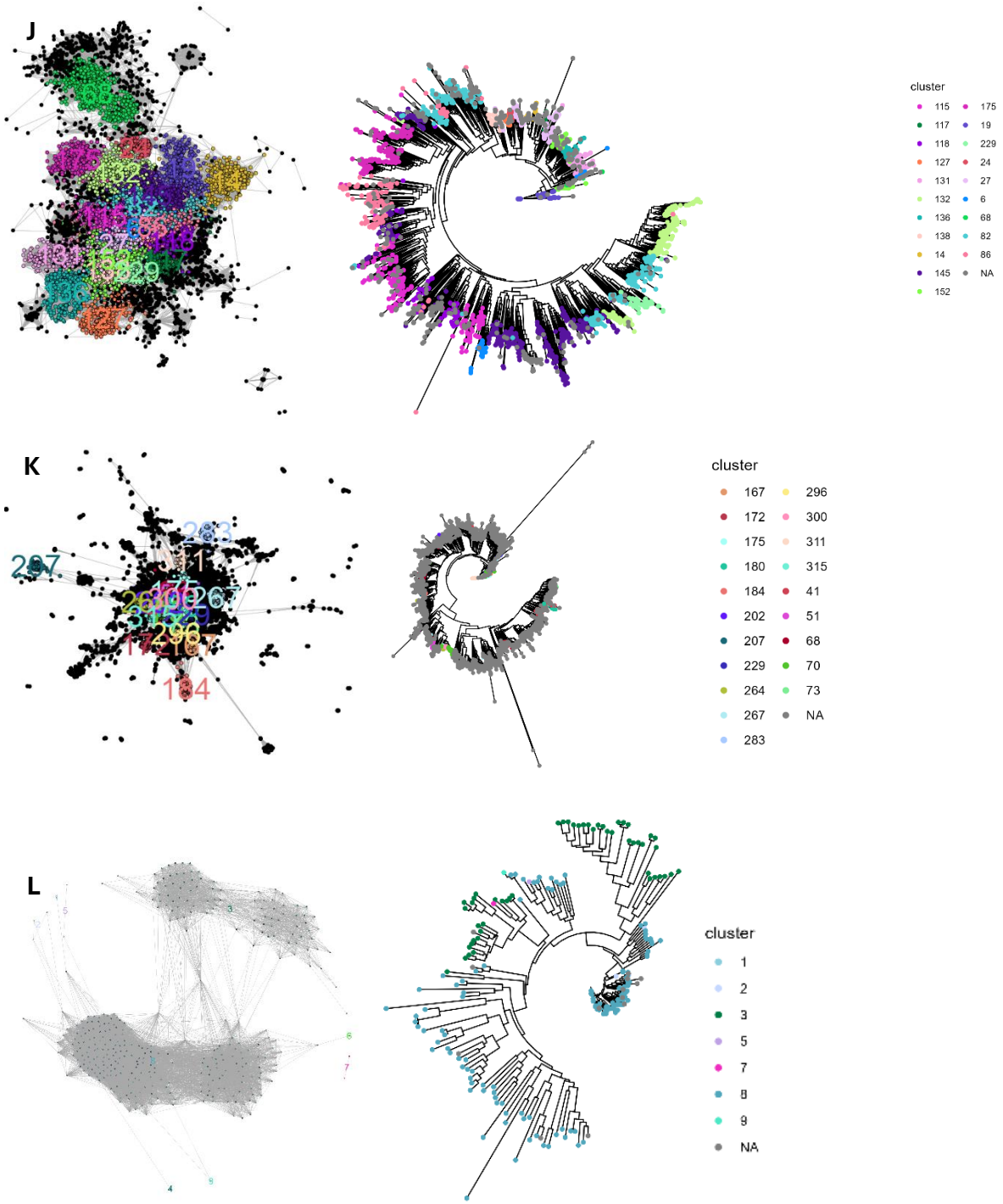
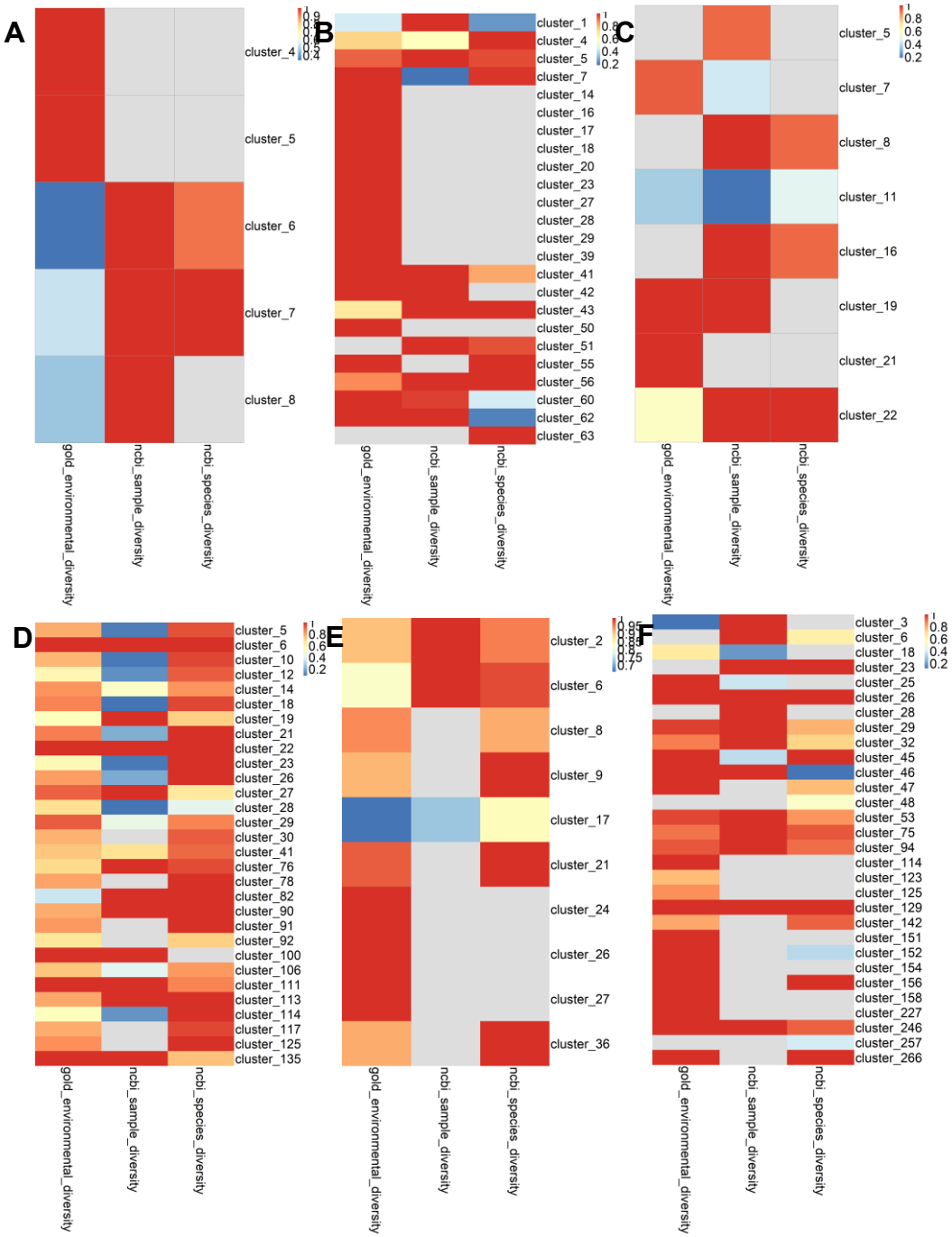
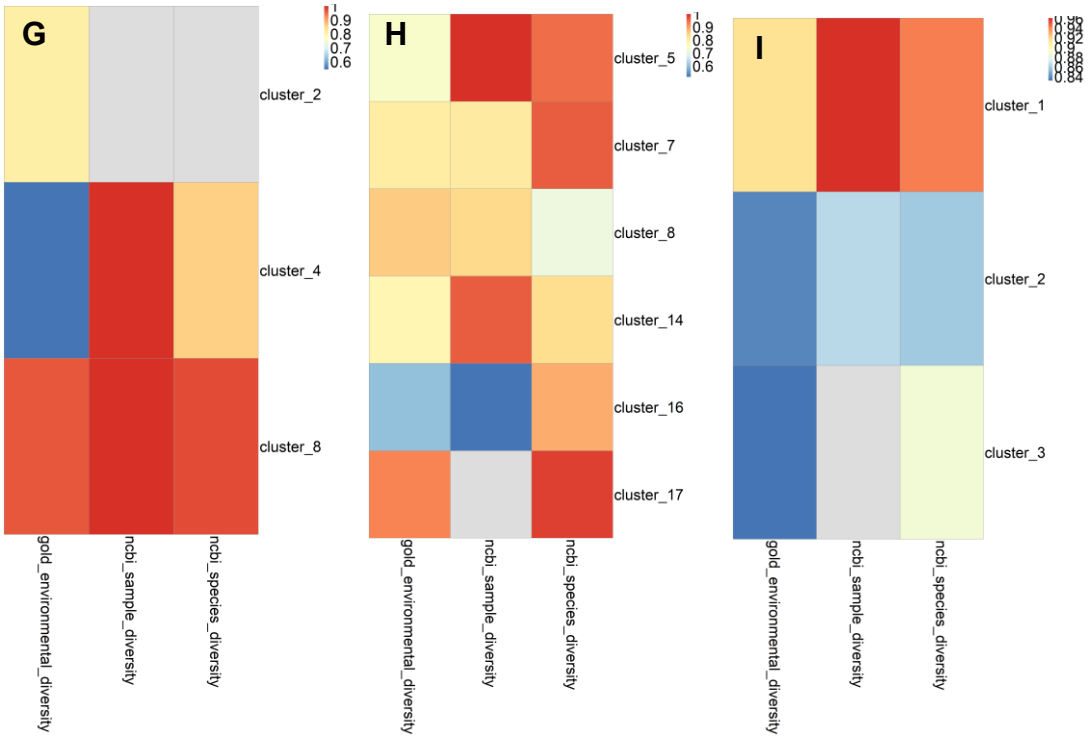
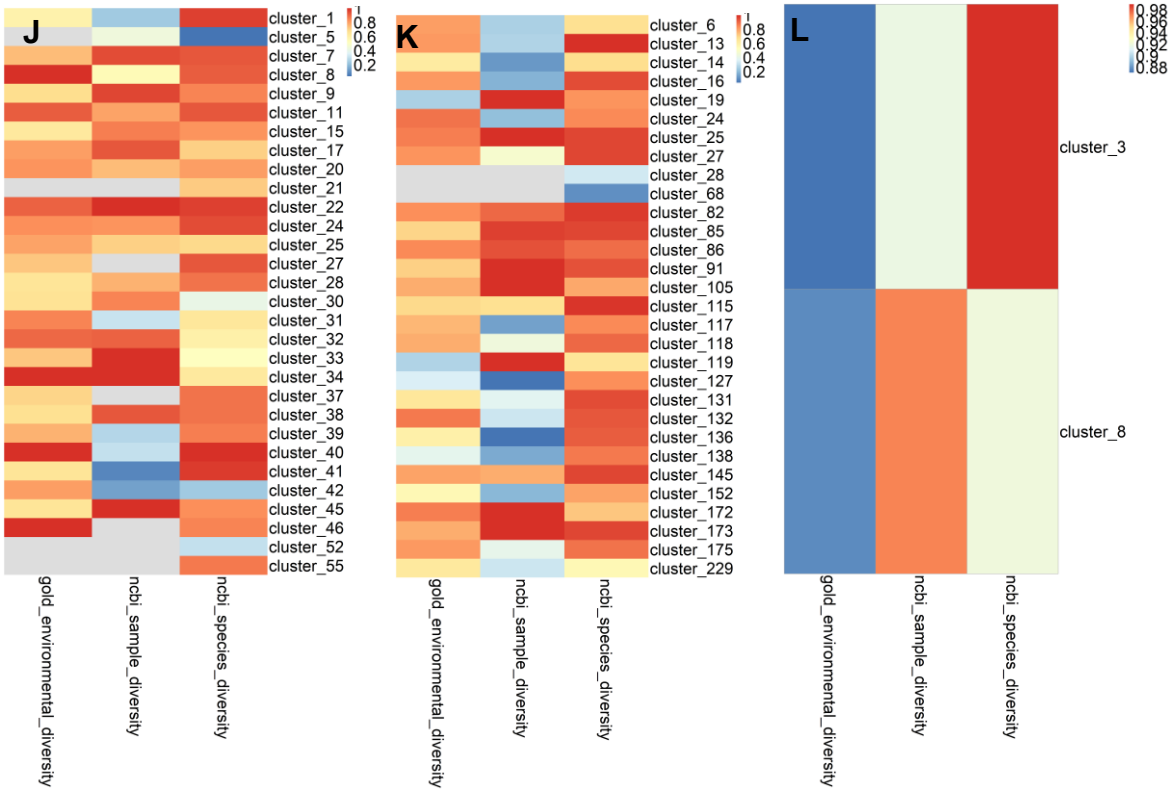


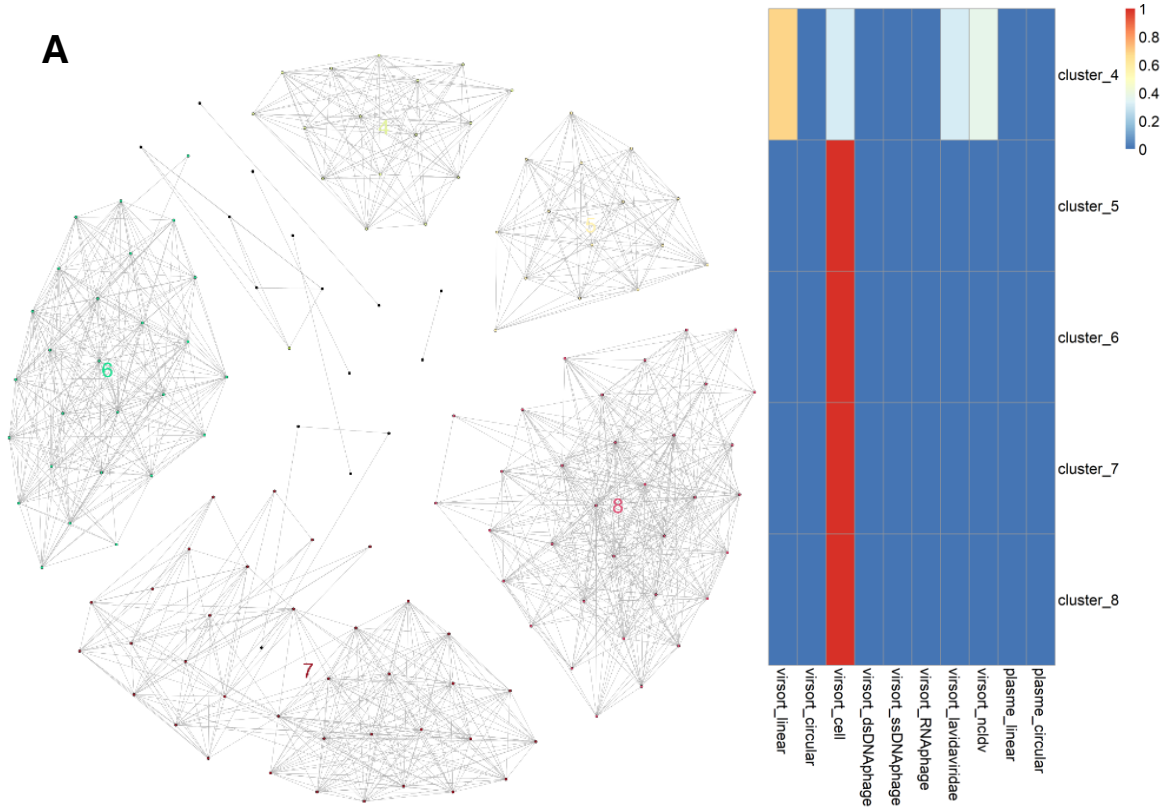
Figure S4.1: ***Comparison of host-encoded CRISPR-Cas subtype sequence differentiation by vConTACT2 gene cluster networks and signature gene phylogenetic trees***. *A) Type V-A, B) Type V-B, C) Type V-F1, D) Type VI-A, E) Type VI-B, F) Type VI-D, G) Type I-A, H) Type I-B, I) Type I-D, J) Type I-F, K) Type III-A, L) Type III-B. In type III-A and I-F systems sequences were pre-clustered at 70% similarity and a single representative sequence taken using mmseqs2*[187] prior to tree generation via IQtree2.

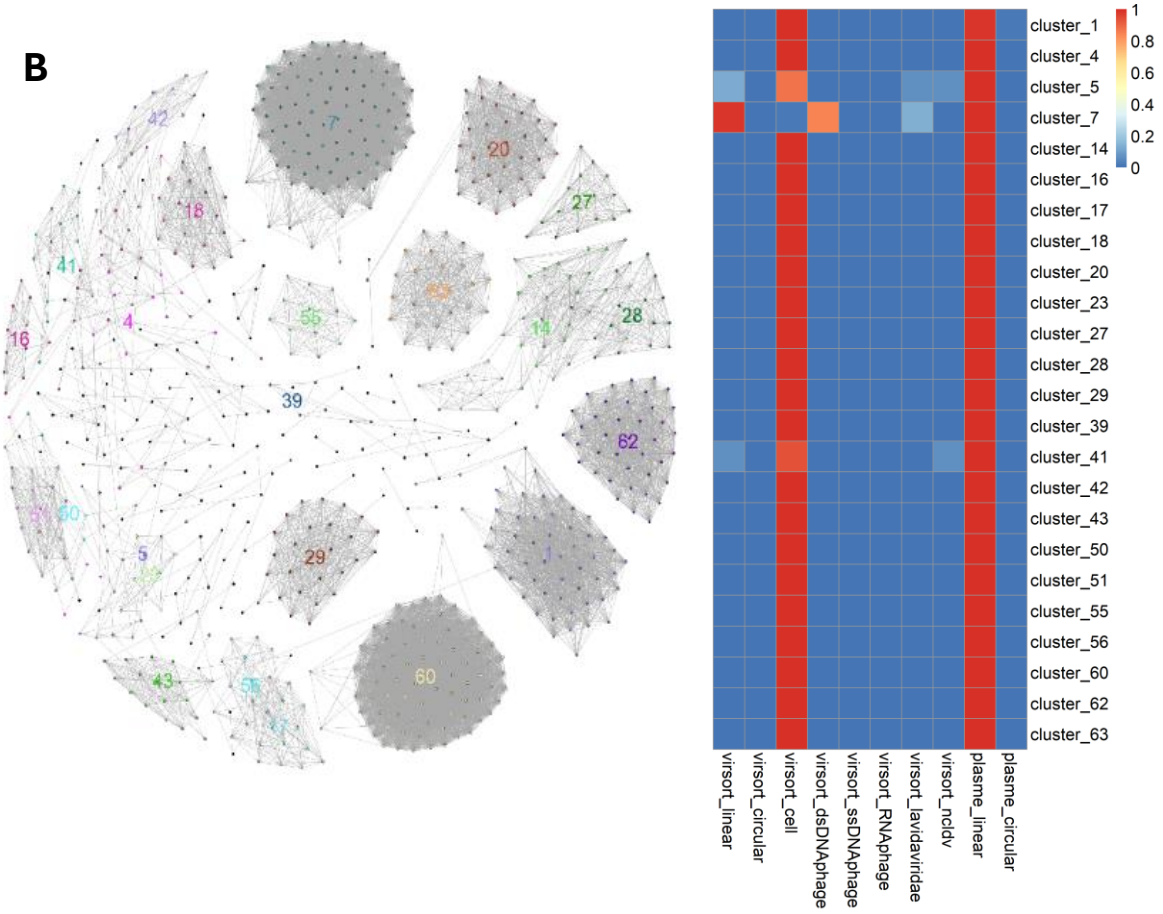

A
cluster_4
cluster_5
cluster_6
cluster_7
cluster_8
gold_environmental_diversity
ncbi_sample_diversity
ncbi_species_diversity
B
cluster_1
cluster_4
cluster_5
cluster_7
cluster_14
cluster_16
cluster_17
cluster_18
cluster_20
cluster_23
cluster_27
cluster_28
cluster_29
cluster_39
cluster_41
cluster_42
cluster_43
cluster_50
cluster_51
cluster_55
cluster_56
cluster_60
cluster_62
cluster_63
C
cluster_5
cluster_7
cluster_8
cluster_11
cluster_16
cluster_19
cluster_21
cluster_22
D
cluster_5
cluster_6
cluster_10
cluster_12
cluster_14
cluster_18
cluster_19
cluster_21
cluster_22
cluster_23
cluster_26
cluster_27
cluster_28
cluster_29
cluster_30
cluster_41
cluster_76
cluster_78
cluster_82
cluster_90
cluster_91
cluster_92
cluster_100
cluster_106
cluster_111
cluster_113
cluster_114
cluster_117
cluster_125
cluster_135
E
cluster_2
cluster_6
cluster_8
cluster_9
cluster_17
cluster_21
cluster_24
cluster_26
cluster_27
cluster_36
F
cluster_3
cluster_6
cluster_18
cluster_23
cluster_25
cluster_26
cluster_28
cluster_29
cluster_32
cluster_45
cluster_46
cluster_47
cluster_48
cluster_53
cluster_75
cluster_94
cluster_114
cluster_123
cluster_125
cluster_129
cluster_142
cluster_151
cluster_152
cluster_154
cluster_156
cluster_158
cluster_227
cluster_246
cluster_257
cluster_266

