KU LEUVEN

FACULTY OF BIOSCIENCE ENGINEERING
FACULTY OF ENGINEERING SCIENCE
FACULTY OF MEDICINE
FACULTY OF SCIENCES

LARGE SCALE OMICS ANALYSIS OF THE BACTERIAL PAN-IMMUNE SYSTEM

Promoter:
Prof. Vera van Noort
Dr. Cédric Lood
Department Microbial and Plant Genetics (CMPG),
Leuven (Arenberg)
Division Departement Microbiële en Moleculaire
Systemen

Dissertation presented in
fulfillment of the requirements
for the degree of Master of Science:
Bioinformatics

Mianyong Ding

August 2022

Acknowledgements

First of all, I am grateful for the opportunity to do the thesis under the supervision of my promoter Prof. Vera van Noort and my co-promoter Dr. Cédric Lood. Thanks for welcoming me to the lab and giving me the opportunity to participate in the group meetings. I would especially like to thank my daily supervisor, Dr. Cédric Lood, for answering my questions and providing me with valuable experience as a bioinformatician.

Secondly, I would like to thank my family for supporting me during my study abroad. Without your encouragement, attention, and support, I would not be able to come this far. The language is not enough to express my gratitude.

Finally, I would like to thank all my friends who accompany me, help me and give me advice in the most difficult times. The good memories with you are also an important harvest during my master studies.

Abstract

The army race between phages and bacteria is unstoppable and has led bacteria to evolve various defenses mechanisms against phages. In recent years, the discovery of defense systems has been greatly accelerated, and a total of more than 60 systems have been discovered. The newly discovered defense systems expand the knowledge of pan-immune systems of bacteria.

One of the most well-known anti-phage defense systems is CRISPR-Cas, an adaptive immune system that defends bacterial cells against invading mobile genetic elements (MGE). CRISPR-Cas has been shown to potentially restrict HGT in *P. aeruginosa* and *P. aeruginosa* strains with CRISPR-Cas tend to have smaller genomes. The distribution of *P. aeruginosa* genome size shows a bimodal pattern, and the presence of CRISPR-Cas may drive the formation of this pattern.

In this work, we collect the complete genomes of the top 10 species with the most abundant genomes from the National Center for Biotechnology Information (NCBI) RefSeq and perform a large-scale genomic analysis to investigate whether the bimodal distribution of genome size is prevalent in other bacteria and whether CRISPR-Cas is associated with smaller genome size. We show that only *P. aeruginosa* has a distinct bimodal distribution of genome size. The association between a smaller genome and the presence of CRISPR-Cas exists only in *P. aeruginosa* and *L. monocytogenes*. Next, we investigate the impacts of anti-CRISPR proteins (Acrs) on genome size and find that the presence of anti-CRISPR proteins (Acrs) is associated with a larger genome in most species. We identify spacer targets and find that only *P. aeruginosa* has a high percentage of genomes (81.8%) with spacers targeting integrative conjugative elements (ICEs). This suggests that HGT is more restricted by CRISPR-Cas in *P. aeruginosa* than in other species, which may explain why *P. aeruginosa* is the only species with a distinct bimodal distribution of genome size.

To better understand the pan-immune system of each species, we next use the newly developed tool-PALDOC to detect defense systems of 10 species and provide a quantitative description of the antiviral arsenal of 10 species. We test whether several factors can influence the abundance and diversity of defense systems, such as genome size and the presence of CRISPR-Cas systems. We find that a larger genome is associated with more diverse and abundant defense systems in most species. The CRISPR-Cas is negatively correlated with some defense systems in some species. We investigate the spacers to see if the CRISPR-Cas can directly target other defense systems, the result shows that the RM systems are common targets for CRISPR-Cas systems.

Overall, our works can contribute to a comprehensive understanding of the bacterial pan-immune system.

List of abbreviations

Abi	Abortive infection
Aca	Anti-CRISPR-associated
Acr	Anti-CRISPR protein
Cas	CRISPR-associated (Cas) proteins
CDS	Coding DNA sequences
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DNA	Deoxyribonucleic acid
GBA	Guilt-By-Association
HGT	Horizontal Gene Transfer
HMM	Hidden Markov Model
HTH	Helix-turn-Helix
ICE	Integrative Conjugative Elements
MGEs	Mobile Genetic Elements
NCBI	National Center for Biotechnology Information
PADLOC	Prokaryotic Antiviral Defence LOCator
RBP	Receptor-Binding Protein
RM	Restriction Modification
RNA	Ribonucleic acid
RT	Reverse Transcriptase
SauCas9	Staphylococcus aureus Cas9
SIE	Superinfection Exclusion

List of tables

List of Figures

Table of Contents

1 Literature Review

1.1 Bacteriophages as the most abundant biological entities

Bacteriophages (or phages) were first discovered in 1915 by Fredrick Twoort and 1917 by Félix d'Hérelle independently (Salmond & Fineran, 2015). They are single-stranded or double stranded RNA or DNA viruses, which can infect and replicate within bacteria. Phages are the most abundant biological entities on Earth. They can be found in almost every biological niche, even in extreme environments, such as hot springs, deep-sea hydrothermal fields, Arctic Sea ice, and hypersaline lake (Gil et al., 2021). It was estimated the total number of phages on Earth is most likely close to 10^{31} (Mushegian, 2020). They also have high diversity in morphology, structure (tailed, non-tailed, enveloped, or filamentous), and genomes(Dion et al., 2020). The most discovered phages have a tailed morphology with dsDNA (*Caudovirales*). Their genome size varies from 2,435 bp (Leuconostoc phage L5) to over 650 kb (mega pages) (Dion et al., 2020), and the most giant identified phage is about 750kb (Al-Shayeb et al., 2020). From the National Center for Biotechnology Information (NCBI), there are more than 11000 complete phage genomes (last accessed on June 30, 2022).

1.1.1 The life cycle of phages

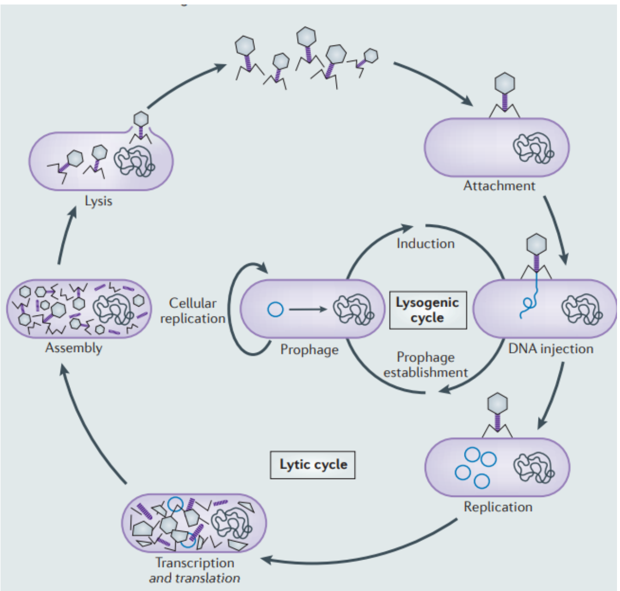


Figure 1.1 The life cycle of phages (Figure from (Salmond & Fineran, 2015)). The life cycle starts from phage attachment, and then phages inject the genomes inside the host. And the phage can go through two cycles, the lysogenetic and lytic cycles. The phage can replicate, transcribe, and assemble into new virions for the lytic cycle. The new virions are released through the lysis of the host. For the lysogenic cycle, phages integrate their genomes in the host chromosome as a prophage. Induced by the stressor, the prophage can exit the lysogenic cycle and enter the lytic cycle to replicate.

The phage infection generally starts from the specific recognition between the receptor-binding protein (RBP) and the receptors on the surface of bacteria (Salmond & Fineran, 2015). After the specific recognition, the phages can bind irreversibly to the cell wall of bacteria and initiate the injection of their genome into the host cell. After bypassing potential host intracellular defenses, the phage can go through two main strategies for its propagation (Figure 1.1) (Salmond & Fineran, 2015).

Lytic cycle: The genetic material of phages is replicated, transcribed, and assembled into new progeny particles. At the end of the lytic life cycle, the new progeny particles are released through lysis, leading to the host's death.

Lysogenic cycle: After entering the hosts, the phage genome can integrate into the bacterial host genomes with the help of the phage-encoded integrases or replicate as an episome (Howard-Varona et al., 2017; Ofir & Sorek, 2018). The phage is then referred to as a prophage in the bacterial genome, and the bacterial cell is called a lysogen. The prophage replicates as part of the bacterial replication. Without killing the bacteria host rapidly, the stable lysogenic life cycle can pass through many generations of cell division processes or horizontally transfer to other species (Howard-Varona et al., 2017).