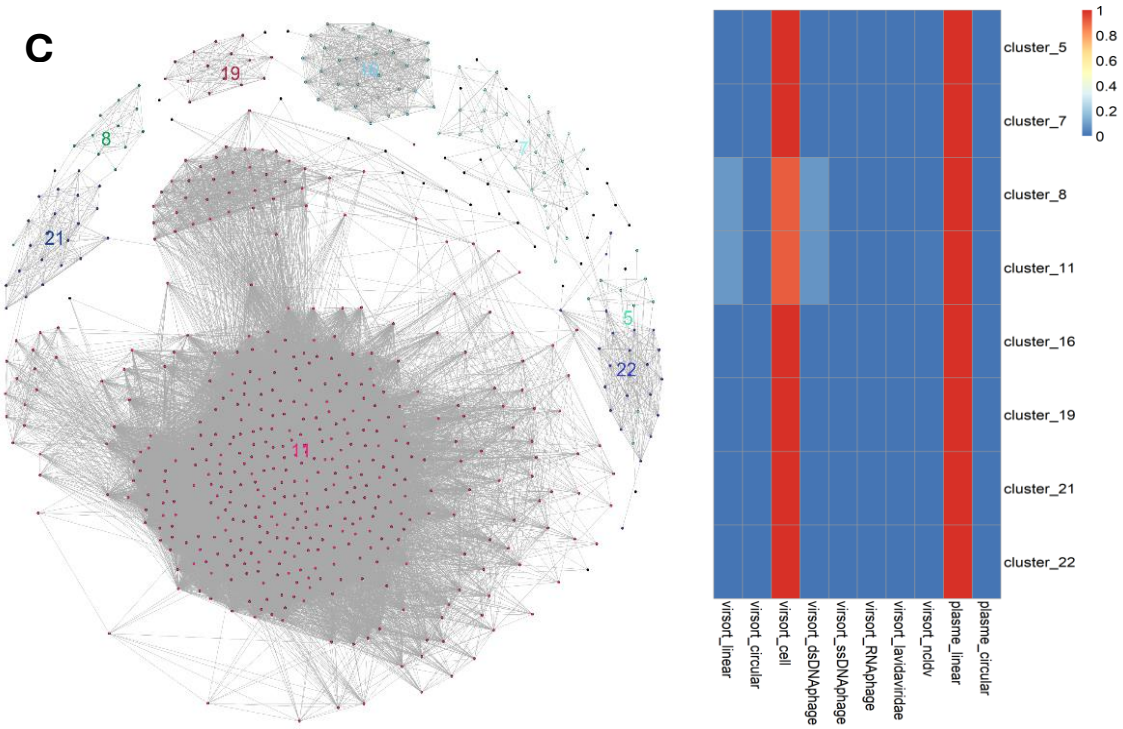

G
cluster_2
cluster_4
cluster_8
gold_environmental_diversity
ncbi_sample_diversity
ncbi_species_diversity
H
cluster_5
cluster_7
cluster_8
cluster_14
cluster_16
cluster_17
I
cluster_1
cluster_2
cluster_3

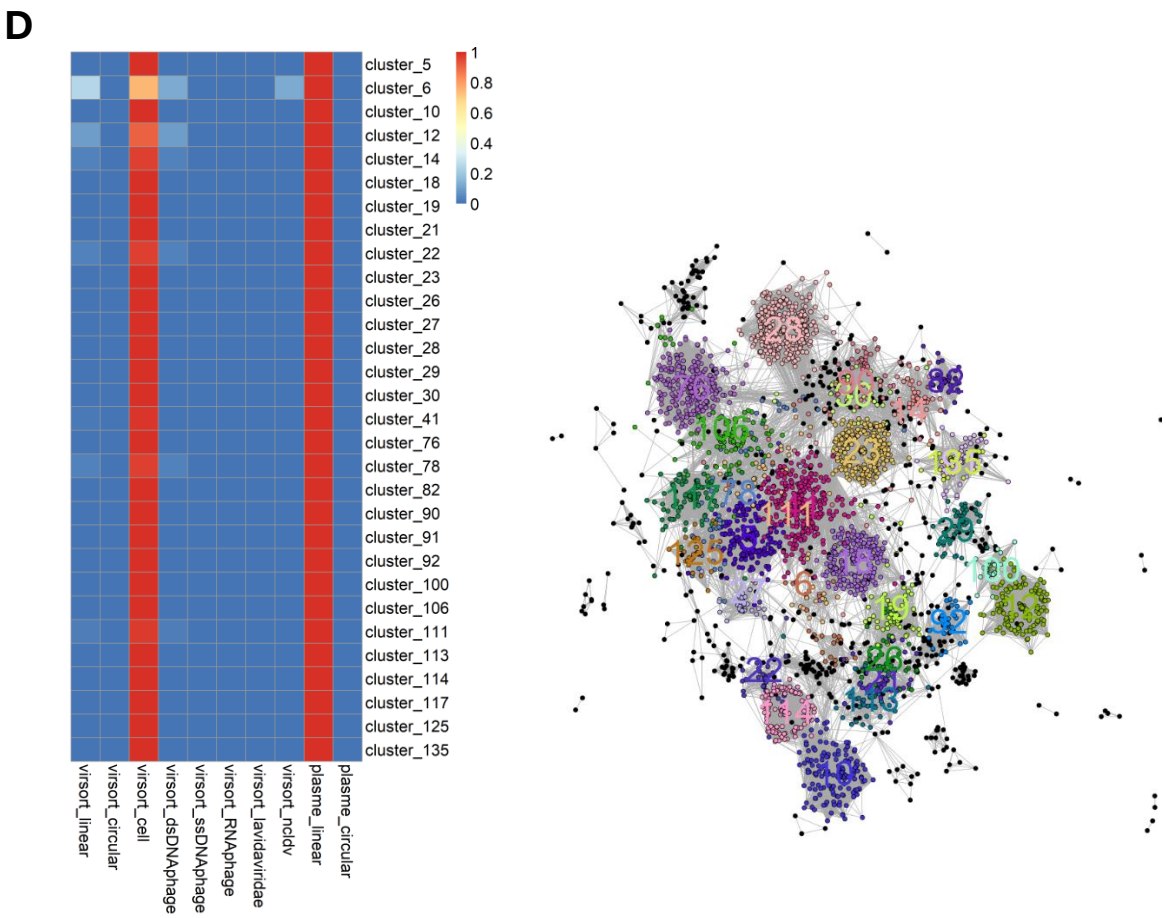

J
K
L
cluster_3
cluster_8

<u>Figure S4.2:</u> *Shannon equitability index of NCBI and JGI (GOLD) metadata derived sequences for each host-encoded CRISPR-Cas subtype investigated.* *Lower (non-zero values mean higher species diversity). A value of 1 means only one species/sample type was present. A value of NA (grey) meant that metadata was not available for a given cluster. The cluster labelled NA represented the diversity of all sequences not assigned to one of the top 30 clusters. In most cases the metadata describing JGI and NCBI sequences was mutually exclusive (i.e. Having JGI metadata implied the absence of JGI data and vice versa). Annotation was not performed for cluster sizes < 6 as these were assumed to be too small to estimate the diversity accurately. (A) Type VI-A, (B) Type VI-B, (C) Type VI-D, (D) Type V-A, (E) Type V-B, (F) Type V-F1, (G) Type I-A, (H) Type I-B, (I) Type I-D, (J) Type I-F, (K) Type III-A, (L) Type III-B.*

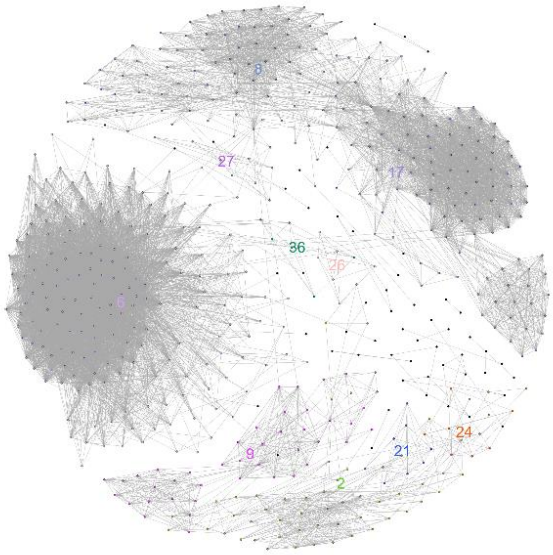


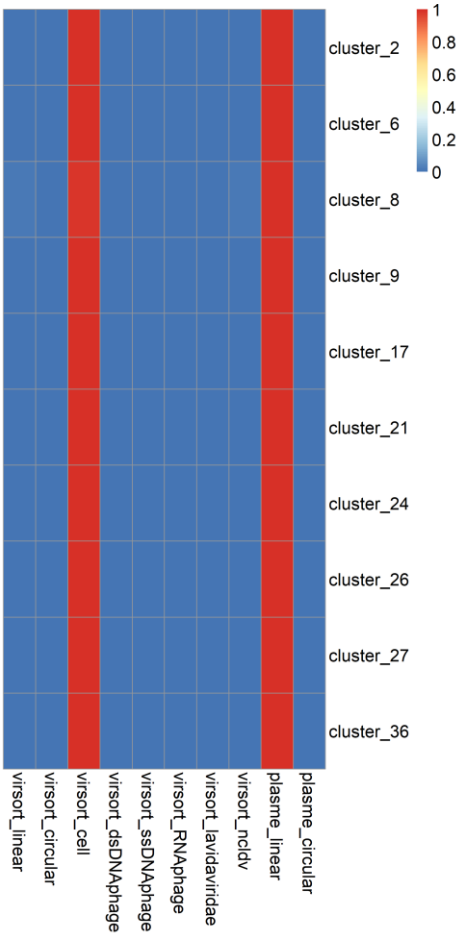

B
7
20
27
18
42
41
4
16
55
14
28
39
62
60
5
29
1
43
60
cluster_1
cluster_4
cluster_5
cluster_7
cluster_14
cluster_16
cluster_17
cluster_18
cluster_20
cluster_23
cluster_27
cluster_28
cluster_29
cluster_39
cluster_41
cluster_42
cluster_43
cluster_50
cluster_51
cluster_55
cluster_56
cluster_60
cluster_62
cluster_63
1
0.8
0.6
0.4
0.2
0
virsort_linear
virsort_circular
virsort_cell
virsort_dsDNAphage
virsort_ssDNAphage
virsort_RNAphage
virsort_lavidaviridae
virsort_ncldv
plasme_linear
plasme_circular


C
19
16
8
7
21
5
22
11
cluster_5
cluster_7
cluster_8
cluster_11
cluster_16
cluster_19
cluster_21
cluster_22
1
0.8
0.6
0.4
0.2
0
virsort_linear
virsort_circular
virsort_cell
virsort_dsDNAphage
virsort_ssDNAphage
virsort_RNAphage
virsort_lavidaviridae
virsort_ncldv
plasme_linear
plasme_circular

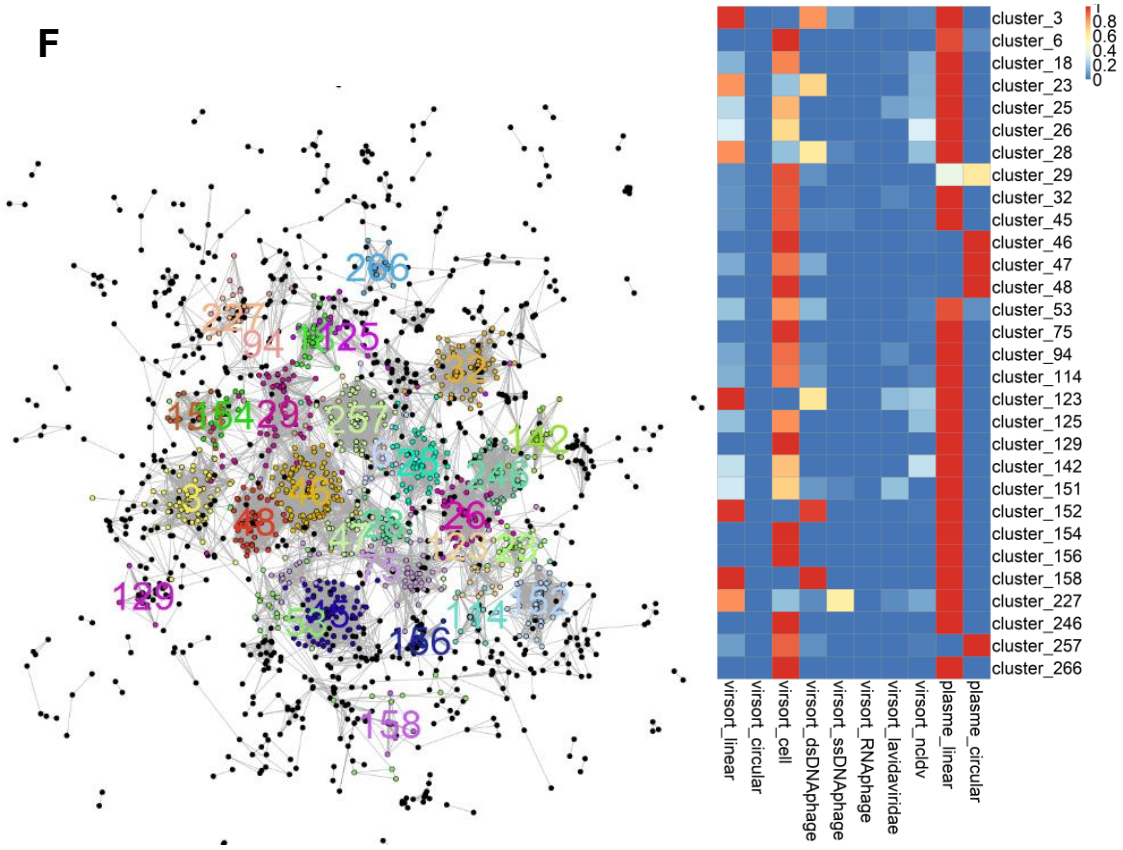

D
cluster_5
cluster_6
cluster_10
cluster_12
cluster_14
cluster_18
cluster_19
cluster_21
cluster_22
cluster_23
cluster_26
cluster_27
cluster_28
cluster_29
cluster_30
cluster_41
cluster_76
cluster_78
cluster_82
cluster_90
cluster_91
cluster_92
cluster_100
cluster_106
cluster_111
cluster_113
cluster_114
cluster_117
cluster_125
cluster_135
1
0.8
0.6
0.4
0.2
0
virsort_linear
virsort_circular
virsort_cell
virsort_dsDNAphage
virsort_ssDNAphage
virsort_RNAphage
virsort_lavidaviridae
virsort_ncldv
plasme_linear
plasme_circular

E

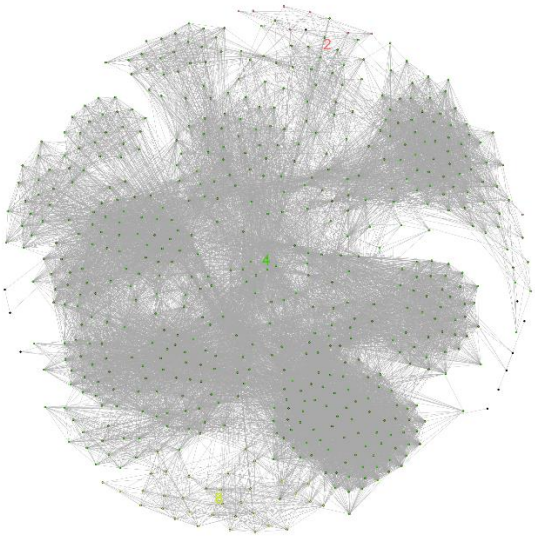

8
27
17
36
26
24
9
21
2

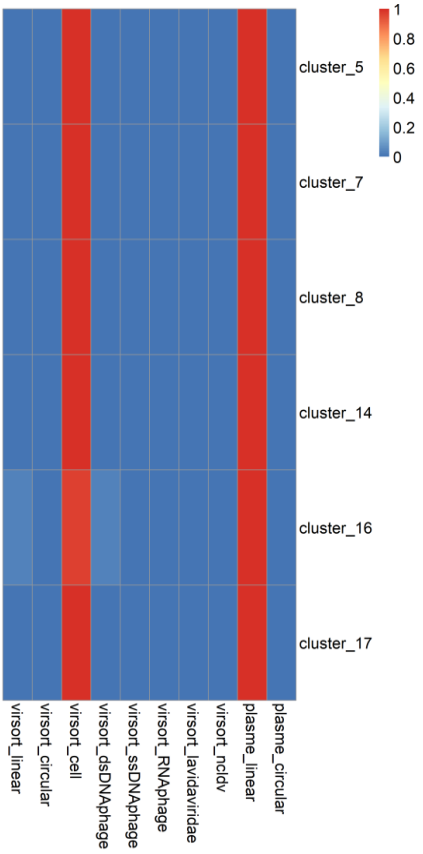

1
0.8
0.6
0.4
0.2
0
cluster_2
cluster_6
cluster_8
cluster_9
cluster_17
cluster_21
cluster_24
cluster_26
cluster_27
cluster_36
virsort_linear
virsort_circular
virsort_cell
virsort_dsDNAphage
virsort_ssDNAphage
virsort_RNAphage
virsort_lavidaviridae
virsort_ncldv
plasme_linear
plasme_circular

F

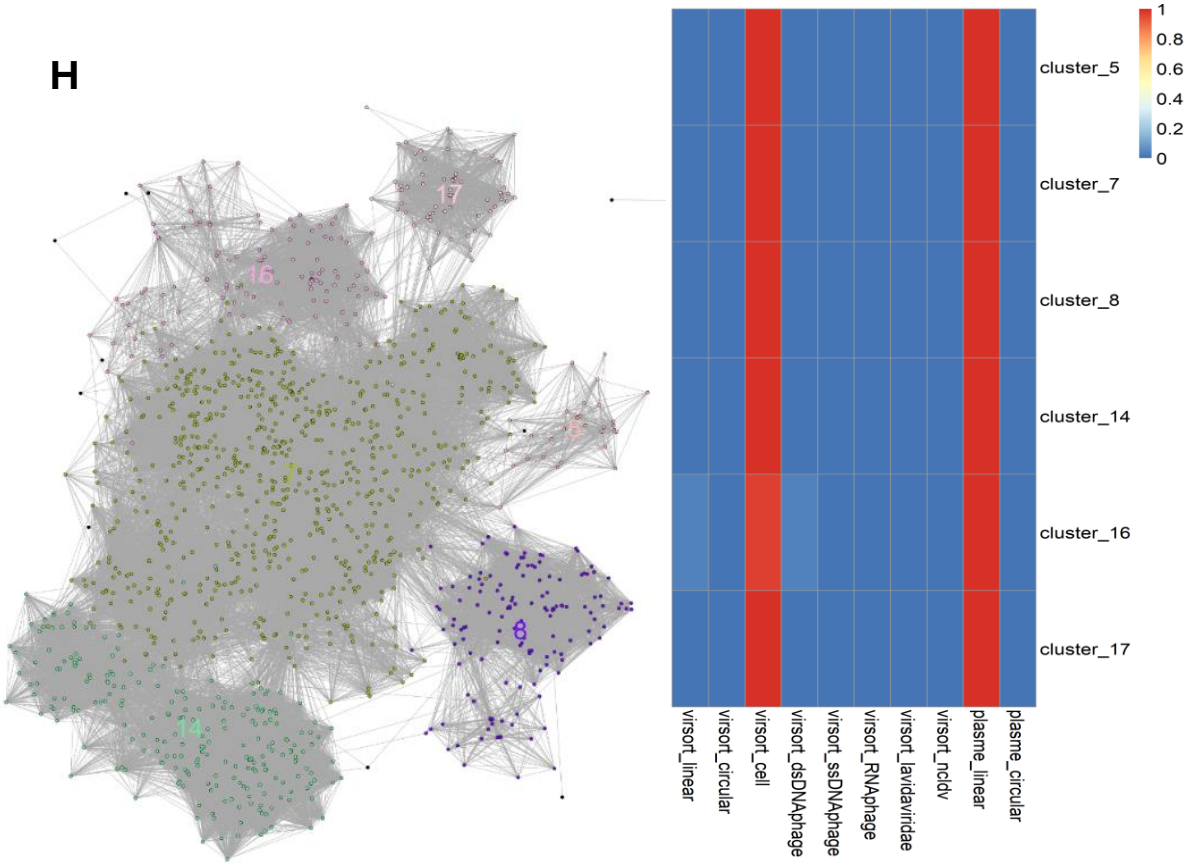

266
94
125
142
26
129
114
158
cluster_3
cluster_6
cluster_18
cluster_23
cluster_25
cluster_26
cluster_28
cluster_29
cluster_32
cluster_45
cluster_46
cluster_47
cluster_48
cluster_53
cluster_75
cluster_94
cluster_114
cluster_123
cluster_125
cluster_129
cluster_142
cluster_151
cluster_152
cluster_154
cluster_156
cluster_158
cluster_227
cluster_246
cluster_257
cluster_266
1
0.8
0.6
0.4
0.2
0
virsort_linear
virsort_circular
virsort_cell
virsort_dsDNAphage
virsort_ssDNAphage
virsort_RNAphage
virsort_lavidaviridae
virsort_ncldv
plasme_linear
plasme_circular