Virulent viruses can only replicate their genomes through the lytic life cycle. By contrast, the temperate phages can either enter the lysogenic life cycle or produce new virions through the lytic cycle. The temperate phages can integrate their genome into the bacterial chromosome or maintain circular (P1) or linear (N15) extrachromosomal plasmid (Howard-Varona et al., 2017). The stressors (nutrients, pH, or temperature) can induce prophages to exit the lysogenic state and re-enter the lytic life cycle. The new virions are produced and released in the lytic life cycle through cell lysis (Salmond & Fineran, 2015). However, not all phages are released through cell lysis; some filamentous phages can be secreted across the host envelope without cell lysis through a productive chronic infection lifestyle (Loh et al., 2019).

Besides, phages may also have a psuedolysogenic component, which is defined (Clokie et al., 2011) as a carrier state in which the phages do not integrate into a stable fashion but wait for certain conditions to trigger them to enter into the lytic or lysogenic life cycle. Their extrachromosomal element can also be unevenly transferred to daughter cells through cell division (Clokie et al., 2011). This strategy is usually occurring in a nutrient-depleted environment and may also benefit from avoiding the damage from physicochemical conditions (such as UV light) (Chevallereau et al., 2022)

1.1.2 The host range of phages

Many phages show high specificity towards their hosts and can only infect a small set of strains within a bacterial species(Malki et al., 2015). Several discovered mechanisms may contribute to the difference. For the surface-adhesion step, the phages can interact with the receptors present on the bacterial surface via their receptor-binding proteins (RBPs). The RBPs expressed in the tips of the tail fibers or tail spikes strongly determine the host range and host specificity. Some phages can express multiple RBP genes simultaneously (Schwarzer et al., 2012) or express one RBP targeting the conserved structures between different hosts to extend their host range (de Jonge et al., 2019). After entering their hosts, mechanisms, such as deactivating defense systems of bacteria or preventing further infection of the same cell by other viruses (clonal or otherwise) via superinfection exclusion (SIE) (Bondy-Denomy et al., 2016). For the temperate phages, they may conduct integrase-mediated integration of their genome into the host chromosomal, which is specific to the sequences of targets. Targeting the conserved integration sites or maintaining them as a plasmid can potentially increase the host range(de Jonge et al., 2019).

The host range can also evolve. Some phages have high genetic diversity of RBP, which encode several RBPs in their genome. But they express only one RBP at one time. Switching the expression of RBPs can modify the host specificity of the host (de Jonge et al., 2019). The host specificity can also be modified genetically, e.g., the mutation of the receptor binding protein (RBP) gene (Perry et al., 2015; Yehl et al., 2019).

Intuitively, the phages with a broad host range can infect more bacteria and have high chances of surviving. However, there may be a tradeoff between the benefits from the broad host range and other costs. Extended host range is typically associated with lower virulence, slower growth rate, and lower thermostability (de Jonge et al., 2019). Environmental factors such as host density and diversity can influence the host range evolution (Chevallereau et al., 2022). For instance, *Pseudomonas* phage ϕ2 is only detected to infect resistant hosts when susceptible hosts are at frequencies between 0.1% and 1% (Benmayor et al., 2009). The phage can also evolve to be more specific if they grow in the environment with the abundance of one good quality host. For example, the Enterobacteria phage λ encoded two RBFs in their genomes (OmpF and LamB). However, grown in an environment with a mix of two receptors, they evolve to express either OmpF or LamB with their different host preferences (Meyer et al., 2016).

1.2 Host-phage interaction

The red queen hypothesis suggests that preys and predictors must keep evolving to maintain their relative fitness. The evolutionary arms race between phage and bacteria never stops. For example, the coevolution experiments find several rounds of coevolution between phage and bacteria when the *Escherichia coli* and lytic bacteriophage T3 are grown together for 30 days (Perry et al., 2015).For the first round, the bacteria develop resistance via the mutation that can influence the surface structure of bacteria for phage targeting, and the phages evolve new varieties (mutation in phage tail fiber genes modifying the host range), which can infect the resistant bacteria. Finally, bacteria develop resistance via mutation in multiple genes involved in phage interaction against the new varieties of phage receptors (Perry et al., 2015).

The host can also recruit a strategy to defend the phages via the competition of MGEs, e.g., the prophage can protect the host and prevent the superinfection of closely related phages (Koonin et al., 2020). The mechanism of superinfection exclusion(SIE) includes the expression of the transcriptional repressors (Koonin et al., 2020) that express the transcription of the competitors and cytoplasmic membrane modification to block phage binding and phage genome injection (Bondy-Denomy et al., 2016).

In the adsorption stage of phage infection, bacteria defend the phage through masking, hiding, mutating host surface receptors, or embedding in a biofilm matrix (Rostol & Marraffini, 2019) as the first front line. After the first front line is failed and the phage genomes are injected into the host cells, the bacteria encode diverse defense systems to defend the mobile genetic elements (MGEs) (see section 1.2.1). Recruiting diverse bacterial antiviral defense systems is a common strategy for bacteria to protect against phages.

The strategies from the first front line, such as cell surface modification, are constitutive (Westra et al., 2015), which means they are always active and independent of phage exposure. Other cellular defense systems, such as the CRISPR-Cas system, are inducible defense strategies and elicited upon phage infection (Westra et al., 2015). Resource environments and phage exposure can impact the tradeoff between constitutive versus inducible immunity (Westra et al., 2015). The constitutive strategy is associated with a fixed cost, which favors high resource environments and parasite-enriched conditions (Westra et al., 2015). The inducible defense strategy has few costs when the phage exposure is absent, but it can be costly with the phage exposure favors the low resource environments and low-parasite conditions(Westra et al., 2015).

1.2.1 The bacterial anti-phage defense system

Facing the diversity and abundance of viruses, the bacteria have evolved diverse antiviral systems. As of November 2021, more than 60 known antiviral defense families have been discovered (Tesson et al., 2022). And the discovery of novel defense systems has been accelerated in recent years, with more than 42 defense systems being discovered in 2022 (Millman et al., 2022; Vassallo et al., 2022). The anti-phage defense systems protect the host with diverse mechanisms, such as the CRISPR-Cas and RM systems can recognize and cleave the foreign nucleic acids. Abi protects the host through the suicide to prevent the spread of new virions (Bernheim & Sorek, 2020).

Defense systems are often discovered in defense islands, a hypothesis that anti-phage defense systems are frequently clustered in the bacterial genome (Makarova et al., 2011). The defense island concept enabled the discovery of new defense systems from the uncharacterized genes in these defense islands E.g., Doron *et al*. selected the candidate defense systems that co-localize with the known defense systems, and nine defense systems were validated to have anti-phage activities (Doron et al., 2018). Since 2018, more than 50 previously unknown defense systems have been discovered via the defense island hypothesis (Millman et al.,2022).

Searching for prophage encoded defense systems is also a strategy. Some defense systems are found within the prophages, which can participate in the inter-viral competition, showing a mutualistic relationship between the helper phage and host (Rousset et al., 2022). Rousset *et al*. discovered that the p2-like phage and p4-like phage satellites encode hotspots (small loci between two conserved genes, and it has a high turnover of genetic material) for anti-phage systems (Rousset et al., 2022). And one Abi-like defense system called PARIS system was validated for its anti-phage activity, which can be triggered by the anti-restriction protein and induce the growth arrest of the host (Rousset et al., 2022).