G

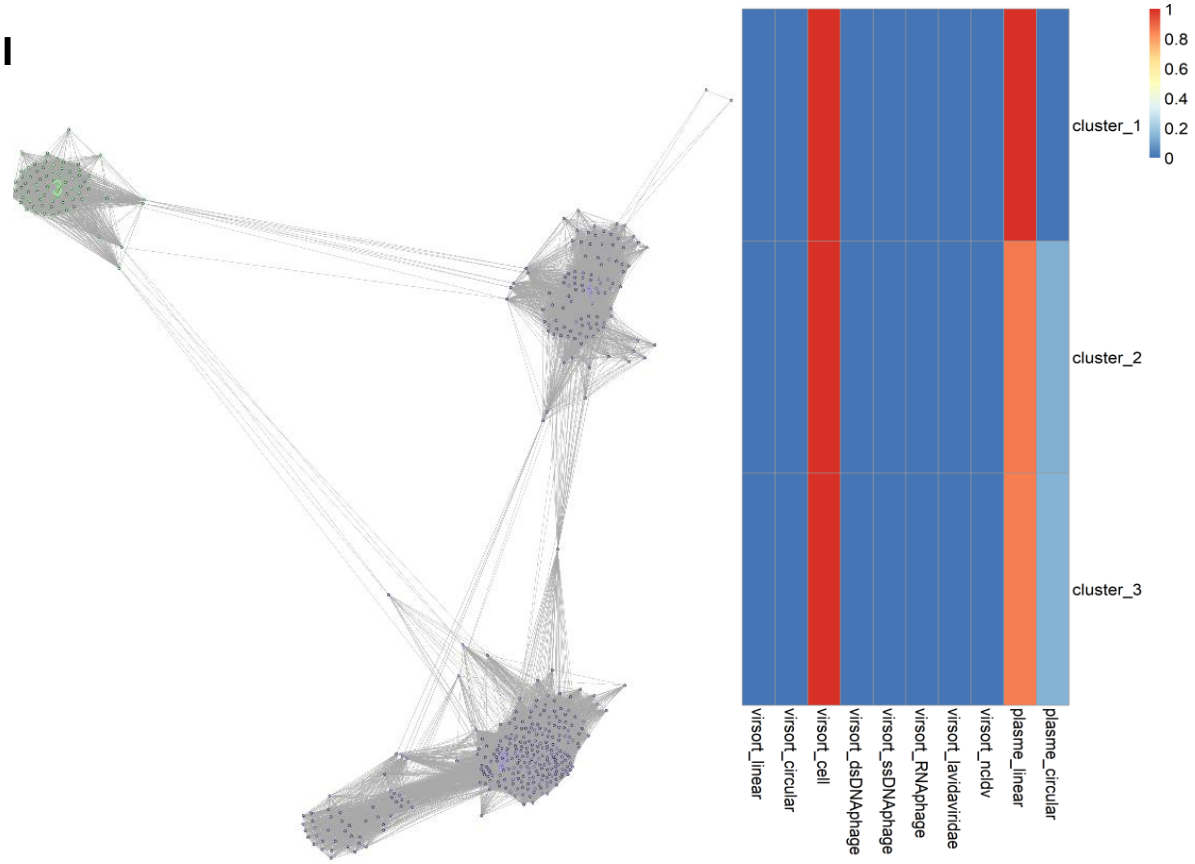



1
0.8
0.6
0.4
0.2
0
cluster_5
cluster_7
cluster_8
cluster_14
cluster_16
cluster_17
virsort_linear
virsort_circular
virsort_cell
virsort_dsDNAphage
virsort_ssDNAphage
virsort_RNAphage
virsort_lavidaviridae
virsort_ncldv
plasme_linear
plasme_circular

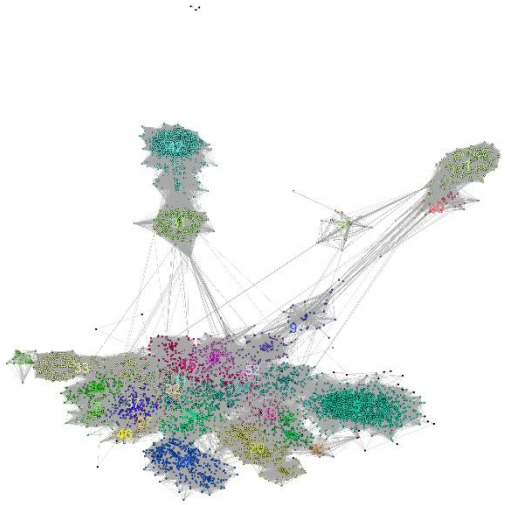
H

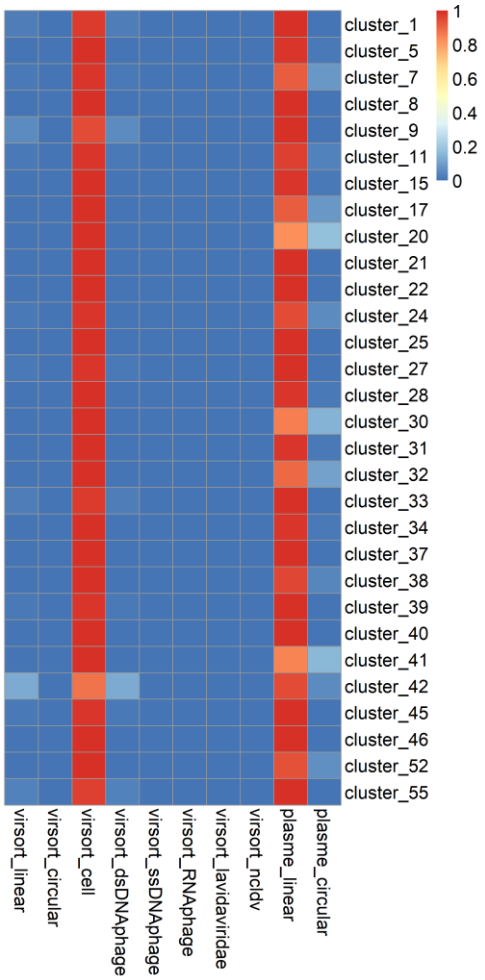

17
16
5
3
8
14
cluster_5
cluster_7
cluster_8
cluster_14
cluster_16
cluster_17
virsort_linear
virsort_circular
virsort_cell
virsort_dsDNAphage
virsort_ssDNAphage
virsort_RNAphage
virsort_lavidaviridae
virsort_ncldv
plasme_linear
plasme_circular
1
0.8
0.6
0.4
0.2
0

I

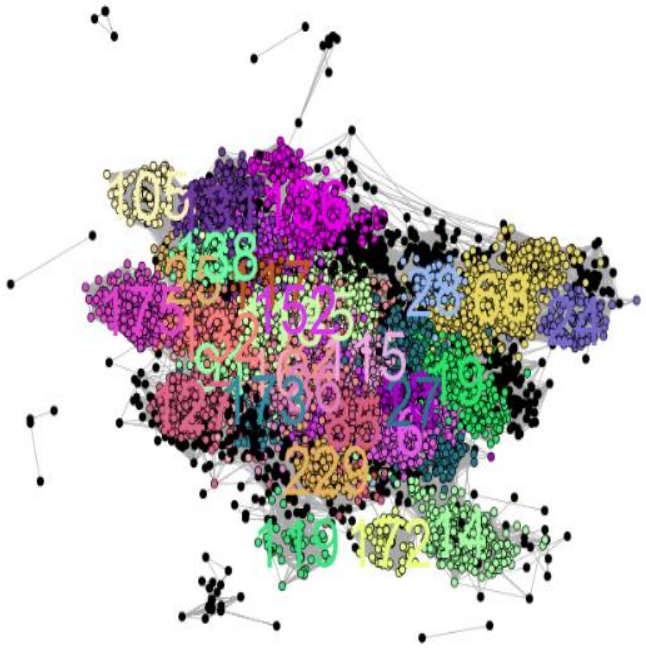

3
cluster_1
cluster_2
cluster_3
virsort_linear
virsort_circular
virsort_cell
virsort_dsDNAphage
virsort_ssDNAphage
virsort_RNAphage
virsort_lavidaviridae
virsort_ncldv
plasme_linear
plasme_circular
1
0.8
0.6
0.4
0.2
0

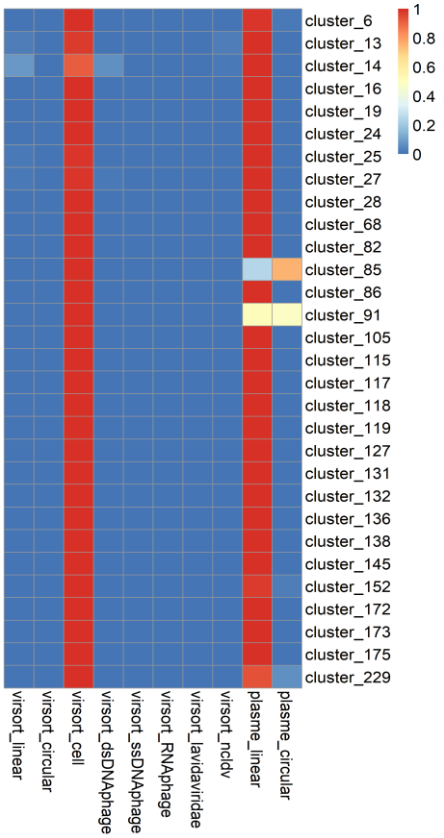
J



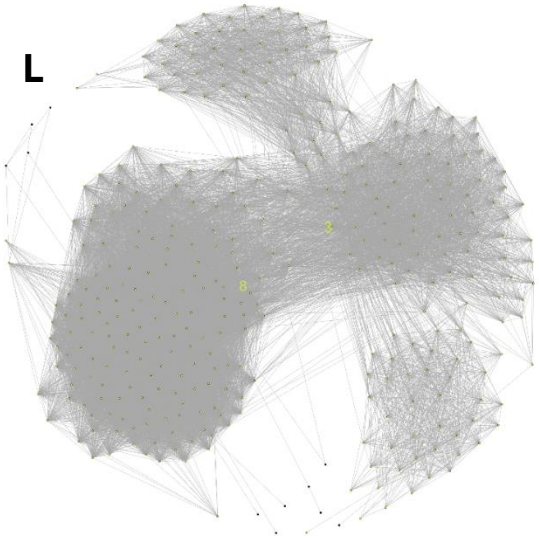

cluster_1
cluster_5
cluster_7
cluster_8
cluster_9
cluster_11
cluster_15
cluster_17
cluster_20
cluster_21
cluster_22
cluster_24
cluster_25
cluster_27
cluster_28
cluster_30
cluster_31
cluster_32
cluster_33
cluster_34
cluster_37
cluster_38
cluster_39
cluster_40
cluster_41
cluster_42
cluster_45
cluster_46
cluster_52
cluster_55
1
0.8
0.6
0.4
0.2
0
virsort_linear
virsort_circular
virsort_cell
virsort_dsDNAphage
virsort_ssDNAphage
virsort_RNAphage
virsort_lavidaviridae
virsort_ncldv
plasme_linear
plasme_circular

K




cluster_6
cluster_13
cluster_14
cluster_16
cluster_19
cluster_24
cluster_25
cluster_27
cluster_28
cluster_68
cluster_82
cluster_85
cluster_86
cluster_91
cluster_105
cluster_115
cluster_117
cluster_118
cluster_119
cluster_127
cluster_131
cluster_132
cluster_136
cluster_138
cluster_145
cluster_152
cluster_172
cluster_173
cluster_175
cluster_229
virsort_linear
virsort_circular
virsort_cell
virsort_dsDNAphage
virsort_ssDNAphage
virsort_RNAphage
virsort_lavidaviridae
virsort_ncldv
plasme_linear
plasme_circular
1
0.8
0.6
0.4
0.2
0

L



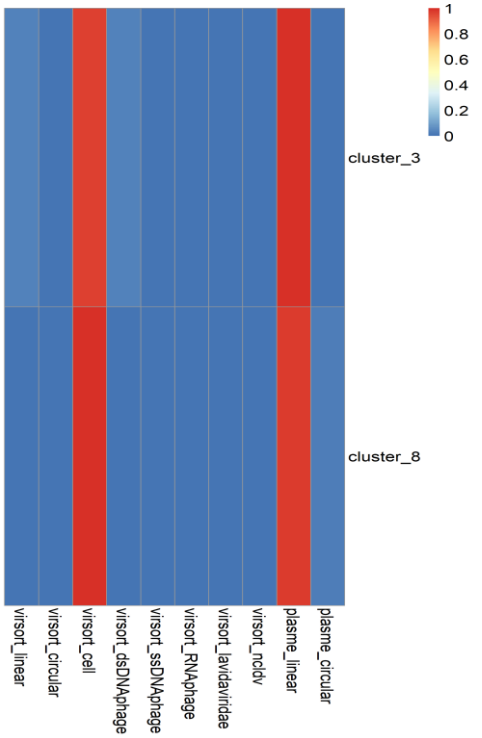
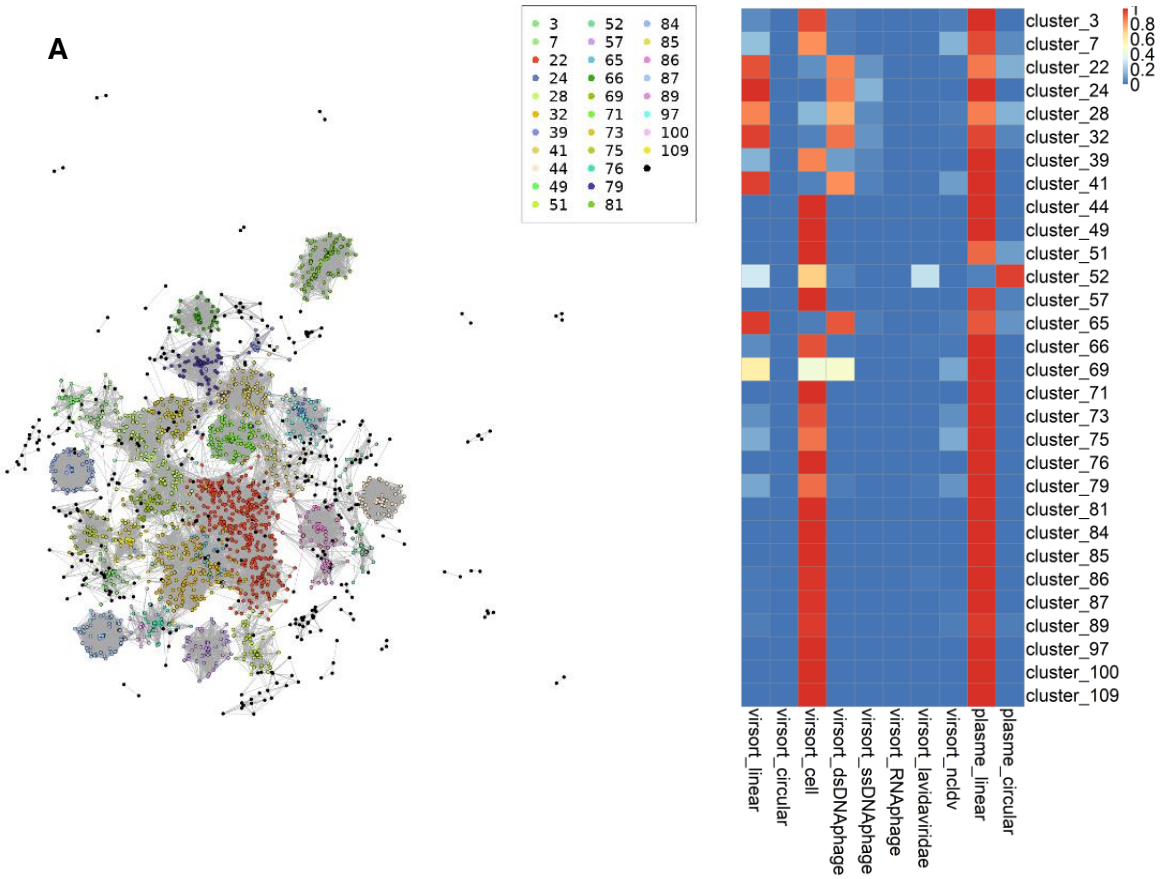
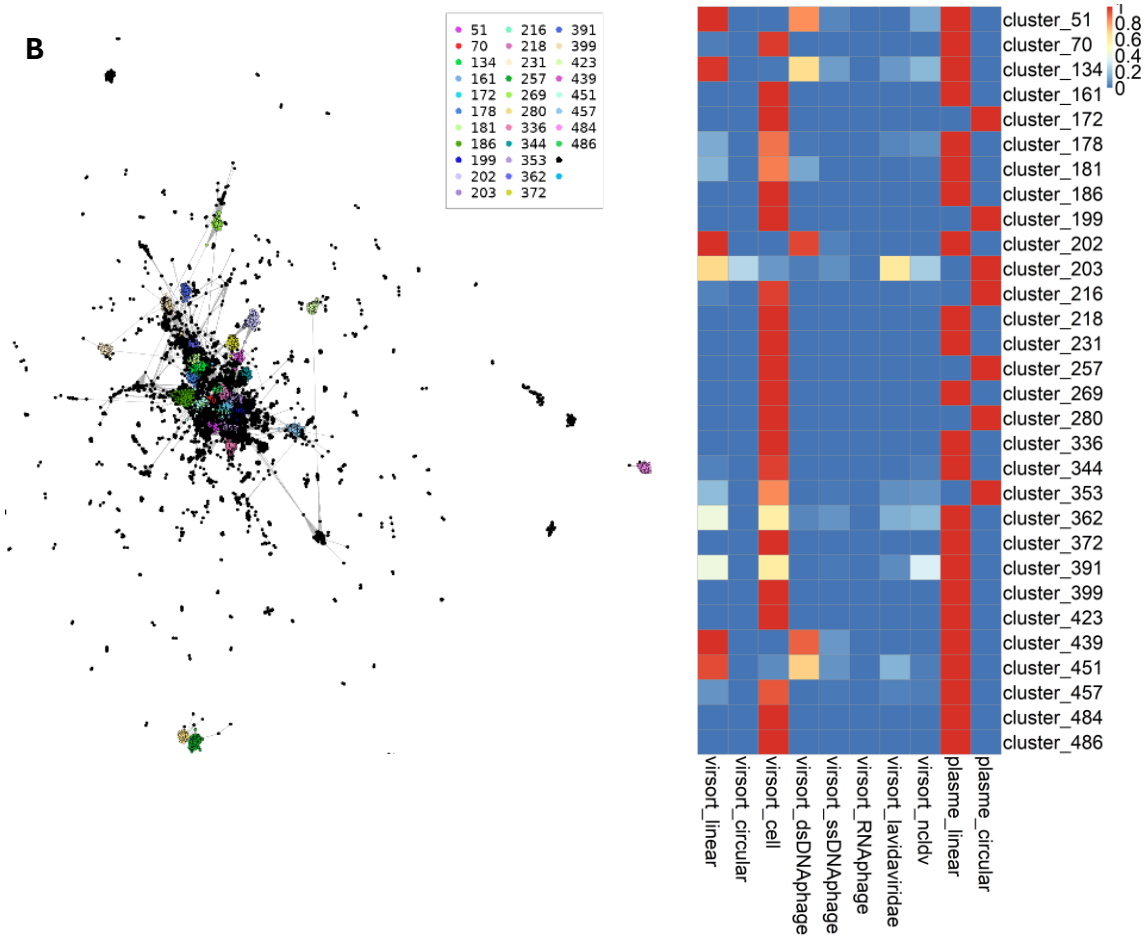
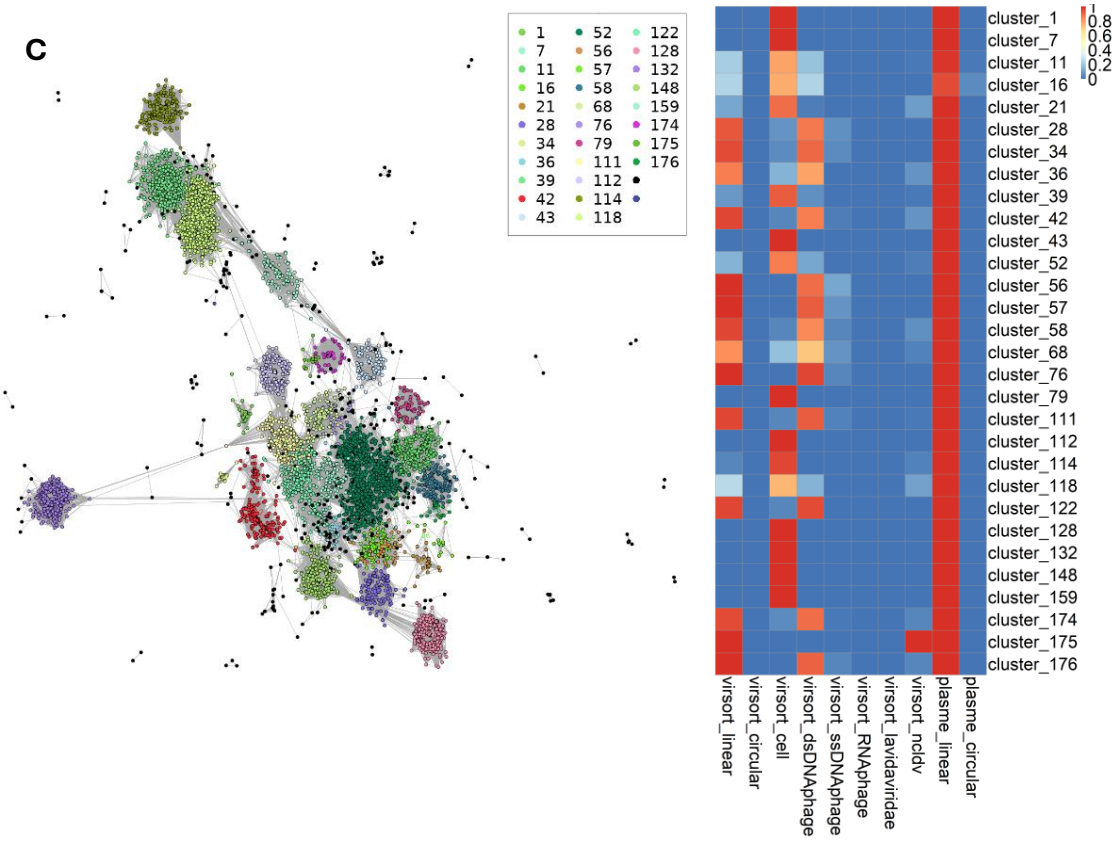
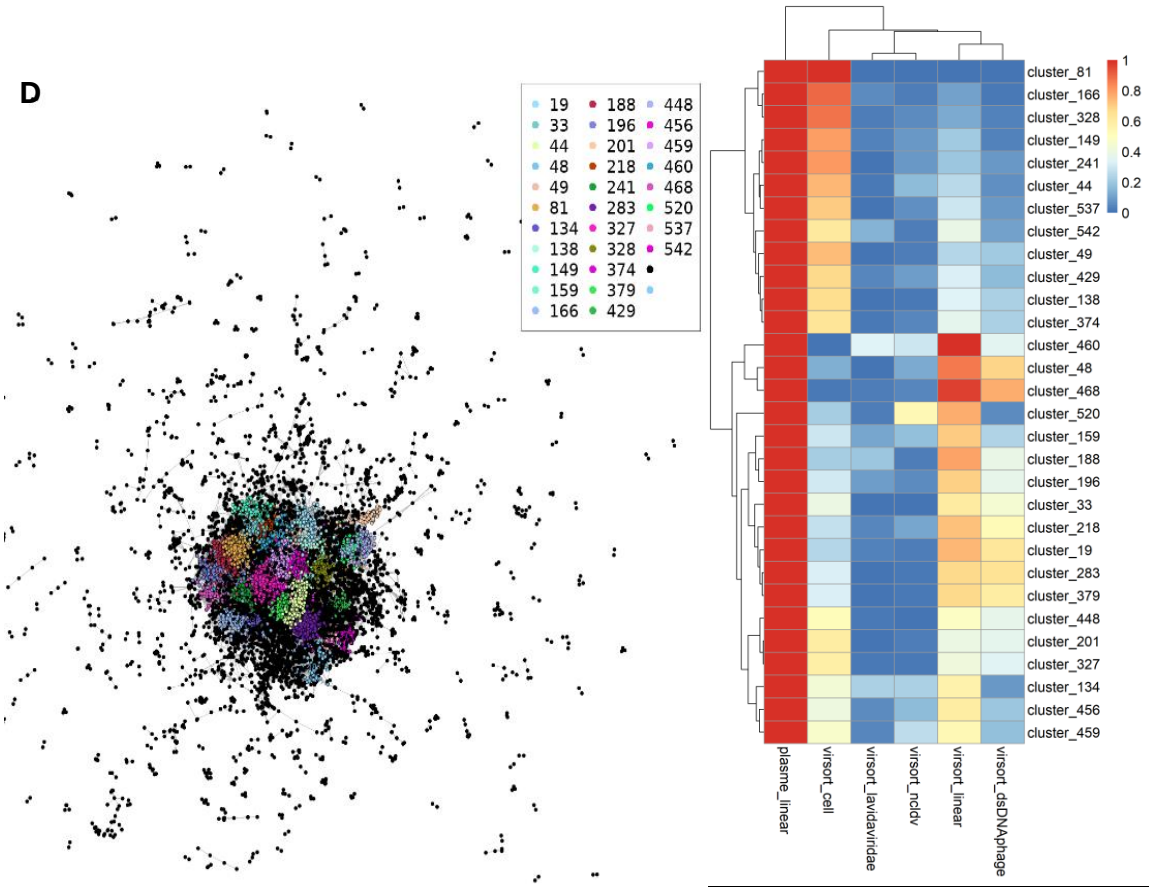