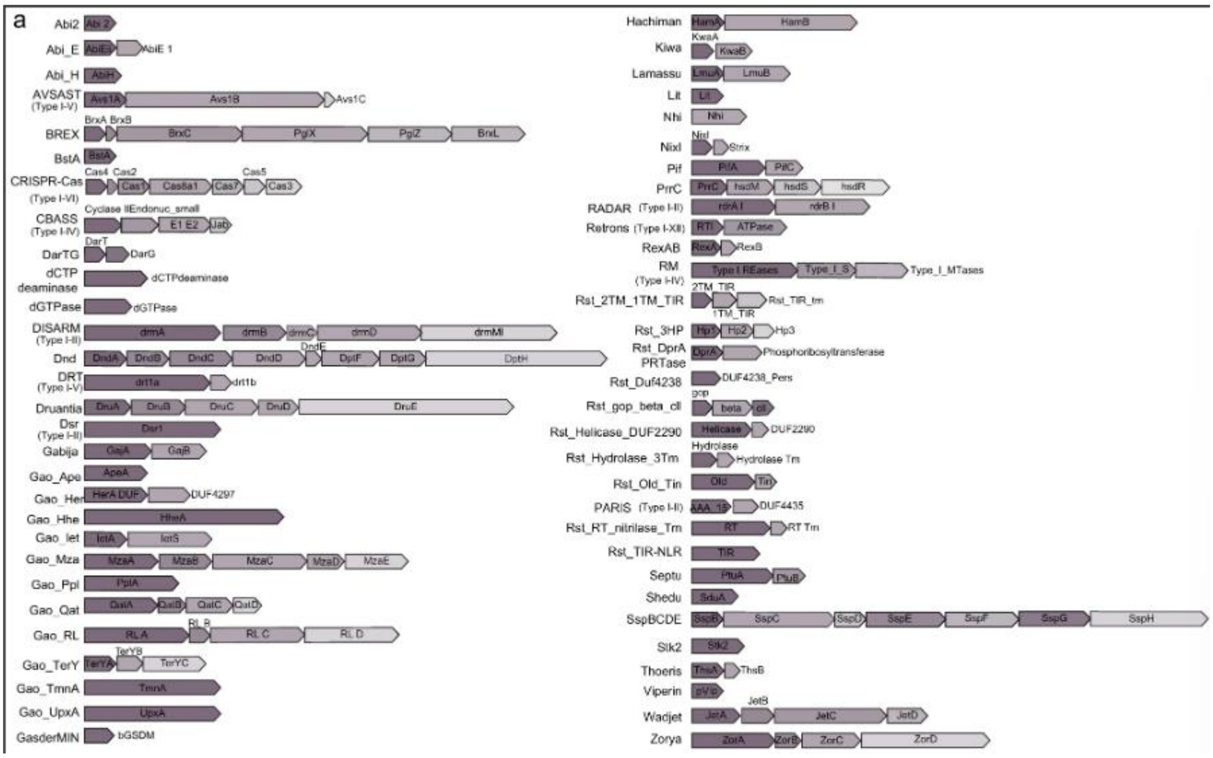


Figure 1.2 The defense systems discovered before November 2021 (the figure is from (Tesson et al., 2022). In total, 60 systems have been identified.

1.2.1.1 CRISPR-Cas system, the adaptive immune system.

The CRISPR (clustered, regularly, interspaced, short, palindromic repeats)–Cas (CRISPR-associated genes) is the defense system of bacteria and archaea to offer adaptive immunity against the invaded virus, external plasmids, or other MGEs (Newsom et al., 2020). The CRISPR-Cas system was first discovered in 1987 in *E. coli* and it is present in 85.2% of archaea and 40% of bacteria (Braz et al., 2020).

Currently, 33 subtypes of CRISPR-Cas systems have been identified and classified into two classes (class 1 and class 2) and six types (Type I, II, III, IV, V, VI) (Makarova et al., 2020). The class I CRISPR-Cas systems, which contain type I, type III, and type IV, are most abundant (Makarova et al., 2020).

The diversity of CRISPR-Cas reflects the various *cas* genes, gene composition, loci architecture, and mechanisms (Makarova et al., 2020). The main difference between class 1 and class 2 is that the class 1 systems recruit multi-proteins as effector modules, and class 2 systems use one single multi-domain protein as effector modules (Makarova et al., 2020). Type I and II can target DNA sequence, and RNA sequence can be targeted and cleaved by type III. Type VI can act as abortive infection ways (Makarova et al., 2020). Type IV CRISPR-Cas is a system without adaptation modules and effector nuclease, mostly encoded on plasmids (Pinilla-Redondo et al., 2022). A large comprehensive analysis of bacterial and archaeal genomes shows that the plasmid-encoding type CRISPR-Cas IV system has a bias target on other plasmids, suggesting the type IV system may be involved in the inter-plasmids competition(Kamruzzaman & Iredell, 2019; Pinilla-Redondo et al., 2022). The discovered frequent co-occurrence between plasmid-encoding type IV and other chromosomal CRISPR-Cas systems suggests that Type IV may utilize adaptation modules from other systems (Kamruzzaman & Iredell, 2019; Pinilla-Redondo et al., 2022).

The CRISPR-Cas system consists of short repeat partial palindromic arrays and unique spaces interspersed between them. And adjacent *cas* operon expresses the effective modules. Generally, the CRISPR array is adjacent to the *cas* operon to have immunity, but there are also some isolated CRISPR arrays (the CRISPR arrays are away from the *cas* operon). The recent bioinformatic analysis suggests three main routes for the evolution of isolated CRISPR arrays (Shmakov, Utkina, et al., 2020) (Figure 1.3):

(1) The CRISPR arrays are kept after the loss of *cas* genes from the complete CRISPR-Cas system.

(2) The CRISPR arrays are transferred by mobile genetic elements

(3) The CRISPR arrays originate from the synthesis of CRISPR array de novo, incorporating off-target spacers into repeat-like sequences.

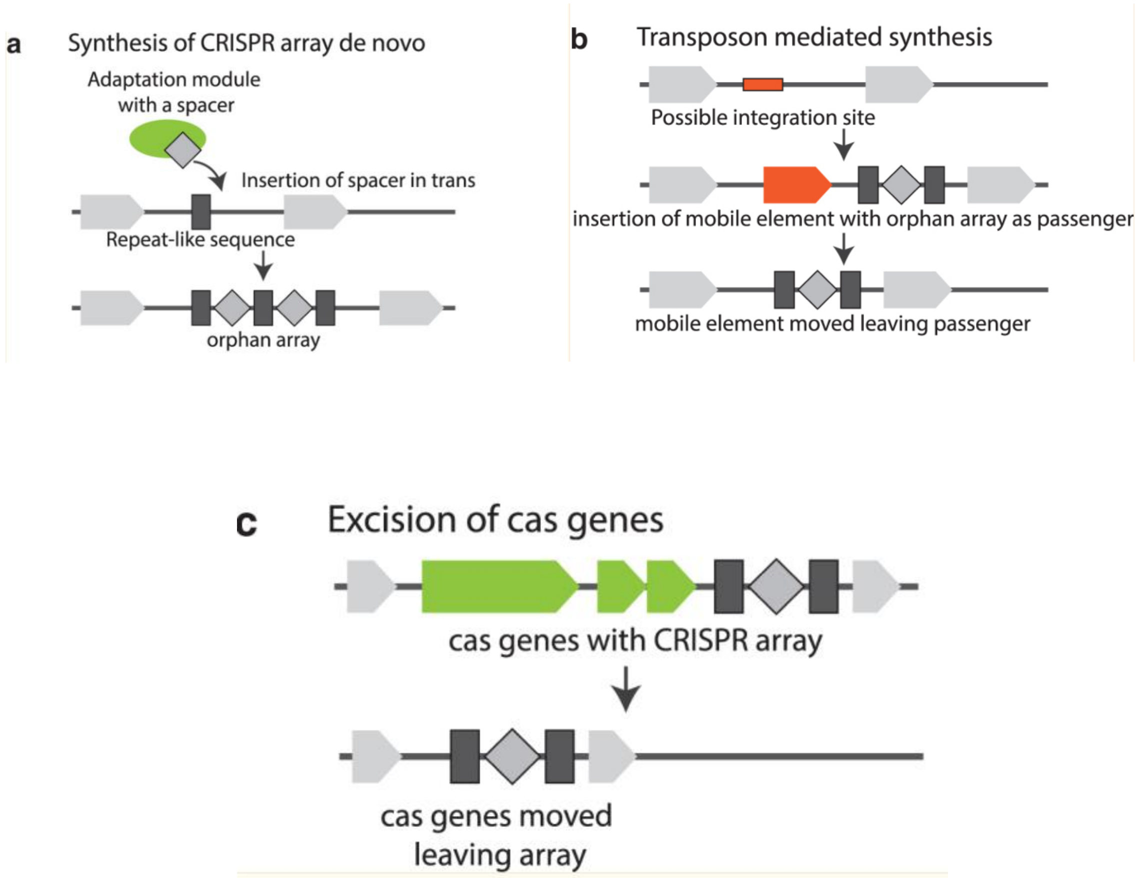

a
Synthesis of CRISPR array de novo
Adaptation module with a spacer
Insertion of spacer in trans
Repeat-like sequence
orphan array
b
Transposon mediated synthesis
Possible integration site
insertion of mobile element with orphan array as passenger
mobile element moved leaving passenger
c
Excision of cas genes
cas genes with CRISPR array
cas genes moved leaving array

Figure 1.3 The three routes for generating orphan CRISPR arrays (adapted from (Shmakov, Utkina, et al., 2020)). The dark gray rectangles are the repeat and diamonds are spacers, orange blocks are transposable elements and green blocks are *cas* genes. a: The de novo formation of CRISPR array by incorporating the off-target spacers from the adaptation module into repeat-like sequence. b: The isolated CRISPR array is formed by transposon insertion. c: The orphan array is formed by the loss of *cas* genes.