cluster_3
cluster_8
virsort_linear
virsort_circular
virsort_cell
virsort_dsDNAphage
virsort_ssDNAphage
virsort_RNAphage
virsort_lavidaviridae
virsort_ncldv
plasme_linear
plasme_circular
1
0.8
0.6
0.4
0.2
0

Figure S4.3: ***Host gene cluster network analysis of selected host-encoded CRISPR-Cas subtypes.*** *For each cluster in the network, the proportion of sequences classified as Viral by Virsorter2 or linear/circular by either Virsorter2 or PLASME, prediction programs, was computed. Sequences not classified as viral by virsorter were assigned as "virsort_cell", while those not classified as circular were classed by default as linear. Viral/plasmid prediction as well as gene composition was calculated for the top 30 largest clusters within each subtype network. Annotations of ORFs were derived from an amalgamation of the Pfam and DEFLOC databases. Where an annotation for a given gene was found from both set of Pfam and DEFLOC profiles, DEFLOC was chosen preferentially, as these profiles possessed more accurate descriptions of the translated gene products than pfam. Singletons or genes too divergent to establish any relations by sequence similarity were excluded by vCONTACT2 during network generation. , (A) Type VI-A, (B) Type VI-B, (C) Type VI-D, (D) Type V-A, (E) Type V-B, (F) Type V-F1, (G) Type I-A, (H) Type I-B, (I) Type I-D, (J) Type I-F, (K) Type III-A, (L) Type III-B.*




B
51 70 134 161 172 178 181 186 199 202 203
216 218 231 257 269 280 336 344 353 362 372
391 399 423 439 451 457 484 486
cluster_51 cluster_70 cluster_134 cluster_161 cluster_172 cluster_178 cluster_181 cluster_186 cluster_199 cluster_202 cluster_203 cluster_216 cluster_218 cluster_231 cluster_257 cluster_269 cluster_280 cluster_336 cluster_344 cluster_353 cluster_362 cluster_372 cluster_391 cluster_399 cluster_423 cluster_439 cluster_451 cluster_457 cluster_484 cluster_486
1 0.8 0.6 0.4 0.2 0
virsort_linear virsort_circular virsort_cell virsort_dsDNAphage virsort_ssDNAphage virsort_RNAphage virsort_lavidaviridae virsort_ncldv plasme_linear plasme_circular


C
1 7 11 16 21 28 34 36 39 42 43
52 56 57 58 68 76 79 111 112 114 118
122 128 132 148 159 174 175 176
cluster_1 cluster_7 cluster_11 cluster_16 cluster_21 cluster_28 cluster_34 cluster_36 cluster_39 cluster_42 cluster_43 cluster_52 cluster_56 cluster_57 cluster_58 cluster_68 cluster_76 cluster_79 cluster_111 cluster_112 cluster_114 cluster_118 cluster_122 cluster_128 cluster_132 cluster_148 cluster_159 cluster_174 cluster_175 cluster_176
1 0.8 0.6 0.4 0.2 0
virsort_linear virsort_circular virsort_cell virsort_dsDNAphage virsort_ssDNAphage virsort_RNAphage virsort_lavidaviridae virsort_ncldv plasme_linear plasme_circular


D
19 188 448
33 196 456
44 201 459
48 218 460
49 241 468
81 283 520
134 327 537
138 328 542
149 374
159 379
166 429
cluster_81
cluster_166
cluster_328
cluster_149
cluster_241
cluster_44
cluster_537
cluster_542
cluster_49
cluster_429
cluster_138
cluster_374
cluster_460
cluster_48
cluster_468
cluster_520
cluster_159
cluster_188
cluster_196
cluster_33
cluster_218
cluster_19
cluster_283
cluster_379
cluster_448
cluster_201
cluster_327
cluster_134
cluster_456
cluster_459
1
0.8
0.6
0.4
0.2
0
plasme_linear
virsort_cell
virsort_lavidaviridae
virsort_ncldv
virsort_linear
virsort_dsDNAphage

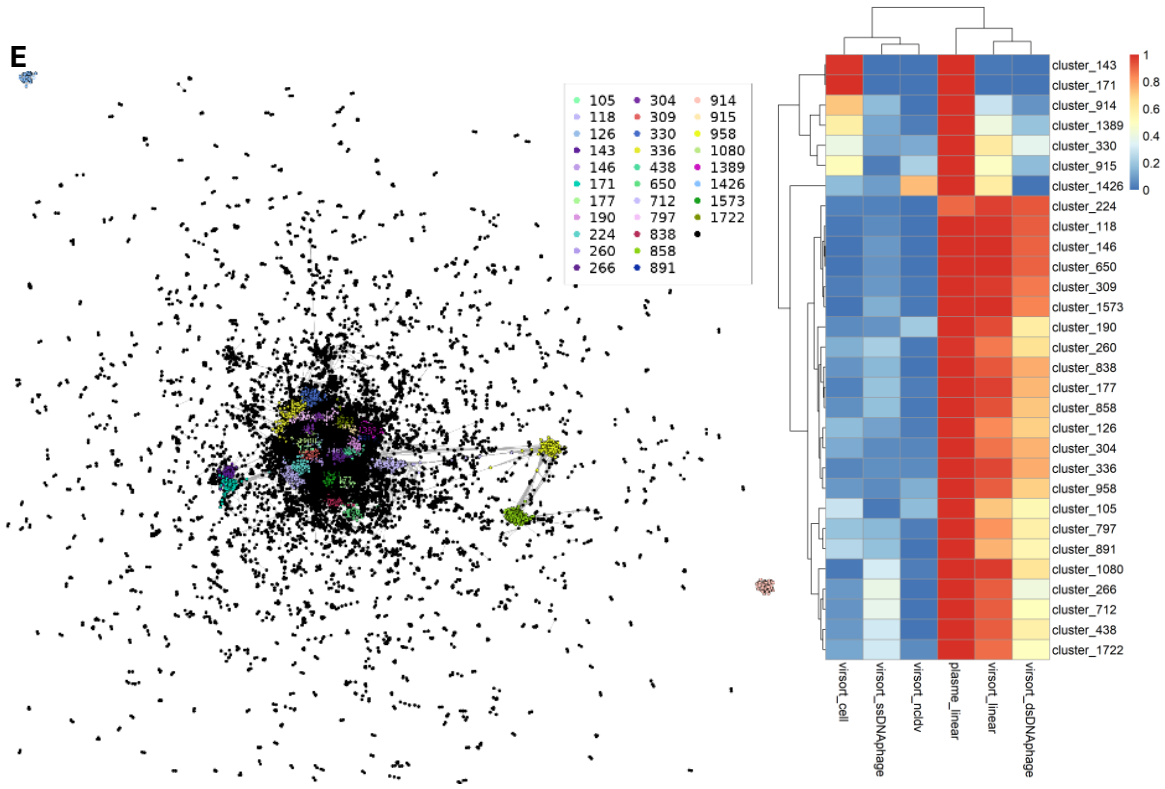
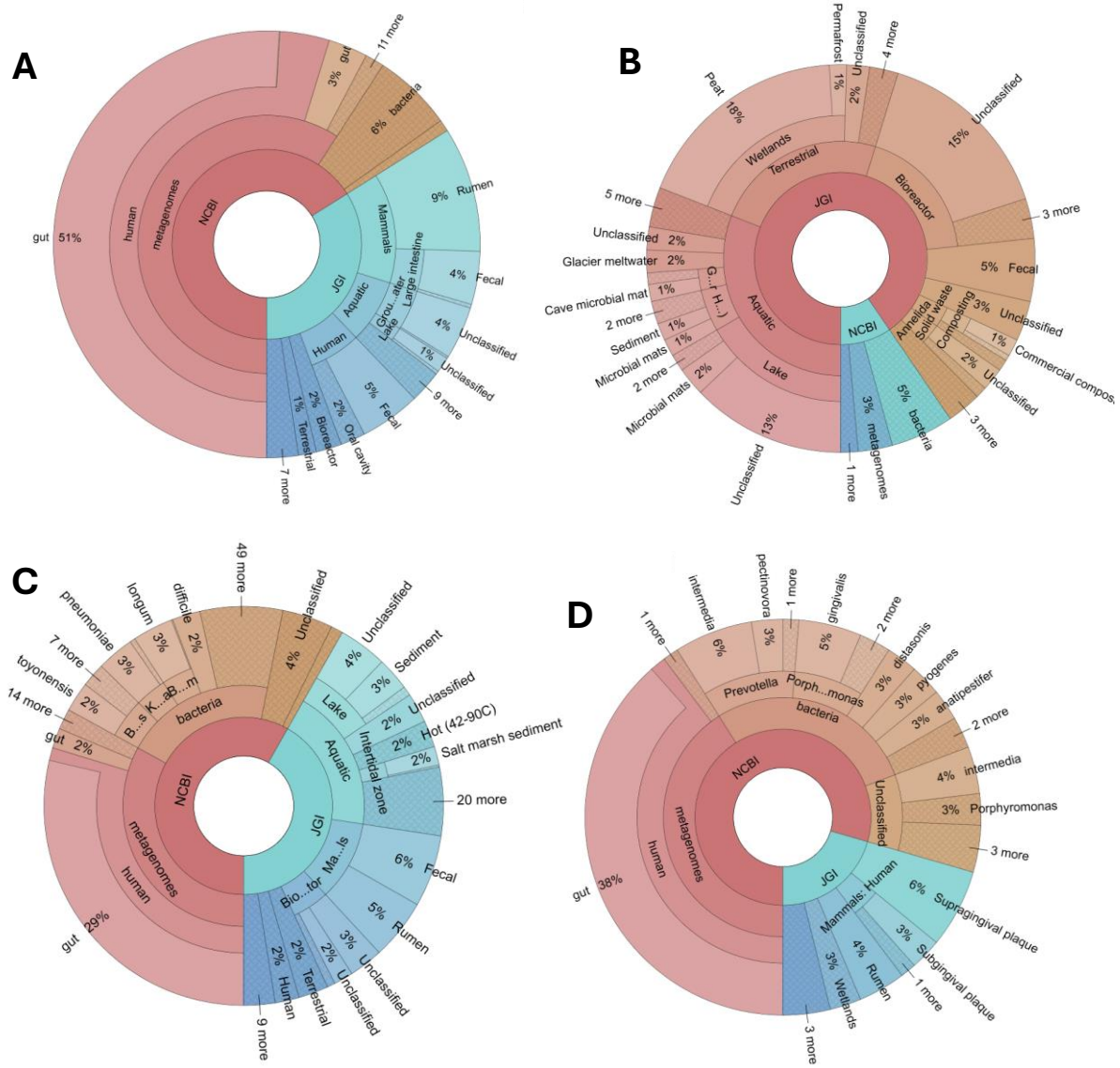

E
105 304 914
118 309 915
126 330 958
143 336 1080
146 438 1389
171 650 1426
177 712 1573
190 797 1722
224 838
260 858
266 891
cluster_143
cluster_171
cluster_914
cluster_1389
cluster_330
cluster_915
cluster_1426
cluster_224
cluster_118
cluster_146
cluster_650
cluster_309
cluster_1573
cluster_190
cluster_260
cluster_838
cluster_177
cluster_858
cluster_126
cluster_304
cluster_336
cluster_958
cluster_105
cluster_797
cluster_891
cluster_1080
cluster_266
cluster_712
cluster_438
cluster_1722
1
0.8
0.6
0.4
0.2
0
virsort_cell
virsort_ssDNAphage
virsort_ncldv
plasme_linear
virsort_linear
virsort_dsDNAphage

Figure S4.4: ***Mapped spacer target gene cluster network analysis from selected CRISPR-Cas subtypes.*** *For each cluster in the network, the proportion of sequences classified as Viral by Virsorter2 or linear/circular by either Virsorter2 or PLASME, prediction programs, was computed. Sequences not classified as viral by virsorter were assigned as "virsort_cell", while those not classified as circular were classed by default as linear. Conservation was defined as the proportion a given gene was encoded compared to the total number of contigs in each subtype as applied in (chapter 1.4). Viral/plasmid prediction as well as gene composition was only calculated for the top 30 largest clusters within each subtype network. Singletons or genes too divergent to establish any relations by sequence similarity were excluded by vCONTACT2 during network generation. (A) Type VI-A,(B) Type VI-B, (C) Type VI-D , (C) Type I-A, (E) Type I-B.*

Supplementary data for Chapter 5

Characterisation of spacer mapping patterns across CRISPR-Cas systems



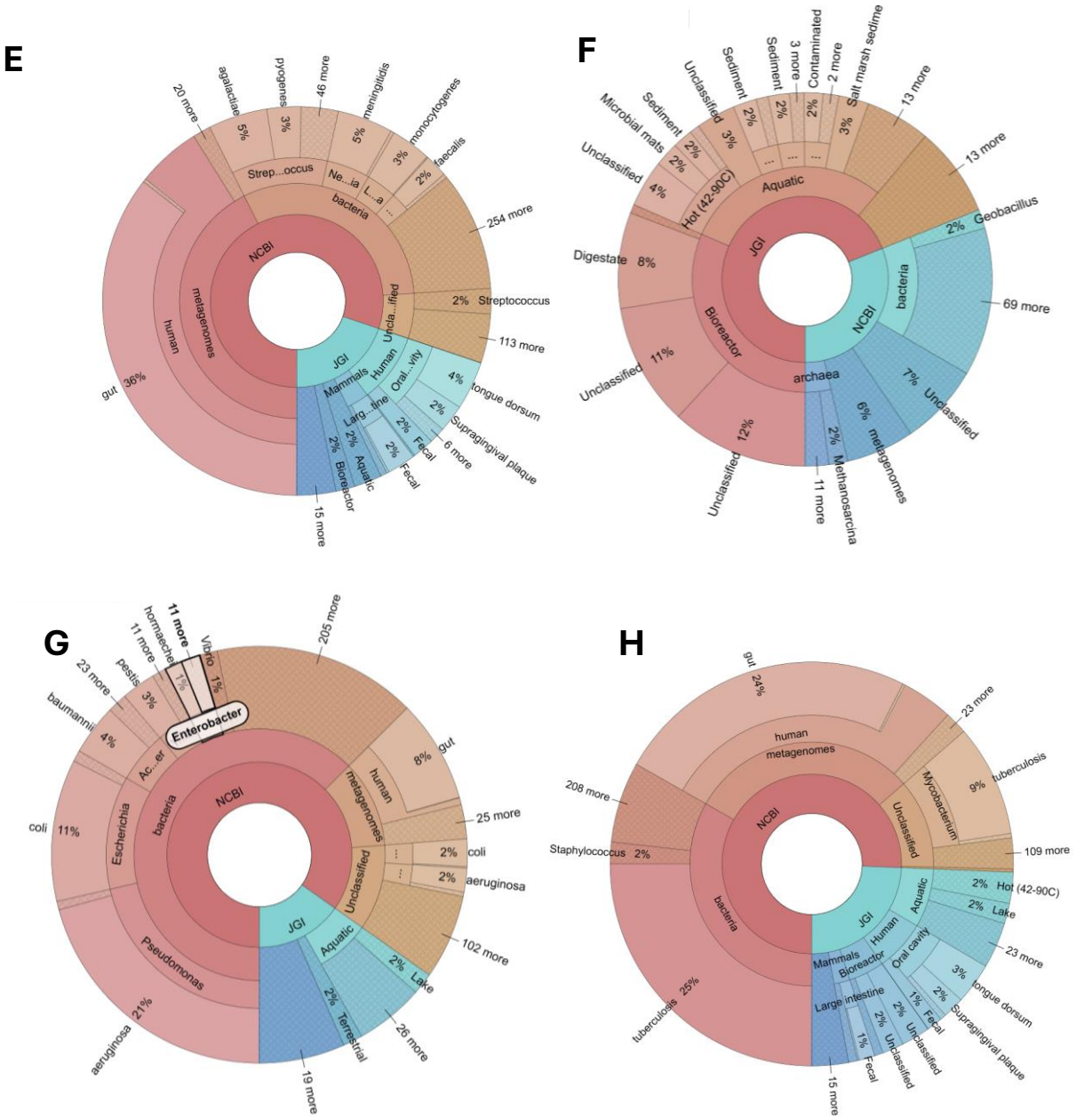
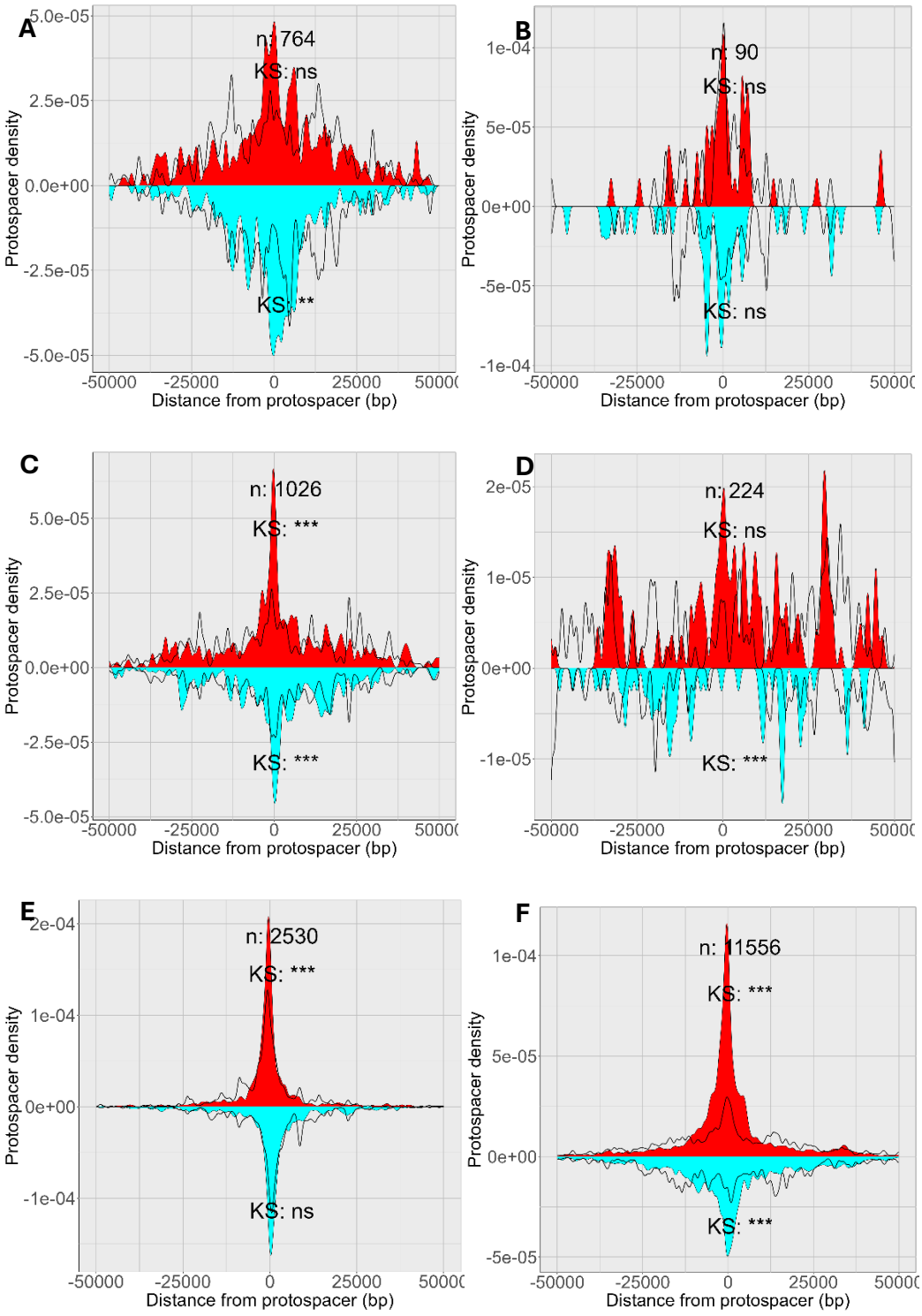
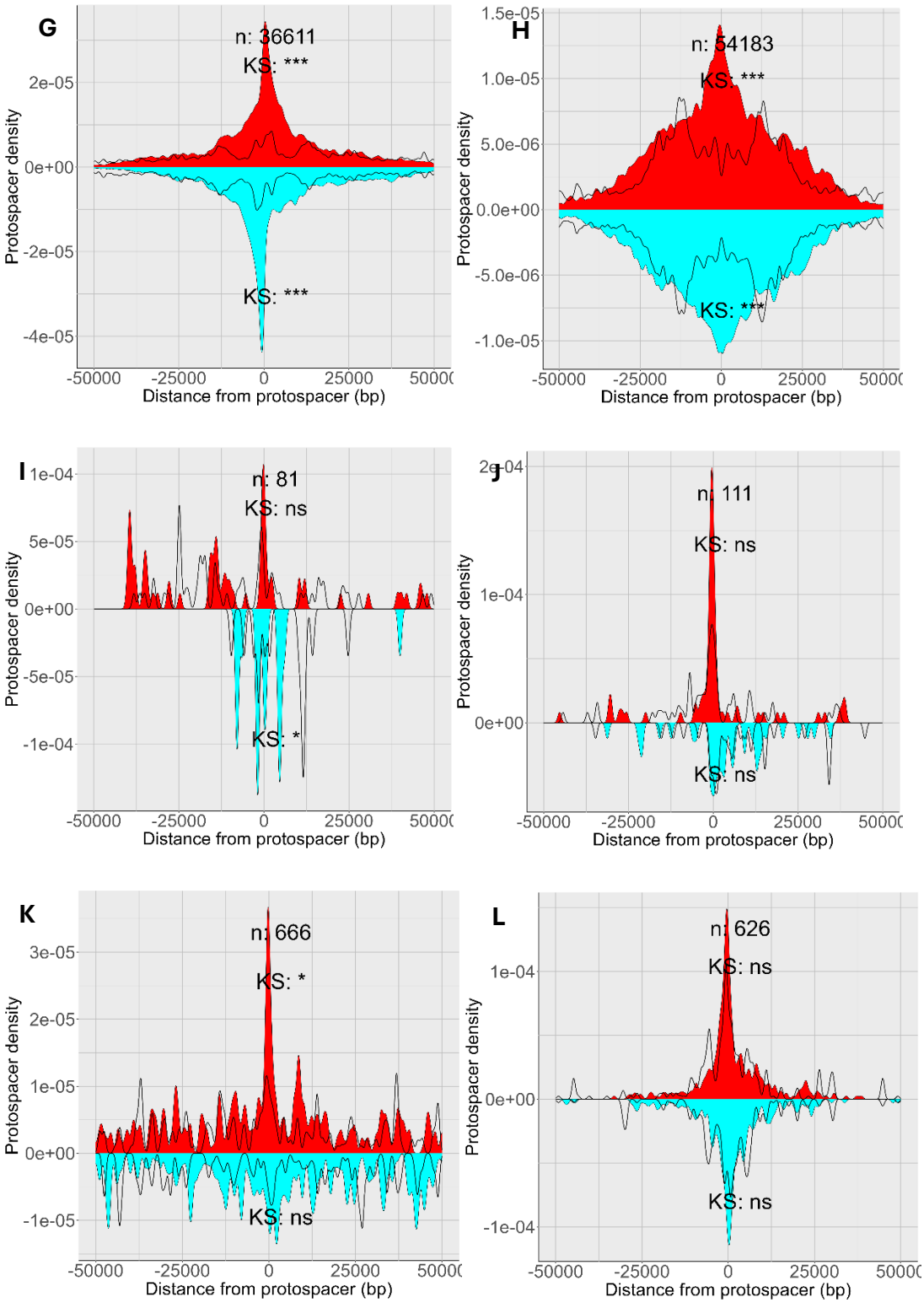


Figure S5.1: Census of CRISPR-Cas subtype host genomes. *(A) Type V-A, (B) Type V-B, (C) Type V-F1,(D) Type VI-B, (E) Type-II, (F) Type I-B, (G) Type I-F, (H) Type III-A*


A
n: 764
KS: ns
KS: **
B
n: 90
KS: ns
KS: ns
C
n: 1026
KS: ***
KS: ***
D
n: 224
KS: ns
KS: ***
E
n: 2530
KS: ***
KS: ns
F
n: 11556
KS: ***
KS: ***
Protospacer density
Distance from protospacer (bp)


G
n: 36611
KS: ***
KS: ***
Protospacer density
2e-05
0e+00
-2e-05
-4e-05
-50000
-25000
0
25000
Distance from protospacer (bp)
H
n: 54183
KS: ***
KS: ***
Protospacer density
1.5e-05
1.0e-05
5.0e-06
0.0e+00
-5.0e-06
-1.0e-05
Distance from protospacer (bp)
I
n: 81
KS: ns
KS: *
Protospacer density
1e-04
5e-05
0e+00
-5e-05
-1e-04
Distance from protospacer (bp)
J
n: 111
KS: ns
KS: ns
Protospacer density
1e-04
0e+00
Distance from protospacer (bp)
K
n: 666
KS: *
KS: ns
Protospacer density
3e-05
2e-05
1e-05
0e+00
-1e-05
Distance from protospacer (bp)
L
n: 626
KS: ns
KS: ns
Protospacer density
1e-04
0e+00
-1e-04
Distance from protospacer (bp)

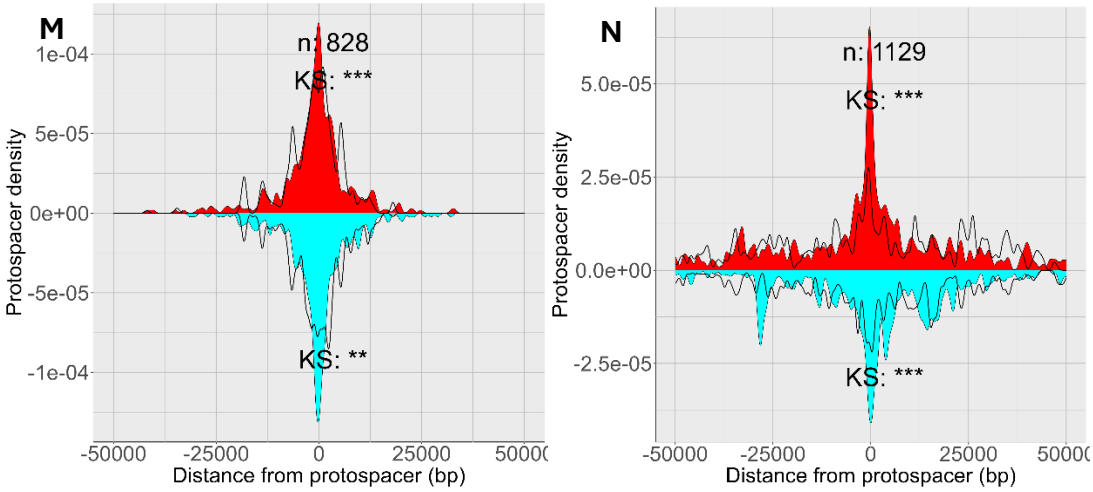


Figure S5.2: Expected vs. mapped spacer distances in representative CRISPR-Cas subtypes (within 50kb). *(A) Type V-A, (B) Type V-B, (C) Type V-F1, (D) Type VI-B, (E) Type I-B , (F) Type III-A, (G) Type I-F, (H) Type II, (I) Type VI-D, (J) Type III-B, (K) Type III-E, (L) Type I-D, (M) Type I-A, (N) Type V-J (Type V-U3)*

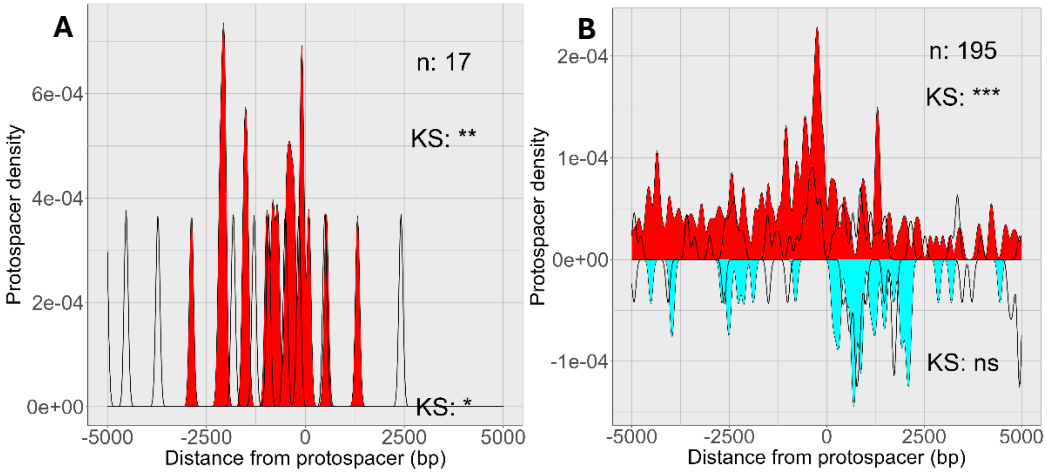


Figure S5.3: Effect of reverse transcriptase (RT) on Type III-A spacer distribution.

Subset of Type III-A spacer distribution based on the presence of and absence of (A) RT-Cas6-Cas1, (B) Cas1 (No RT). 17 PPS-spacer mapping pairs were identified at random from the Type III-A spacer distribution for which homology to previously identified RT-Cas6-Cas1[267] ortholog [PDB: 7KFU_1] could be found fr. 195 PPS-spacer mapping pairs were identified at random to contigs from the Type III-A spacer distribution lacking RT. Sample contigs were screened for the presence or absence of RT using Motif finder. Searches for RT by similarity were conducted using BLAST (e-value 10^{-10}).

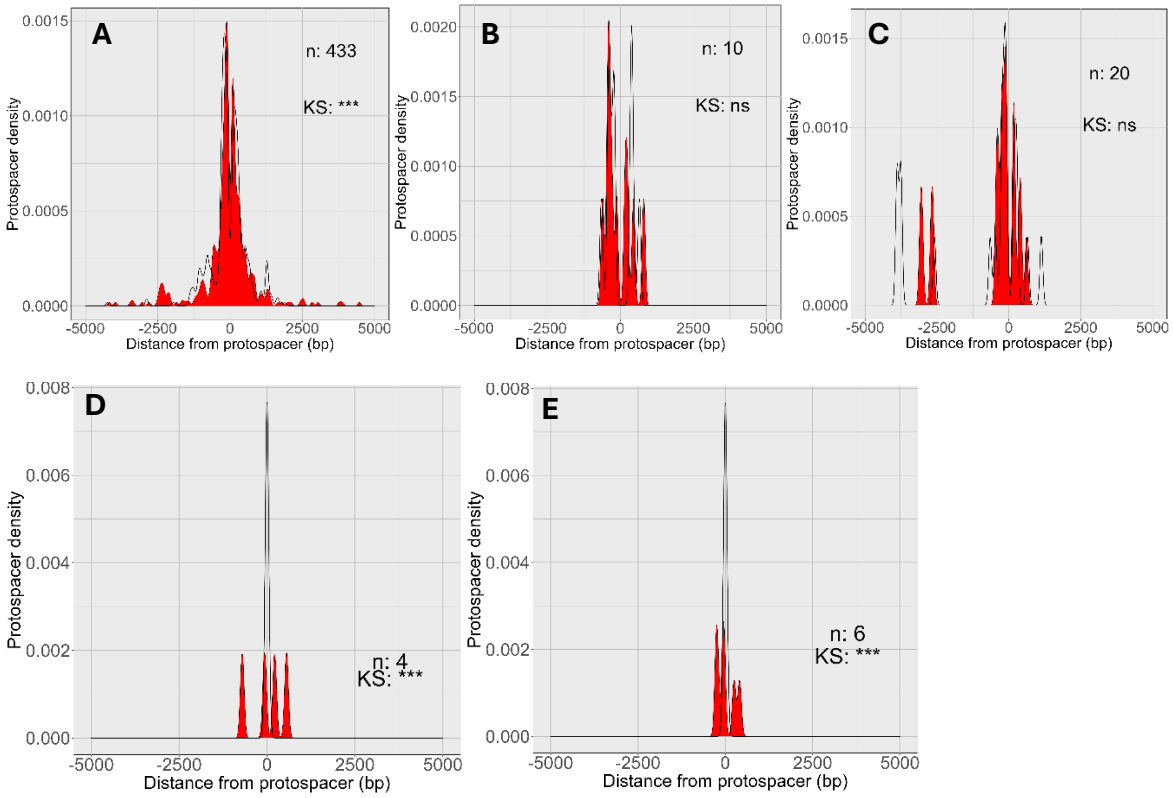


Figure S5.4: ***Intra-ORF mapping distributions in RNA – targeting systems.*** *(A) Type III-A, (B) Type III-B, (C) Type III-E, (D) Type VI-B, (E) Type VI-D*

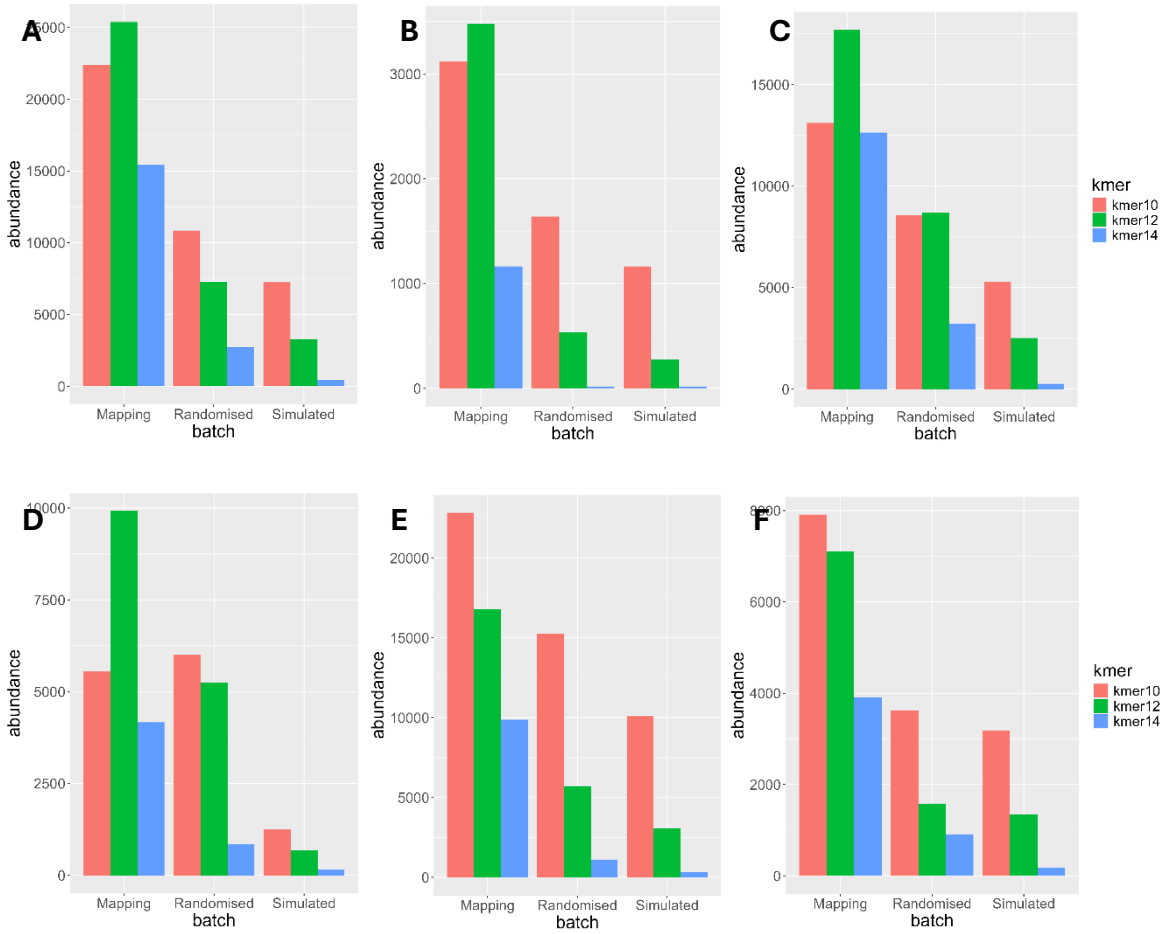


Figure S5.5: Comparison of spacer detection frequencies between test, randomized and synthetic controls in representative CRISPR-Cas subtypes. *A) Type V-A, B) Type V-B, C) Type V-F1, D) Type VI-B, E) type-IB, (F) type I-D*