It is unclear if all the orphan arrays away from the *cas* genes are functionally active or not. Still, several pieces of evidence suggest that at least some orphan CRISPR arrays can be functionally active, such as some orphan CRISPR arrays that can still capture new spacers. Some of them can still contribute to the production of mature crRNAs (Shmakov, Utkina, et al., 2020). And the discovery of CRISPR mini arrays encoded in the phage genome also suggests the CRISPR array can contribute to the inter-virus competition (Medvedeva et al., 2019). The phage encodes the CRISPR arrays, which contain only 1-2 spacer homologous sequences of the related virus. And the phage can recruit the host Cas protein to inhibit the targeted virus (Medvedeva et al., 2019).

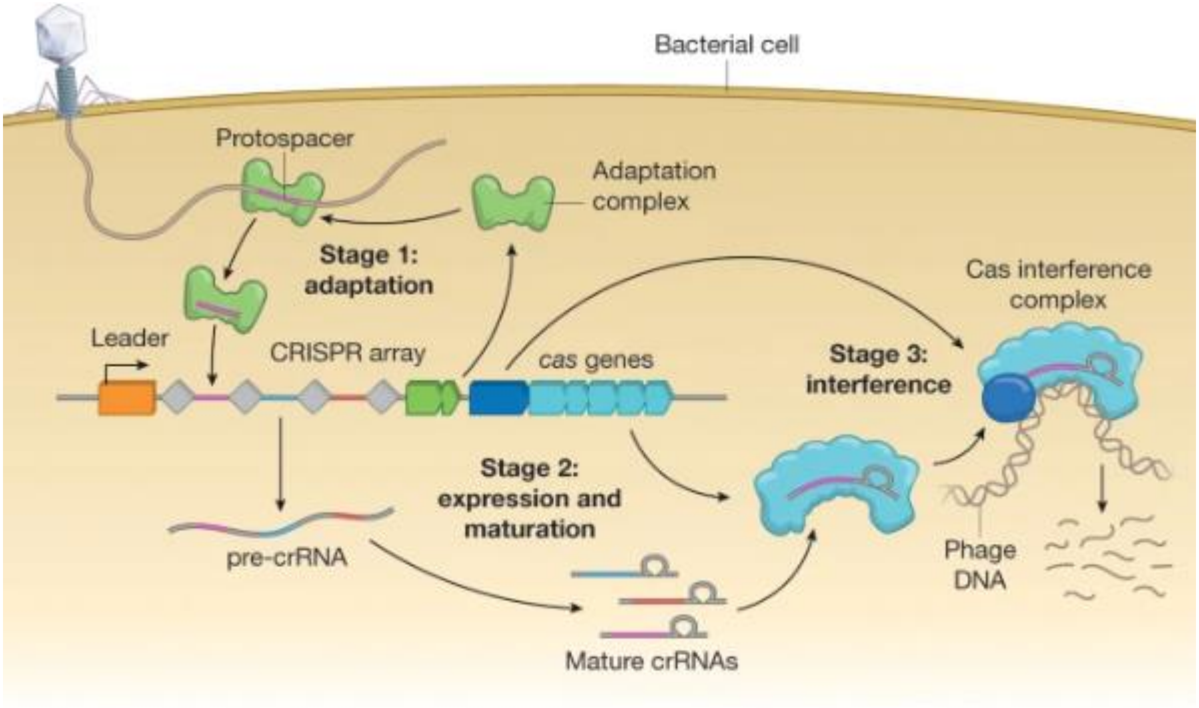


Figure 1.4 The three stages of CRISPR-Cas immunity. The figure is from(Hampton et al., 2020). The three stages of CRISPR-Cas immunity: adaptation, expression and maturation, and interference.

The classical CRISPR-Cas immunity comprises three stages: Adaptation, Expression, and Interference (Figure 1.4)(Hampton et al., 2020):

Adaptation: The foreign DNA is targeted by the Cas proteins complex (acquisition machinery), often after recognizing a short and distinct motif called protospacer-adjacent motif (PAM). The foreign DNA is cleaved into a protospacer, and the protospacer is inserted between CRISPR arrays as a spacer. Through the adaptation, the host can memorize the previously invaded phages and rapidly respond to the subsequent infections by the same phages.