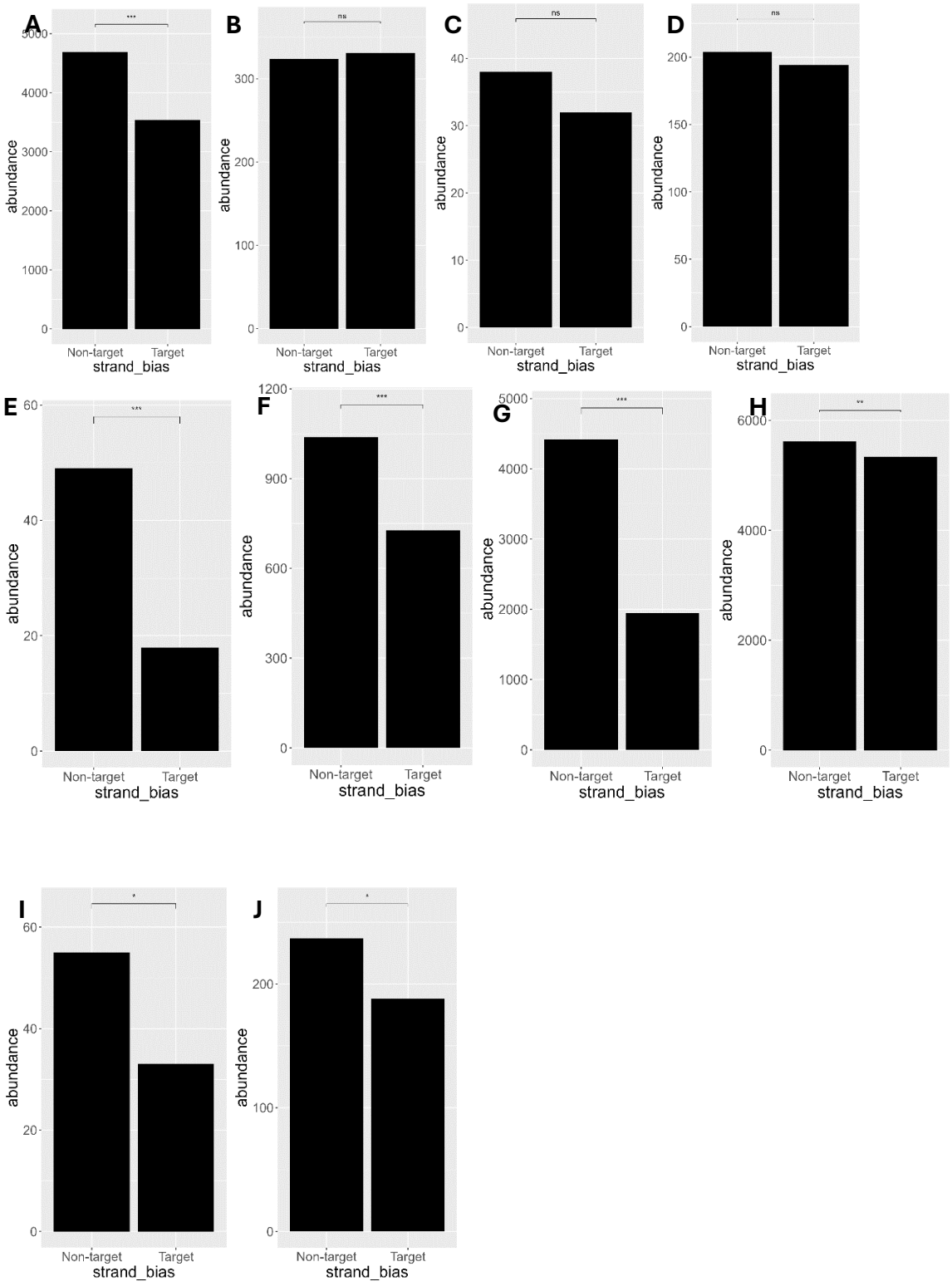


Figure S5.6: ***Spacer acquisition bias when kmer size = 14 by target/non-target strand of class II CRISPR-Cas systems.*** *(A) Type II, (B) Type V-A, (C) Type V-B, (D) Type V-F1, (E) Type VI-B, (F) Type I-B , (G) Type III-A, (H) Type I-F, (I) Type III-B (J) Type I-D. Significance was determined using binomial probability.*

Subtype	D (sense)	D (antisense)
V-A	0.056331	0.088044
V-B	0.15755	0.13031
V-F1	0.073555	0.075721
VI-B	0.048567	0.13009
VI-D	0.084038	0.30026
I-A	0.10371	0.088209
I-B	0.084548	0.037388
I-D	0.053771	0.035037
I-F	0.19907	0.21457
III-A	0.33524	0.074262
III-B	0.1146	0.087945
II	0.053263	0.051259

Table S5.1: *Cohen's D scores for each subtype upon which spacer distribution analysis was conducted.*

Subtype	N (sense)	N (antisense)
V-A	135	131
V-B	22	21
V-F1	207	159
VI-B	33	3
VI-D	15	11
I-A	300	237
I-B	1118	647
I-D	200	167
I-F	6885	7029
III-A	4993	1149
III-B	46	16
II	7126	5884

Table S5.2: *Cohen's D scores for each subtype upon which spacer distribution analysis was conducted.*

Stat power	n	Non-target strand	Target Strand
Cas12a	266	7%	38%
Cas12b	43	8%	4%
Cas12f1	366	48%	41%
IIIB	62	7%	11%
IIIE	143	13%	7%
13b	36	4%	21%
13d	26	2.60%	27%
IA	537	91%	61%
ID	368	1%	1%
IB	1765	100%	100%

Table S5.3: *Statistical power calculations for the acquisition biases presented in Figures 3-5. Subtypes with large sample sizes (1000+) were excluded from the power calculation as sample sizes were assumed to be well powered.*

Stat power	n	Non-target strand	Target Strand
IB	1765	93%	1%
IIIB	88	5%	4%
ID	425	1%	0.50%
Cas12a	655	28%	31%
Cas12b	70	58%	2.60%
Cas13b	67	8%	25%
Cas12f1	398	100%	100%

Table S5.4: *Statistical power calculations for the acquisition biases for the union of complete and partial spacer matches presented in Figure 9.*

	Kmer size (bp)			
Subtype	**10**	**12**	**14**	**Whole match**
I-B	**4542**	**3333**	**1765**	**1765**
V-A	**5033**	**1686**	**655**	**266**
V-B	**569**	**234**	**70**	**43**
V-F1	**2232**	**1068**	**398**	**366**
VI-B	**866**	**220**	**67**	**36**
III-B	**541**	**320**	**88**	**62**
I-D	**1223**	**913**	**425**	**368**
II	**24404**	**13831**	**8225**	**13010**
I-F	**24273**	**11943**	**10957**	**13914**
III-A	**27082**	**15276**	**6365**	**6140**
III-E	**N/A**	**N/A**	**N/A**	**143**

Table S5.5: *Number of pooled spacer mappings for kmer size = 10,12,14 and complete spacer mappings respectively.*

References

1 Suttle, C. A. Marine viruses — major players in the global ecosystem. *Nature Reviews Microbiology* **5**, 801-812 (2007). https://doi.org/10.1038/nrmicro1750

2 Comeau, A. M. *et al.* Exploring the prokaryotic virosphere. *Research in Microbiology* **159**, 306-313 (2008). https://doi.org/https://doi.org/10.1016/j.resmic.2008.05.001

3 Suttle, C. A. Viruses in the sea. *Nature* **437**, 356-361 (2005). https://doi.org/10.1038/nature04160

4 Fuhrman, J. A. Marine viruses and their biogeochemical and ecological effects. *Nature* **399**, 541-548 (1999). https://doi.org/10.1038/21119

5 Falkowski, P. *et al.* The Global Carbon Cycle: A Test of Our Knowledge of Earth as a System. *Science* **290**, 291-296 (2000). https://doi.org/10.1126/science.290.5490.291

6 Heneghan, R. F. *et al.* The global distribution and climate resilience of marine heterotrophic prokaryotes. *Nature Communications* **15**, 6943 (2024). https://doi.org/10.1038/s41467-024-50635-z

7 Nguyen, H. M. *et al.* RNA and Single-Stranded DNA Phages: Unveiling the Promise from the Underexplored World of Viruses. *International Journal of Molecular Sciences* **24** (2023).

8 Callanan, J. *et al.* Expansion of known ssRNA phage genomes: From tens to over a thousand. *Science Advances* **6**, eaay5981 https://doi.org/10.1126/sciadv.aay5981

9 Tabib-Salazar, A., Mulvenna, N., Severinov, K., Matthews, S. J. & Wigneshweraraj, S. Xenogeneic Regulation of the Bacterial Transcription Machinery. *Journal of Molecular Biology* **431**, 4078-4092 (2019). https://doi.org/https://doi.org/10.1016/j.jmb.2019.02.008

10 Mueser, T. C., Hinerman, J. M., Devos, J. M., Boyer, R. A. & Williams, K. J. Structural analysis of bacteriophage T4 DNA replication: a review in the Virology Journal series on bacteriophage T4 and its relatives. *Virology Journal* **7**, 359 (2010). https://doi.org/10.1186/1743-422X-7-359

11 Belanger, K. G., Mirzayan, C., Kreuzer, H. E., Alberts, B. M. & Kreuzer, K. N. Two-Dimensional Gel Analysis of Rolling Circle Replication in the Presence and Absence of Bacteriophage T4 Primase. *Nucleic Acids Research* **24**, 2166-2175 (1996). https://doi.org/10.1093/nar/24.11.2166

12 Rao, V. B., Fokine, A., Fang, Q. & Shao, Q. Bacteriophage T4 Head: Structure, Assembly, and Genome Packaging. *Viruses* **15** (2023).

13 Krieger, I. V. *et al.* The Structural Basis of T4 Phage Lysis Control: DNA as the Signal for Lysis Inhibition. *Journal of Molecular Biology* **432**, 4623-4636 (2020). https://doi.org/https://doi.org/10.1016/j.jmb.2020.06.013

14 Ramanculov, E. & Young, R. Functional analysis of the phage T4 holin in a λ context. *Molecular Genetics and Genomics* **265**, 345-353 (2001). https://doi.org/10.1007/s004380000422

15 Ramanculov, E. & Young, R. Genetic analysis of the T4 holin: timing and topology. *Gene* **265**, 25-36 (2001). https://doi.org/https://doi.org/10.1016/S0378-1119(01)00365-1

16 Landy, A. The λ Integrase Site-specific Recombination Pathway. *Microbiology Spectrum* **3**, 10.1128/microbiolspec.mdna1123-0051-2014 (2015). https://doi.org/10.1128/microbiolspec.mdna3-0051-2014

17 Roberts, J. W. & Roberts, C. W. Proteolytic cleavage of bacteriophage lambda repressor in induction. *Proceedings of the National Academy of Sciences* **72**, 147-151 (1975). https://doi.org/10.1073/pnas.72.1.147

18 Meyer, B. J. & Ptashne, M. Gene regulation at the right operator (OR) of bacteriophage λ: III. λ Repressor directly activates gene transcription. *Journal of Molecular Biology* **139**, 195-205 (1980). https://doi.org/https://doi.org/10.1016/0022-2836(80)90304-6

19 Maurer, R., Meyer, B. J. & Ptashne, M. Gene regulation at the right operator (OR) of bacteriophage λ: I. OR3 and autogenous negative control by repressor. *Journal of Molecular Biology* **139**, 147-161 (1980). https://doi.org/https://doi.org/10.1016/0022-2836(80)90302-2

20 Bobay, L.-M., Rocha, E. P. C. & Touchon, M. The Adaptation of Temperate Bacteriophages to Their Host Genomes. *Molecular Biology and Evolution* **30**, 737-751 (2013). https://doi.org/10.1093/molbev/mss279

21 Millman, A. *et al.* An expanded arsenal of immune systems that protect bacteria from phages. *Cell Host & Microbe* **30**, 1556-1569.e1555 (2022). https://doi.org/https://doi.org/10.1016/j.chom.2022.09.017

22 Doron, S. *et al.* Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, eaar4120 (2018). https://doi.org/doi:10.1126/science.aar4120

23 Wu, Y. *et al.* Bacterial defense systems exhibit synergistic anti-phage activity. *Cell Host & Microbe* **32**, 557-572.e556 (2024). https://doi.org/https://doi.org/10.1016/j.chom.2024.01.015

24 Oude Essink, B. B. & Berkhout, B. The restriction enzyme Ban l is inhibited by dcm - methylation of the GGCGC m5 C site. *Nucleic Acids Research* **22**, 108-108 (1994). https://doi.org/10.1093/nar/22.1.108

25 Xu, Q., Morgan, R. D., Roberts, R. J. & Blaser, M. J. Identification of type II restriction and modification systems in Helicobacter pylori reveals their substantial diversity among strains. *Proceedings of the National Academy of Sciences* **97**, 9671-9676 (2000). https://doi.org/10.1073/pnas.97.17.9671

26 Roer, L., Aarestrup Frank, M. & Hasman, H. The EcoKI Type I Restriction-Modification System in Escherichia coli Affects but Is Not an Absolute Barrier for Conjugation. *Journal of Bacteriology* **197**, 337-342 (2014). https://doi.org/10.1128/jb.02418-14

27 Roberts, R. J., Vincze, T., Posfai, J. & Macelis, D. REBASE—a database for DNA restriction and modification: enzymes, genes and genomes. *Nucleic Acids Research* **38**, D234-D236 (2010). https://doi.org/10.1093/nar/gkp874

28 Pingoud, A., Wilson, G. G. & Wende, W. Type II restriction endonucleases—a historical perspective and more. *Nucleic Acids Research* **42**, 7489-7527 (2014). https://doi.org/10.1093/nar/gku447

29 Went, S. C. *et al.* Structure and rational engineering of the PglX methyltransferase and specificity factor for BREX phage defence. *Nature Communications* **15**, 7236 (2024). https://doi.org/10.1038/s41467-024-51629-7

30 Goldfarb, T. *et al.* BREX is a novel phage resistance system widespread in microbial genomes. *The EMBO Journal* **34**, 169-183-183 (2015). https://doi.org/https://doi.org/10.15252/embj.201489455

31 Shen, B. W. *et al.* Structure, substrate binding and activity of a unique AAA+ protein: the BrxL phage restriction factor. *Nucleic Acids Research* **51**, 3513-3528 (2023). https://doi.org/10.1093/nar/gkad083

32 Ofir, G. *et al.* DISARM is a widespread bacterial defence system with broad anti-phage activities. *Nature Microbiology* **3**, 90-98 (2018). https://doi.org/10.1038/s41564-017-0051-0

33 Bravo, J. P. K., Aparicio-Maldonado, C., Nobrega, F. L., Brouns, S. J. J. & Taylor, D. W. Structural basis for broad anti-phage immunity by DISARM. *Nature Communications* **13**, 2987 (2022). https://doi.org/10.1038/s41467-022-30673-1

34 Xiong, X. *et al.* SspABCD–SspE is a phosphorothioation-sensing bacterial defence system with broad anti-phage activities. *Nature Microbiology* **5**, 917-928 (2020). https://doi.org/10.1038/s41564-020-0700-6

35 Wang, L. *et al.* Phosphorothioation of DNA in bacteria by dnd genes. *Nature Chemical Biology* **3**, 709-710 (2007). https://doi.org/10.1038/nchembio.2007.39
36 Zhou, X. *et al.* A novel DNA modification by sulphur. *Molecular Microbiology* **57**, 1428-1438 (2005). https://doi.org/https://doi.org/10.1111/j.1365-2958.2005.04764.x
37 Makarova, K. S., Wolf, Y. I., van der Oost, J. & Koonin, E. V. Prokaryotic homologs of Argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. *Biology Direct* **4**, 29 (2009). https://doi.org/10.1186/1745-6150-4-29
38 Kuzmenko, A. *et al.* DNA targeting and interference by a bacterial Argonaute nuclease. *Nature* **587**, 632-637 (2020). https://doi.org/10.1038/s41586-020-2605-1
39 Esyunina, D. *et al.* Specific targeting of plasmids with Argonaute enables genome editing. *Nucleic Acids Research* **51**, 4086-4099 (2023). https://doi.org/10.1093/nar/gkad191
40 Lu, X., Xiao, J., Wang, L., Zhu, B. & Huang, F. The nuclease-associated short prokaryotic Argonaute system nonspecifically degrades DNA upon activation by target recognition. *Nucleic Acids Research* **52**, 844-855 (2024). https://doi.org/10.1093/nar/gkad1145
41 Koopal, B. *et al.* Short prokaryotic Argonaute systems trigger cell death upon detection of invading DNA. *Cell* **185**, 1471-1486.e1419 (2022). https://doi.org/10.1016/j.cell.2022.03.012
42 Zaremba, M. *et al.* Short prokaryotic Argonautes provide defence against incoming mobile genetic elements through NAD+ depletion. *Nature Microbiology* **7**, 1857-1869 (2022). https://doi.org/10.1038/s41564-022-01239-0
43 Hu, H. *et al.* Structure and mechanism of Zorya anti-phage defense system. *bioRxiv*, 2023.2012.2018.572097 (2023). https://doi.org/10.1101/2023.12.18.572097
44 Lopatina, A., Tal, N. & Sorek, R. Abortive Infection: Bacterial Suicide as an Antiviral Immune Strategy. *Annual Review of Virology* **7**, 371-384 (2020). https://doi.org/https://doi.org/10.1146/annurev-virology-011620-040628
45 Burman, N. *et al.* A virally encoded tRNA neutralizes the PARIS antiviral defence system. *Nature* **634**, 424-431 (2024). https://doi.org/10.1038/s41586-024-07874-3
46 Cheng, R. *et al.* Prokaryotic Gabija complex senses and executes nucleotide depletion and DNA cleavage for antiviral defense. *Cell Host & Microbe* **31**, 1331-1344.e1335 (2023). https://doi.org/https://doi.org/10.1016/j.chom.2023.06.014
47 Dy, R. L., Przybilski, R., Semeijn, K., Salmond, G. P. C. & Fineran, P. C. A widespread bacteriophage abortive infection system functions through a Type IV toxin–antitoxin mechanism. *Nucleic Acids Research* **42**, 4590-4605 (2014). https://doi.org/10.1093/nar/gkt1419
48 Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I. & Koonin, E. V. 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 (2006). https://doi.org/10.1186/1745-6150-1-7
49 Bibi-Triki, S. *et al.* Functional and Structural Analysis of HicA3-HicB3, a Novel Toxin-Antitoxin System of Yersinia pestis. *Journal of Bacteriology* **196**, 3712-3723 (2014). https://doi.org/10.1128/jb.01932-14
50 Manav, M. C. *et al.* The E. coli HicB Antitoxin Contains a Structurally Stable Helix-Turn-Helix DNA Binding Domain. *Structure* **27**, 1675-1685.e1673 (2019). https://doi.org/10.1016/j.str.2019.08.008
51 Kazlauskiene, M., Kostiuk, G., Venclovas, Č., Tamulaitis, G. & Siksnys, V. A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. *Science* **357**, 605-609 (2017). https://doi.org/10.1126/science.aao0100

52 Guegler, C. K. & Laub, M. T. Shutoff of host transcription triggers a toxin-antitoxin system to cleave phage RNA and abort infection. *Molecular Cell* **81**, 2361-2373.e2369 (2021). https://doi.org/10.1016/j.molcel.2021.03.027
53 Millman, A. *et al.* Bacterial Retrons Function In Anti-Phage Defense. *Cell* **183**, 1551-1561.e1512 (2020). https://doi.org/https://doi.org/10.1016/j.cell.2020.09.065
54 Cohen, D. *et al.* Cyclic GMP–AMP signalling protects bacteria against viral infection. *Nature* **574**, 691-695 (2019). https://doi.org/10.1038/s41586-019-1605-5
55 Garb, J. *et al.* Multiple phage resistance systems inhibit infection via SIR2-dependent NAD+ depletion. *Nature Microbiology* **7**, 1849-1856 (2022). https://doi.org/10.1038/s41564-022-01207-8
56 Baca, C. F. *et al.* The CRISPR effector Cam1 mediates membrane depolarization for phage defence. *Nature* **625**, 797-804 (2024). https://doi.org/10.1038/s41586-023-06902-y
57 Makarova, K. S. *et al.* Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nature Reviews Microbiology* **18**, 67-83 (2020). https://doi.org/10.1038/s41579-019-0299-x
58 Garcia-Doval, C. *et al.* Activation and self-inactivation mechanisms of the cyclic oligoadenylate-dependent CRISPR ribonuclease Csm6. *Nature Communications* **11**, 1596 (2020). https://doi.org/10.1038/s41467-020-15334-5
59 Molina, R. *et al.* Structure of Csx1-cOA4 complex reveals the basis of RNA decay in Type III-B CRISPR-Cas. *Nature Communications* **10**, 4302 (2019). https://doi.org/10.1038/s41467-019-12244-z
60 Makarova, K. S. *et al.* Evolutionary and functional classification of the CARF domain superfamily, key sensors in prokaryotic antivirus defense. *Nucleic Acids Research* **48**, 8828-8847 (2020). https://doi.org/10.1093/nar/gkaa635
61 Leavitt Justin, C. *et al.* Bacteriophage P22 SieA-mediated superinfection exclusion. *mBio* **15**, e02169-02123 (2024). https://doi.org/10.1128/mbio.02169-23
62 Mojica, F. J. M., Ferrer, C., Juez, G. & Rodríguez-Valera, F. Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning. *Molecular Microbiology* **17**, 85-93 (1995). https://doi.org/https://doi.org/10.1111/j.1365-2958.1995.mmi_17010085.x
63 Mojica, F. J. M., Díez-Villaseñor, C. s., García-Martínez, J. & Soria, E. Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. *Journal of Molecular Evolution* **60**, 174-182 (2005). https://doi.org/10.1007/s00239-004-0046-3
64 Jansen, R., Embden, J. D. A. v., Gaastra, W. & Schouls, L. M. Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular Microbiology* **43**, 1565-1575 (2002). https://doi.org/https://doi.org/10.1046/j.1365-2958.2002.02839.x
65 Barrangou, R. *et al.* CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. *Science* **315**, 1709-1712 (2007). https://doi.org/10.1126/science.1138140
66 Yosef, I., Goren, M. G. & Qimron, U. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Research* **40**, 5569-5576 (2012). https://doi.org/10.1093/nar/gks216
67 Swarts, D. C., Mosterd, C., van Passel, M. W. J. & Brouns, S. J. J. CRISPR Interference Directs Strand Specific Spacer Acquisition. *PLOS ONE* **7**, e35888 (2012). https://doi.org/10.1371/journal.pone.0035888
68 Datsenko, K. A. *et al.* Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nature Communications* **3**, 945 (2012). https://doi.org/10.1038/ncomms1937

69 Nunez, J. K. *et al.* Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat Struct Mol Biol* **21**, 528-534 (2014). https://doi.org/10.1038/nsmb.2820
70 Nunez, J. K., Lee, A. S., Engelman, A. & Doudna, J. A. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* **519**, 193-198 (2015). https://doi.org/10.1038/nature14237
71 Brouns, S. J. J. *et al.* Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes. *Science* **321**, 960-964 (2008). https://doi.org/10.1126/science.1159689
72 Lee, H., Dhingra, Y. & Sashital, D. G. The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. *eLife* **8**, e44248 (2019). https://doi.org/10.7554/eLife.44248
73 Kim, S. *et al.* Selective loading and processing of prespacers for precise CRISPR adaptation. *Nature* **579**, 141-145 (2020). https://doi.org/10.1038/s41586-020-2018-1
74 Nunez, J. K., Bai, L., Harrington, L. B., Hinder, T. L. & Doudna, J. A. CRISPR Immunological Memory Requires a Host Factor for Specificity. *Mol Cell* **62**, 824-833 (2016). https://doi.org/10.1016/j.molcel.2016.04.027
75 Deltcheva, E. *et al.* CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* **471**, 602-607 (2011). https://doi.org/10.1038/nature09886
76 Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012). https://doi.org/10.1126/science.1225829
77 Nishimasu, H. *et al.* Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935-949 (2014). https://doi.org/10.1016/j.cell.2014.02.001
78 Makarova, K. S. *et al.* An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* **13**, 722-736 (2015). https://doi.org/10.1038/nrmicro3569
79 Calhoun, S. *et al.* Prediction of enzymatic pathways by integrative pathway mapping. *eLife* **7**, e31097 (2018). https://doi.org/10.7554/eLife.31097
80 Moreno-Hagelsieb, G. & Santoyo, G. in *Prokaryotic Systems Biology* (eds PhD Nevan J. Krogan & PhD Mohan Babu) 97-106 (Springer International Publishing, 2015).
81 Shmakov, S. *et al.* Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. *Molecular Cell* **60**, 385-397 (2015). https://doi.org/https://doi.org/10.1016/j.molcel.2015.10.008
82 Shmakov, S. *et al.* Diversity and evolution of class 2 CRISPR-Cas systems. *Nat Rev Microbiol* **15**, 169-182 (2017). https://doi.org/10.1038/nrmicro.2016.184
83 Al-Shayeb, B. *et al.* Clades of huge phages from across Earth's ecosystems. *Nature* **578**, 425-431 (2020). https://doi.org/10.1038/s41586-020-2007-4
84 Burstein, D. *et al.* New CRISPR–Cas systems from uncultivated microbes. *Nature* **542**, 237-241 (2017). https://doi.org/10.1038/nature21059
85 Andreas, M. P. & Giessen, T. W. Large-scale computational discovery and analysis of virus-derived microbial nanocompartments. *Nature Communications* **12**, 4748 (2021). https://doi.org/10.1038/s41467-021-25071-y
86 Tesson, F. *et al.* Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nature Communications* **13**, 2561 (2022). https://doi.org/10.1038/s41467-022-30269-9
87 Smargon, A. A. *et al.* Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28. *Molecular Cell* **65** (2017/02/16). https://doi.org/10.1016/j.molcel.2016.12.023
88 Yan, W. X. *et al.* Cas13d Is a Compact RNA-Targeting Type VI CRISPR Effector Positively Modulated by a WYL-Domain-Containing Accessory Protein. *Mol Cell* (2018). https://doi.org/10.1016/j.molcel.2018.02.028
89 Konermann, S. *et al.* Transcriptome Engineering with RNA-Targeting Type VI-D CRISPR Effectors. *Cell* (2018). https://doi.org/10.1016/j.cell.2018.02.033

90 Shmakov, S. A. *et al.* Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. *Proceedings of the National Academy of Sciences* **115** (2018-6-5). https://doi.org/10.1073/pnas.1803440115
91 Shmakov, S. A. *et al.* Systematic prediction of functionally linked genes in bacterial and archaeal genomes. *Nature Protocols* **14**, 3013-3031 (2019). https://doi.org/10.1038/s41596-019-0211-1
92 Rousset, F. *et al.* Phages and their satellites encode hotspots of antiviral systems. *Cell Host & Microbe* **30**, 740-753.e745 (2022). https://doi.org/10.1016/j.chom.2022.02.018
93 Edgar, R. C. PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics* **8**, 18 (2007). https://doi.org/10.1186/1471-2105-8-18
94 Bland, C. *et al.* CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* **8**, 209 (2007). https://doi.org/10.1186/1471-2105-8-209
95 Biswas, A., Staals, R. H. J., Morales, S. E., Fineran, P. C. & Brown, C. M. CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics* **17**, 356 (2016). https://doi.org/10.1186/s12864-016-2627-0
96 Grissa, I., Vergnaud, G. & Pourcel, C. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Research* **35**, W52-W57 (2007). https://doi.org/10.1093/nar/gkm360
97 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021). https://doi.org/10.1038/s41586-021-03819-2
98 van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nature Biotechnology* **42**, 243-246 (2024). https://doi.org/10.1038/s41587-023-01773-0
99 Duan, N., Hand, E., Pheko, M., Sharma, S. & Emiola, A. Structure-guided discovery of anti-CRISPR and anti-phage defense proteins. *Nature Communications* **15**, 649 (2024). https://doi.org/10.1038/s41467-024-45068-7
100 Yoon, P. H. *et al.* Structure-guided discovery of ancestral CRISPR-Cas13 ribonucleases. *Science* **385**, 538-543 (2024). https://doi.org/10.1126/science.adq0553
101 Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010). https://doi.org/10.1186/1471-2105-11-119
102 Altae-Tran, H. *et al.* Uncovering the functional diversity of rare CRISPR-Cas systems with deep terascale clustering. *Science* **382** (2023-11-24). https://doi.org/10.1126/science.adi1910
103 Sousa, T. d. J. *et al.* Re-sequencing and optical mapping reveals misassemblies and real inversions on Corynebacterium pseudotuberculosis genomes. *Scientific Reports* **9**, 16387 (2019). https://doi.org/10.1038/s41598-019-52695-4
104 Vassallo, C. N., Doering, C. R., Littlehale, M. L., Teodoro, G. I. C. & Laub, M. T. Author Correction: A functional selection reveals previously undetected anti-phage defence systems in the E. coli pangenome. *Nature Microbiology* **9**, 2760-2761 (2024). https://doi.org/10.1038/s41564-024-01724-8
105 Trigodet, F., Sachdeva, R., Banfield, J. F. & Eren, A. M. Assemblies of long-read metagenomes suffer from diverse errors. *bioRxiv*, 2025.2004.2022.649783 (2025). https://doi.org/10.1101/2025.04.22.649783
106 Chen, Q., Zobel, J. & Verspoor, K. Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study. *Database* **2017** (2017). https://doi.org/10.1093/database/baw163
107 Sayers, Eric W. *et al.* Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Research* **53**, D20-D29 (2025). https://doi.org/10.1093/nar/gkae979

108 Lasken, R. S. & McLean, J. S. Recent advances in genomic DNA sequencing of microbial species from single cells. *Nature Reviews Genetics* **15**, 577-584 (2014). https://doi.org/10.1038/nrg3785

109 Huang, J. *et al.* Discovery of deaminase functions by structure-based protein clustering. *Cell* **186**, 3182-3195.e3114 (2023). https://doi.org/https://doi.org/10.1016/j.cell.2023.05.041

110 Yan, W. X. *et al.* Functionally diverse type V CRISPR-Cas systems. *Science* **363**, 88-91 (2019). https://doi.org/doi:10.1126/science.aav7271

111 Xu, C. *et al.* Programmable RNA editing with compact CRISPR–Cas13 systems from uncultivated microbes. *Nature Methods* **18**, 499-506 (2021). https://doi.org/10.1038/s41592-021-01124-4

112 Faure, G. *et al.* CRISPR–Cas in mobile genetic elements: counter-defence and beyond. *Nature Reviews Microbiology* **17**, 513-525 (2019). https://doi.org/10.1038/s41579-019-0204-7

113 Koonin, E. V. & Makarova, K. S. Origins and evolution of CRISPR-Cas systems. *Philosophical Transactions of the Royal Society B: Biological Sciences* **374**, 20180087 (2019). https://doi.org/10.1098/rstb.2018.0087

114 Zhang, X., Garrett, S., Graveley, B. R. & Terns, M. P. Unique properties of spacer acquisition by the type III-A CRISPR-Cas system. *Nucleic Acids Research* **50**, 1562-1582 (2022). https://doi.org/10.1093/nar/gkab1193

115 Lin, J., Shen, Y., Ni, J. & She, Q. A type III-A CRISPR–Cas system mediates co-transcriptional DNA cleavage at the transcriptional bubbles in close proximity to active effectors. *Nucleic Acids Research* **49**, 7628-7643 (2021). https://doi.org/10.1093/nar/gkab590

116 Liu, T. Y., Liu, J.-J., Aditham, A. J., Nogales, E. & Doudna, J. A. Target preference of Type III-A CRISPR-Cas complexes at the transcription bubble. *Nature Communications* **10**, 3001 (2019). https://doi.org/10.1038/s41467-019-10780-2

117 Kazlauskiene, M., Tamulaitis, G., Kostiuk, G., Venclovas, Č. & Siksnys, V. Spatiotemporal Control of Type III-A CRISPR-Cas Immunity: Coupling DNA Degradation with the Target RNA Recognition. *Molecular Cell* **62**, 295-306 (2016). https://doi.org/10.1016/j.molcel.2016.03.024

118 Samai, P. *et al.* Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. *Cell* **161**, 1164-1174 (2015). https://doi.org/10.1016/j.cell.2015.04.027

119 Taylor, H. N. *et al.* Positioning Diverse Type IV Structures and Functions Within Class 1 CRISPR-Cas Systems. *Frontiers in Microbiology* **12** (2021). https://doi.org/10.3389/fmicb.2021.671522

120 Altae-Tran, H. *et al.* Diversity, evolution, and classification of the RNA-guided nucleases TnpB and Cas12. *Proceedings of the National Academy of Sciences* **120** (2023-11-20). https://doi.org/10.1073/pnas.2308224120

121 Meers, C. *et al.* Transposon-encoded nucleases use guide RNAs to promote their selfish spread. *Nature* **622**, 863-871 (2023). https://doi.org/10.1038/s41586-023-06597-1

122 Wiegand, T. *et al.* TnpB homologues exapted from transposons are RNA-guided transcription factors. *Nature* **631**, 439-448 (2024). https://doi.org/10.1038/s41586-024-07598-4

123 Harrington, L. B. *et al.* Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. *Science* **362**, 839-842 (2018). https://doi.org/10.1126/science.aav4294

124 Altae-Tran, H. *et al.* The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases. *Science* **374**, 57-65 (2021). https://doi.org/10.1126/science.abj6856

125 Kato, K. *et al.* Structure of the IscB–ωRNA ribonucleoprotein complex, the likely ancestor of CRISPR-Cas9. *Nature Communications* **13**, 6719 (2022). https://doi.org/10.1038/s41467-022-34378-3
126 Xiao, Y. *et al.* Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System. *Cell* **170**, 48-60 e11 (2017). https://doi.org/10.1016/j.cell.2017.06.012
127 Richter, C. *et al.* Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. *Nucleic Acids Research* **42**, 8516-8526 (2014). https://doi.org/10.1093/nar/gku527
128 Sinkunas, T. *et al.* Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *The EMBO Journal* **30**, 1335-1342-1342 (2011). https://doi.org/https://doi.org/10.1038/emboj.2011.41
129 Beloglazova, N. *et al.* CRISPR RNA binding and DNA target recognition by purified Cascade complexes from Escherichia coli. *Nucleic Acids Research* **43**, 530-543 (2015). https://doi.org/10.1093/nar/gku1285
130 Gong, L. *et al.* Primed adaptation tolerates extensive structural and size variations of the CRISPR RNA guide in Haloarcula hispanica. *Nucleic Acids Research* **47**, 5880-5891 (2019). https://doi.org/10.1093/nar/gkz244
131 Staals, R. H. J. *et al.* Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR–Cas system. *Nature Communications* **7**, 12853 (2016). https://doi.org/10.1038/ncomms12853
132 Semenova, E. *et al.* Highly efficient primed spacer acquisition from targets destroyed by the Escherichia coli type I-E CRISPR-Cas interfering complex. *Proceedings of the National Academy of Sciences* **113**, 7626-7631 (2016). https://doi.org/10.1073/pnas.1602639113
133 Dillard, K. E. *et al.* Assembly and Translocation of a CRISPR-Cas Primed Acquisition Complex. *Cell* **175**, 934-946.e915 (2018). https://doi.org/https://doi.org/10.1016/j.cell.2018.09.039
134 Peters, J. E., Makarova, K. S., Shmakov, S. & Koonin, E. V. Recruitment of CRISPR-Cas systems by Tn7-like transposons. *Proceedings of the National Academy of Sciences* **114**, E7358-E7366 (2017). https://doi.org/10.1073/pnas.1709035114
135 Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219-225 (2019). https://doi.org/10.1038/s41586-019-1323-z
136 Vo, P. L. H. *et al.* CRISPR RNA-guided integrases for high-efficiency, multiplexed bacterial genome engineering. *Nature Biotechnology* **39**, 480-489 (2021). https://doi.org/10.1038/s41587-020-00745-y
137 Lin, J., Feng, M., Zhang, H. & She, Q. Characterization of a novel type III CRISPR-Cas effector provides new insights into the allosteric activation and suppression of the Cas10 DNase. *Cell Discovery* **6**, 29 (2020). https://doi.org/10.1038/s41421-020-0160-4
138 Sheppard, N. F., Glover, C. V. C., Terns, R. M. & Terns, M. P. The CRISPR-associated Csx1 protein of Pyrococcus furiosus is an adenosine-specific endoribonuclease. *RNA* **22**, 216-224 (2016). https://doi.org/10.1261/rna.039842.113
139 Rostøl, J. T. *et al.* The Card1 nuclease provides defence during type III CRISPR immunity. *Nature* **590**, 624-629 (2021). https://doi.org/10.1038/s41586-021-03206-x
140 Hoikkala, V. *et al.* Cooperation between Different CRISPR-Cas Types Enables Adaptation in an RNA-Targeting System. *mBio* **12**, 10.1128/mbio.03338-03320 (2021). https://doi.org/10.1128/mbio.03338-20
141 Steens, J. A., Salazar, C. R. P. & Staals, R. H. J. The diverse arsenal of type III CRISPR–Cas-associated CARF and SAVED effectors. *Biochemical Society Transactions* **50**, 1353-1364 (2022). https://doi.org/10.1042/BST20220289

142 Pinilla-Redondo, R. *et al.* Type IV CRISPR–Cas systems are highly diverse and involved in competition between plasmids. *Nucleic Acids Research* **48**, 2000-2012 (2020). https://doi.org/10.1093/nar/gkz1197
143 Cui, N. *et al.* Type IV-A CRISPR-Csf complex: Assembly, dsDNA targeting, and CasDinG recruitment. *Molecular Cell* **83**, 2493-2508.e2495 (2023). https://doi.org/https://doi.org/10.1016/j.molcel.2023.05.036
144 Crowley, V. M. *et al.* A Type IV-A CRISPR-Cas System in Pseudomonas aeruginosa Mediates RNA-Guided Plasmid Interference In Vivo. *The CRISPR Journal* **2**, 434-440 (2019). https://doi.org/10.1089/crispr.2019.0048
145 Yang, J. *et al.* Structural basis for the activity of the type VII CRISPR–Cas system. *Nature* **633**, 465-472 (2024). https://doi.org/10.1038/s41586-024-07815-0
146 Tenjo-Castaño, F. *et al.* Structure of the TnsB transposase-DNA complex of type V-K CRISPR-associated transposon. *Nature Communications* **13**, 5792 (2022). https://doi.org/10.1038/s41467-022-33504-5
147 Abudayyeh, O. O. *et al.* RNA targeting with CRISPR–Cas13. *Nature* **550**, 280-284 (2017). https://doi.org/10.1038/nature24049
148 Molina-Sánchez, M. D., Martínez-Abarca, F., Casamayor, V. M., Mestre, M. R. & Toro, N. Spacer acquisition in type VI CRISPR-Cas systems associated with reverse transcriptase-Cas1 fusion proteins. *bioRxiv*, 2024.2003.2012.584598 (2024). https://doi.org/10.1101/2024.03.12.584598
149 VanderWal, A. R. *et al.* Csx28 is a membrane pore that enhances CRISPR-Cas13b–dependent antiphage defense. *Science* **380**, 410-415 (2023). https://doi.org/doi:10.1126/science.abm1184
150 Nakagawa, R. *et al.* Structure and engineering of the minimal type VI CRISPR-Cas13bt3. *Molecular Cell* **82**, 3178-3192.e3175 (2022). https://doi.org/10.1016/j.molcel.2022.08.001
151 Iranzo, J., Krupovic, M. & Koonin Eugene, V. The Double-Stranded DNA Virosphere as a Modular Hierarchical Network of Gene Sharing. *mBio* **7**, 10.1128/mbio.00978-00916 (2016). https://doi.org/10.1128/mbio.00978-16
152 Bin Jang, H. *et al.* Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nature Biotechnology* **37**, 632-639 (2019). https://doi.org/10.1038/s41587-019-0100-8
153 Siddell, S. G. *et al.* Virus taxonomy and the role of the International Committee on Taxonomy of Viruses (ICTV). *Journal of General Virology* **104** (2023). https://doi.org/https://doi.org/10.1099/jgv.0.001840
154 Pfeifer, E. & Rocha, E. P. C. Phage-plasmids promote recombination and emergence of phages and plasmids. *Nature Communications* **15**, 1545 (2024). https://doi.org/10.1038/s41467-024-45757-3
155 Simmonds, P. *et al.* Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology* **15**, 161-168 (2017). https://doi.org/10.1038/nrmicro.2016.177
156 Iranzo, J., Koonin Eugene, V., Prangishvili, D. & Krupovic, M. Bipartite Network Analysis of the Archaeal Virosphere: Evolutionary Connections between Viruses and Capsidless Mobile Elements. *Journal of Virology* **90**, 11043-11055 (2016). https://doi.org/10.1128/jvi.01622-16
157 Zafar, N., Mazumder, R. & Seto, D. CoreGenes: A computational tool for identifying and cataloging "core" genes in a set of small genomes. *BMC Bioinformatics* **3**, 12 (2002). https://doi.org/10.1186/1471-2105-3-12
158 Edwards, R. A. & Rohwer, F. Viral metagenomics. *Nature Reviews Microbiology* **3**, 504-510 (2005). https://doi.org/10.1038/nrmicro1163

159 Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 (2008). https://doi.org/10.1088/1742-5468/2008/10/P10008
160 Adler, B. A. *et al.* Broad-spectrum CRISPR-Cas13a enables efficient phage genome editing. *Nature Microbiology* **7**, 1967-1979 (2022). https://doi.org/10.1038/s41564-022-01258-x
161 Mendoza, S. D. *et al.* A bacteriophage nucleus-like compartment shields DNA from CRISPR nucleases. *Nature* **577**, 244-248 (2020). https://doi.org/10.1038/s41586-019-1786-y
162 Malone, L. M. *et al.* A jumbo phage that forms a nucleus-like structure evades CRISPR–Cas DNA targeting but is vulnerable to type III RNA-based immunity. *Nature Microbiology* **5**, 48-55 (2020). https://doi.org/10.1038/s41564-019-0612-5
163 Marino, N. D., Pinilla-Redondo, R. & Bondy-Denomy, J. CRISPR-Cas12a targeting of ssDNA plays no detectable role in immunity. *Nucleic Acids Research* **50**, 6414-6422 (2022). https://doi.org/10.1093/nar/gkac462
164 Künne, T. *et al.* Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. *Molecular Cell* **63**, 852-864 (2016). https://doi.org/10.1016/j.molcel.2016.07.011
165 Jackson, S. A., Birkholz, N., Malone, L. M. & Fineran, P. C. Imprecise Spacer Acquisition Generates CRISPR-Cas Immune Diversity through Primed Adaptation. *Cell Host & Microbe* **25**, 250-260.e254 (2019). https://doi.org/https://doi.org/10.1016/j.chom.2018.12.014
166 Semenova, E. *et al.* Highly efficient primed spacer acquisition from targets destroyed by the Escherichia coli type I-E CRISPR-Cas interfering complex. *Proceedings of the National Academy of Sciences* **113**, 7626-7631 (2016). https://doi.org/doi:10.1073/pnas.1602639113
167 Ramachandran, A. & Bailey, S. Memory Upgrade: Insights into Primed Adaptation by CRISPR-Cas Immune Systems. *Molecular Cell* **64**, 641-642 (2016). https://doi.org/10.1016/j.molcel.2016.11.008
168 Xue, C., Whitis, N. R. & Sashital, D. G. Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity. *Molecular Cell* **64**, 826-834 (2016). https://doi.org/10.1016/j.molcel.2016.09.033
169 George, N. A. & Hug, L. A. CRISPR-resolved virus-host interactions in a municipal landfill include non-specific viruses, hyper-targeted viral populations, and interviral conflicts. *Scientific Reports* **13**, 5611 (2023). https://doi.org/10.1038/s41598-023-32078-6
170 Li, M., Wang, R., Zhao, D. & Xiang, H. Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. *Nucleic Acids Research* **42**, 2483-2492 (2014). https://doi.org/10.1093/nar/gkt1154
171 Li, M., Wang, R. & Xiang, H. Haloarcula hispanica CRISPR authenticates PAM of a target sequence to prime discriminative adaptation. *Nucleic Acids Research* **42**, 7226-7235 (2014). https://doi.org/10.1093/nar/gku389
172 Musharova, O. *et al.* Prespacers formed during primed adaptation associate with the Cas1–Cas2 adaptation complex and the Cas3 interference nuclease–helicase. *Proceedings of the National Academy of Sciences* **118**, e2021291118 (2021). https://doi.org/10.1073/pnas.2021291118
173 Vorontsova, D. *et al.* Foreign DNA acquisition by the I-F CRISPR–Cas system requires all components of the interference machinery. *Nucleic Acids Research* **43**, 10848-10860 (2015). https://doi.org/10.1093/nar/gkv1261
174 Fineran, P. C. *et al.* Degenerate target sites mediate rapid primed CRISPR adaptation. *Proceedings of the National Academy of Sciences* **111**, E1629-E1638 (2014). https://doi.org/10.1073/pnas.1400071111

175 Nicholson, T. J. *et al.* Bioinformatic evidence of widespread priming in type I and II CRISPR-Cas systems. *RNA Biology* **16**, 566-576 (2019). https://doi.org/10.1080/15476286.2018.1509662
176 Westra, Edze R. *et al.* CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3. *Molecular Cell* **46**, 595-605 (2012). https://doi.org/10.1016/j.molcel.2012.03.018
177 Morisaka, H. *et al.* CRISPR-Cas3 induces broad and unidirectional genome editing in human cells. *Nature Communications* **10**, 5302 (2019). https://doi.org/10.1038/s41467-019-13226-x
178 Kim, Do Y., Lee, So Y., Ha, Hyun J. & Park, Hyun H. Structural basis of Cas3 activation in type I-C CRISPR-Cas system. *Nucleic Acids Research* **52**, 10563-10574 (2024). https://doi.org/10.1093/nar/gkae723
179 Levy, A. *et al.* CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* **520**, 505-510 (2015). https://doi.org/10.1038/nature14302
180 Radovčić, M. *et al.* CRISPR–Cas adaptation in Escherichia coli requires RecBCD helicase but not nuclease activity, is independent of homologous recombination, and is antagonized by 5' ssDNA exonucleases. *Nucleic Acids Research* **46**, 10173-10183 (2018). https://doi.org/10.1093/nar/gky799
181 Wu, W. Y. *et al.* Adaptation by Type V-A and V-B CRISPR-Cas Systems Demonstrates Conserved Protospacer Selection Mechanisms Between Diverse CRISPR-Cas Types. *The CRISPR Journal* **5**, 536-547 (2022). https://doi.org/10.1089/crispr.2021.0150
182 Aviram, N., Thornal, Ashley N., Zeevi, D. & Marraffini, Luciano A. Different modes of spacer acquisition by the Staphylococcus epidermidis type III-A CRISPR-Cas system. *Nucleic Acids Research* **50**, 1661-1672 (2022). https://doi.org/10.1093/nar/gkab1299
183 Silas, S. *et al.* Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase–Cas1 fusion protein. *Science* **351**, aad4234 (2016). https://doi.org/10.1126/science.aad4234
184 Silas, S. *et al.* On the Origin of Reverse Transcriptase-Using CRISPR-Cas Systems and Their Hyperdiverse, Enigmatic Spacer Repertoires. *mBio* **8**, 10.1128/mbio.00897-00817 (2017). https://doi.org/10.1128/mbio.00897-17
185 Grigoriev, I. V. *et al.* The Genome Portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Research* **40**, D26-D32 (2012). https://doi.org/10.1093/nar/gkr947
186 Chen, I. M. A. *et al.* The IMG/M data management and analysis system v.7: content updates and new features. *Nucleic Acids Research* **51**, D723-D732 (2023). https://doi.org/10.1093/nar/gkac976
187 Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35**, 1026-1028 (2017). https://doi.org/10.1038/nbt.3988
188 Castillo, J. A., Secaira-Morocho, H., Maldonado, S. & Sarmiento, K. N. Diversity and Evolutionary Dynamics of Antiphage Defense Systems in Ralstonia solanacearum Species Complex. *Frontiers in Microbiology* **Volume 11 - 2020** (2020). https://doi.org/10.3389/fmicb.2020.00961
189 Sun, C. L. *et al.* Phage mutations in response to CRISPR diversification in a bacterial population. *Environmental Microbiology* **15**, 463-470 (2013). https://doi.org/https://doi.org/10.1111/j.1462-2920.2012.02879.x
190 Couvin, D. *et al.* CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Research* **46**, W246-W251 (2018). https://doi.org/10.1093/nar/gky425

191 Payne, L. J. *et al.* Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. *Nucleic Acids Research* **49**, 10868-10878 (2021). https://doi.org/10.1093/nar/gkab883
192 Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530-1534 (2020). https://doi.org/10.1093/molbev/msaa015
193 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**, 539 (2011). https://doi.org/https://doi.org/10.1038/msb.2011.75
194 Sievers, F. & Higgins, D. G. Clustal Omega for making accurate alignments of many protein sequences. *Protein Science* **27**, 135-145 (2018). https://doi.org/https://doi.org/10.1002/pro.3290
195 Alkhnbashi, O. S. *et al.* Characterizing leader sequences of CRISPR loci. *Bioinformatics* **32**, i576-i585 (2016). https://doi.org/10.1093/bioinformatics/btw454
196 Biswas, A., Fineran, P. C. & Brown, C. M. Accurate computational prediction of the transcribed strand of CRISPR non-coding RNAs. *Bioinformatics* **30**, 1805-1813 (2014). https://doi.org/10.1093/bioinformatics/btu114
197 Alkhnbashi, O. S. *et al.* CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci. *Bioinformatics* **30**, i489-i496 (2014). https://doi.org/10.1093/bioinformatics/btu459
198 Biswas, A., Gagnon, J. N., Brouns, S. J. J., Fineran, P. C. & Brown, C. M. CRISPRTarget. *RNA Biology* **10**, 817-827 (2013). https://doi.org/10.4161/rna.24046
199 Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE* **11**, e0163962 (2016). https://doi.org/10.1371/journal.pone.0163962
200 Tang, X., Shang, J., Ji, Y. & Sun, Y. PLASMe: a tool to identify PLASMid contigs from short-read assemblies using transformer. *Nucleic Acids Research* **51**, e83-e83 (2023). https://doi.org/10.1093/nar/gkad578
201 Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods* **9**, 173-175 (2012). https://doi.org/10.1038/nmeth.1818
202 Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports* **9**, 5233 (2019). https://doi.org/10.1038/s41598-019-41695-z
203 Fruchterman, T. M. J. & Reingold, E. M. Graph drawing by force-directed placement. *Software: Practice and Experience* **21**, 1129-1164 (1991). https://doi.org/https://doi.org/10.1002/spe.4380211102
204 Csárdi, G. & Nepusz, T.
205 Hale, C. R. *et al.* RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex. *Cell* **139** (2009/11/25). https://doi.org/10.1016/j.cell.2009.07.040
206 Jiang, K. *et al.* Programmable RNA-guided DNA endonucleases are widespread in eukaryotes and their viruses. *Science Advances* **9** (2023-09-29). https://doi.org/10.1126/sciadv.adk0171
207 Shmakov, S. A. *et al.* Widespread CRISPR-derived RNA regulatory elements in CRISPR-Cas systems. *Nucleic Acids Research* **51** (2023/08/25). https://doi.org/10.1093/nar/gkad495
208 Jain, I. *et al.* tRNA anticodon cleavage by target-activated CRISPR-Cas13a effector. *Science Advances* **10**, eadl0164 (2024). https://doi.org/doi:10.1126/sciadv.adl0164
209 Yirmiya, E. *et al.* Phages overcome bacterial immunity via diverse anti-defence proteins. *Nature* **625**, 352-359 (2024). https://doi.org/10.1038/s41586-023-06869-w

210 Hochhauser, D., Millman, A. & Sorek, R. The defense island repertoire of the Escherichia coli pan-genome. *PLOS Genetics* **19**, e1010694 (2023). https://doi.org/10.1371/journal.pgen.1010694
211 Mitrofanov, A. *et al.* CRISPRidentify: identification of CRISPR arrays using machine learning approach. *Nucleic Acids Research* **49**, e20-e20 (2021). https://doi.org/10.1093/nar/gkaa1158
212 Arita, M., Karsch-Mizrachi, I., Cochrane, G. & Collaboration, o. b. o. t. I. N. S. D. The international nucleotide sequence database collaboration. *Nucleic Acids Research* **49**, D121-D124 (2020). https://doi.org/10.1093/nar/gkaa967
213 Hein, S., Scholz, I., Voß, B. & Hess, W. R. Adaptation and modification of three CRISPR loci in two closely related cyanobacteria. *RNA Biology* **10**, 852-864 (2013). https://doi.org/10.4161/rna.24160
214 Makarova, K. S., Anantharaman, V., Grishin, N. V., Koonin, E. V. & Aravind, L. CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems. *Frontiers in Genetics* **5** (2014). https://doi.org/10.3389/fgene.2014.00102
215 Athukoralage, J. S. *et al.* An anti-CRISPR viral ring nuclease subverts type III CRISPR immunity. *Nature* **577**, 572-575 (2020). https://doi.org/10.1038/s41586-019-1909-5
216 Laugel, B. *et al.* Engineering of Isogenic Cells Deficient for MR1 with a CRISPR/Cas9 Lentiviral System: Tools To Study Microbial Antigen Processing and Presentation to Human MR1-Restricted T Cells. *J Immunol* **197**, 971-982 (2016). https://doi.org/10.4049/jimmunol.1501402
217 Néron, B. *et al.* MacSyFinder v2: Improved modelling and search engine to identify molecular systems in genomes. *Peer Community Journal* **3** (2023). https://doi.org/10.24072/pcjournal.250
218 Chi, H. *et al.* Antiviral type III CRISPR signalling via conjugation of ATP and SAM. *Nature* **622**, 826-833 (2023). https://doi.org/10.1038/s41586-023-06620-5
219 Mukherjee, I. A., Gabel, C., Noinaj, N., Bondy-Denomy, J. & Chang, L. Structural basis of AcrIF24 as an anti-CRISPR protein and transcriptional suppressor. *Nature Chemical Biology* **18**, 1417-1424 (2022). https://doi.org/10.1038/s41589-022-01137-w
220 Picton, D. M. *et al.* A widespread family of WYL-domain transcriptional regulators co-localizes with diverse phage defence systems and islands. *Nucleic Acids Research* **50**, 5191-5207 (2022). https://doi.org/10.1093/nar/gkac334
221 Blankenchip, C. L. & Corbett, K. D. Bacterial WYL domain transcriptional repressors sense single-stranded DNA to control gene expression. *Nucleic Acids Research* **52**, 13723-13732 (2024). https://doi.org/10.1093/nar/gkae1101
222 Turnbull, K. J. & Gerdes, K. HicA toxin of scherichia coli derepresses hic transcription to selectively produce HicB antitoxin. *Molecular Microbiology* **104**, 781-792 (2017). https://doi.org/https://doi.org/10.1111/mmi.13662
223 EBI, E. B. I. *Ìndex of /pub/databases/Pfam/releases/Pfam35.0*, 2024).
224 Crawley, A. B., Henriksen, E. D., Stout, E., Brandt, K. & Barrangou, R. Characterizing the activity of abundant, diverse and active CRISPR-Cas systems in lactobacilli. *Scientific Reports* **8**, 11544 (2018). https://doi.org/10.1038/s41598-018-29746-3
225 Yoon, P. H. *et al.* Structure-guided discovery of ancestral CRISPR-Cas13 ribonucleases. *Science* **0**, eadq0553 https://doi.org/doi:10.1126/science.adq0553
226 Altae-Tran, H. *et al.* Diversity, evolution, and classification of the RNA-guided nucleases TnpB and Cas12. *Proceedings of the National Academy of Sciences* **120**, e2308224120 (2023). https://doi.org/10.1073/pnas.2308224120
227 Georjon, H. & Bernheim, A. The highly diverse antiphage defence systems of bacteria. *Nature Reviews Microbiology* **21**, 686-700 (2023). https://doi.org/10.1038/s41579-023-00934-x

228 Koonin Eugene, V., Krupovic, M. & Agol Vadim, I. The Baltimore Classification of Viruses 50 Years Later: How Does It Stand in the Light of Virus Evolution? *Microbiology and Molecular Biology Reviews* **85**, 10.1128/mmbr.00053-00021 (2021). https://doi.org/10.1128/mmbr.00053-21
229 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990). https://doi.org/10.1016/S0022-2836(05)80360-2
230 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009). https://doi.org/10.1186/1471-2105-10-421
231 Guo, J. *et al.* VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome* **9**, 37 (2021). https://doi.org/10.1186/s40168-020-00990-y
232 Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal* **27**, 379-423 (1948). https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
233 Zhu, W. *et al.* The CRISPR ancillary effector Can2 is a dual-specificity nuclease potentiating type III CRISPR defence. *Nucleic Acids Research* **49**, 2777-2789 (2021). https://doi.org/10.1093/nar/gkab073
234 Camargo, A. P. *et al.* Identification of mobile genetic elements with geNomad. *Nature Biotechnology* **42**, 1303-1312 (2024). https://doi.org/10.1038/s41587-023-01953-y
235 Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology* **36**, 996-1004 (2018). https://doi.org/10.1038/nbt.4229
236 Zhang, H., Dong, C., Li, L., Wasney, G. A. & Min, J. Structural insights into the modulatory role of the accessory protein WYL1 in the Type VI-D CRISPR-Cas system. *Nucleic Acids Research* **47**, 5420-5428 (2019). https://doi.org/10.1093/nar/gkz269
237 Margolis, S. R. & Meeske, A. J. Crosstalk between three CRISPR-Cas types enables primed type VI-A adaptation in <em>Listeria seeligeri</em&gt. *bioRxiv*, 2024.2010.2025.620265 (2024). https://doi.org/10.1101/2024.10.25.620265
238 Hilbert, B. J., Hayes, J. A., Stone, N. P., Xu, R.-G. & Kelch, B. A. The large terminase DNA packaging motor grips DNA with its ATPase domain for cleavage by the flexible nuclease domain. *Nucleic Acids Research* **45**, 3591-3605 (2017). https://doi.org/10.1093/nar/gkw1356
239 Zhang, L. *et al.* Structural and functional characterization of deep-sea thermophilic bacteriophage GVE2 HNH endonuclease. *Scientific Reports* **7**, 42542 (2017). https://doi.org/10.1038/srep42542
240 Kala, S. *et al.* HNH proteins are a widespread component of phage DNA packaging machines. *Proceedings of the National Academy of Sciences* **111**, 6022-6027 (2014). https://doi.org/10.1073/pnas.1320952111
241 Quiles-Puchalt, N. *et al.* Staphylococcal pathogenicity island DNA packaging system involving cos-site packaging and phage-encoded HNH endonucleases. *Proceedings of the National Academy of Sciences* **111**, 6016-6021 (2014). https://doi.org/10.1073/pnas.1320538111
242 Pinilla-Redondo, R. *et al.* CRISPR-Cas systems are widespread accessory elements across bacterial and archaeal plasmids. *Nucleic Acids Research* **50**, 4315-4328 (2022). https://doi.org/10.1093/nar/gkab859
243 Siedentop, B., Rüegg, D., Bonhoeffer, S. & Chabas, H. My host's enemy is my enemy: plasmids carrying CRISPR-Cas as a defence against phages. *Proceedings of the Royal Society B: Biological Sciences* **291**, 20232449 (2024). https://doi.org/10.1098/rspb.2023.2449

244 Stern, A., Keren, L., Wurtzel, O., Amitai, G. & Sorek, R. Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends Genet* **26**, 335-340 (2010). https://doi.org/10.1016/j.tig.2010.05.008
245 Ensminger, C. R. D. C. A. W. Priming in a permissive type I-C CRISPR–Cas system reveals distinct dynamics of spacer acquisition and loss. *RNA* **23**, 1525-1538 (2017). https://doi.org/10.1261/rna.062083.117
246 Nussenzweig, P. M., McGinn, J. & Marraffini, L. A. Cas9 Cleavage of Viral Genomes Primes the Acquisition of New Immunological Memories. *Cell Host & Microbe* **26**, 515-526.e516 (2019). https://doi.org/10.1016/j.chom.2019.09.002
247 Xiao, Y., Ng, S., Nam, K. H. & Ke, A. How type II CRISPR-Cas establish immunity through Cas1-Cas2-mediated spacer integration. *Nature* **550**, 137-141 (2017). https://doi.org/10.1038/nature24020
248 Karneyeva, K. *et al.* Interference Requirements of Type III CRISPR-Cas Systems from Thermus thermophilus. *Journal of Molecular Biology* **436**, 168448 (2024). https://doi.org/https://doi.org/10.1016/j.jmb.2024.168448
249 Wei, Y., Chesne, M. T., Terns, R. M. & Terns, M. P. Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in Streptococcus thermophilus. *Nucleic Acids Research* **43**, 1749-1758 (2015). https://doi.org/10.1093/nar/gku1407
250 Paez-Espino, D. *et al.* IMG/VR v.2.0: an integrated data management and analysis system for cultivated and environmental viral genomes. *Nucleic Acids Research* **47**, D678-D686 (2018). https://doi.org/10.1093/nar/gky1127
251 Camargo, A. P. *et al.* IMG/VR v4: an expanded database of uncultivated virus genomes within a framework of extensive functional, taxonomic, and ecological metadata. *Nucleic Acids Research* **51**, D733-D743 (2022). https://doi.org/10.1093/nar/gkac1037
252 Paez-Espino, D. *et al.* CRISPR Immunity Drives Rapid Phage Genome Evolution in Streptococcus thermophilus. *mBio* **6**, 10.1128/mbio.00262-00215 (2015). https://doi.org/10.1128/mbio.00262-15
253 Shiriaeva, A. A. *et al.* Detection of spacer precursors formed in vivo during primed CRISPR adaptation. *Nature Communications* **10**, 4603 (2019). https://doi.org/10.1038/s41467-019-12417-w
254 Shiriaeva, A. A. *et al.* Host nucleases generate prespacers for primed adaptation in the E. coli type I-E CRISPR-Cas system. *Science Advances* **8**, eabn8650 (2022). https://doi.org/doi:10.1126/sciadv.abn8650
255 Schmidt, F., Cherepkova, M. Y. & Platt, R. J. Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* **562**, 380-385 (2018). https://doi.org/10.1038/s41586-018-0569-1
256 Savitskaya, E., Semenova, E., Dedkov, V., Metlitskaya, A. & Severinov, K. High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in E. coli. *RNA Biology* **10**, 716-725 (2013). https://doi.org/10.4161/rna.24325
257 Lu, M. *et al.* Structure and genome editing of type I-B CRISPR-Cas. *Nature Communications* **15**, 4126 (2024). https://doi.org/10.1038/s41467-024-48598-2
258 Nimkar, S. & Anand, B. Cas3/I-C mediated target DNA recognition and cleavage during CRISPR interference are independent of the composition and architecture of Cascade surveillance complex. *Nucleic Acids Research* **48**, 2486-2501 (2020). https://doi.org/10.1093/nar/gkz1218
259 Hochstrasser, M. L., Taylor, D. W., Kornfeld, J. E., Nogales, E. & Doudna, J. A. DNA Targeting by a Minimal CRISPR RNA-Guided Cascade. *Molecular Cell* **63**, 840-851 (2016). https://doi.org/10.1016/j.molcel.2016.07.027
260 Lin, J. *et al.* DNA targeting by subtype I-D CRISPR–Cas shows type I and type III features. *Nucleic Acids Research* **48**, 10470-10478 (2020). https://doi.org/10.1093/nar/gkaa749

261 Rollie, C., Graham, S., Rouillon, C. & White, M. F. Prespacer processing and specific integration in a Type I-A CRISPR system. *Nucleic Acids Res* **46**, 1007-1020 (2018). https://doi.org/10.1093/nar/gkx1232

262 Hossain, A. A., McGinn, J., Meeske, A. J., Modell, J. W. & Marraffini, L. A. Viral recombination systems limit CRISPR-Cas targeting through the generation of escape mutations. *Cell Host & Microbe* **29**, 1482-1495.e1412 (2021). https://doi.org/https://doi.org/10.1016/j.chom.2021.09.001

263 Wu, X., Zhu, J., Tao, P. & Rao Venigalla, B. Bacteriophage T4 Escapes CRISPR Attack by Minihomology Recombination and Repair. *mBio* **12**, 10.1128/mbio.01361-01321 (2021). https://doi.org/10.1128/mbio.01361-21

264 Schelling, M. A., Nguyen, G. T. & Sashital, D. G. CRISPR-Cas effector specificity and cleavage site determine phage escape outcomes. *PLOS Biology* **21**, e3002065 (2023). https://doi.org/10.1371/journal.pbio.3002065

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

266 Letunic, I. & Bork, P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Research* **52**, W78-W82 (2024). https://doi.org/10.1093/nar/gkae268

267 Wang, J. Y. *et al*. Structural coordination between active sites of a CRISPR reverse transcriptase-integrase complex. *Nature Communications* **12**, 2571 (2021). https://doi.org/10.1038/s41467-021-22900-y