Expression: The CRISPR array is transcribed into a single long precursor CRISPR RNA (pre-crRNA). Then, the pre-crRNA is processed by effective modules into a mature crRNA.

Interference: The crRNA and the Cas protein form the complex can recognize the nucleotide sequence, which is homologous to the spacer sequence. And the foreign nucleic acids are cleaved by Cas nuclease.

1.2.1.1.1 The uneven distribution of CRISPR-Cas systems

The CRISPR-Cas is not evenly distributed even within the same species. For instance, 0.8 % Staphylococcus aureus (Cruz-Lopez et al., 2021), 41.2% *Klebsiella. pneumoniae* (Li et al., 2018) and 50% *Pseudomonas aeruginosa*(Wheatley & MacLean, 2021) encode an active CRISPR-Cas system. The tradeoff between the benefits of antiviral defense and the cost of possessing CRISPR-Cas may explain the uneven distribution of CRISPR-Cas systems in the prokaryotic genomes. The costs include the maintenance of the Cas protein, the auto-immune response (Wimmer & Beisel, 2019), and the cost as the barrier for HGT(Zheng et al., 2020), resulting in frequent loss of CRISPR-Cas in bacteria.

Auto-immunity issue: The spacer sequence can match the phage genome, plasmid, archaeal virus, other organisms, and the host sequence (Wimmer & Beisel, 2019). With the self-targeting spacers, the CRISPR-based interference can target the host chromosomal and cut the sequence, potentially leading to the host cell death. The self-targeting spacer is estimated to exist in around 7% of genomes (Zhang et al., 2018). The cytotoxic impact caused by self-targeting CRISPR-Cas is termed the auto-immunity effect. However, some bacteria can still survive when the self-targeting spacers exist through several mechanisms. For instance, bacteria possess intrinsic DNA mechanisms that can repair the damage caused by CRISPR-based interference, such as homology-directed repair (HDR) (Wimmer & Beisel, 2019). The mutation of targeting sites, CRISPR array or *cas* genes ,and the presence of anti-CRISPR-Cas protein can also prevent the damage from the CRISPR-Cas (Wimmer & Beisel, 2019).

As a Barrier to HGT: The CRISPR-Cas have been proposed to limit the horizontal gene transfer (HGT) and prevent the acquisition of MGEs (Marraffini & Sontheimer, 2008). The role of CRISPR-Cas as the barrier to HGT also suggests it can prevent the acquisition of new beneficial exogenous genes to adapt to new environments. In *Bacillus cereus*, the CRISPR-Cas system is found as a barrier to HGT (Zheng et al., 2020). The strains of *B. cereus* with the active CRISPR-Cas system have restricted niche distribution and lower adaptation under extreme conditions compared to the strains without active CRISPR-Cas systems.

However, the influence of CRISPR-Cas on HGT is still controversial. The CRISPR-Cas have been found to inhibit the conjugation (Westra et al., 2013) and transformation (Bikard et al., 2012), which are two of three most important routes of HGT. Wheatley & MacLean found that more than 80% of *P. aeruginosa* have spacer targeting the integrative conjugative elements ,and the genome size is negatively associated with the presence of active CRISPR-Cas, suggesting the CRISPR-Cas systems can restrict the HGT (Wheatley & MacLean, 2021). The presence of anti-CRISPR proteins (Acrs), which can inhibit the activity of CRISPR-Cas is found to enhance the conjugation of the plasmids (Mahendra et al., 2020), also suggest that CRISPR-Cas can act as a barrier to HGT. However, as a third route of HGT, the phage-mediated transduction can be enhanced by CRISPR-Cas (Du Toit, 2018). Gophna *et al*. found no correlation between the number of HGT events and the CRISPR-Cas activity, suggesting the CRISPR-Cas does not influence the HGT over evolutionary timescales (Gophna et al., 2015).

The ecological and environmental factors can also influence the distribution of the CRISPR-Cas systems, such as the CRISPR-Cas are abundant in thermophiles and hyperthermophiles (Lan et al., 2022). Increased viral mutation rate and high genetic diversities of phages can lead the microbes to lose CRISPR-Cas systems as the bacteria only have few benefits from encoding the CRISPR-Cas with low chances of encountering phages (Weinberger et al., 2012). Weissman *et al.* (Weissman et al., 2019) used machine learning methods to determine the most important ecological factors shaping the abundance of CRISPR-Cas. The oxygen level and the temperature are found as the most critical factors (Weissman et al., 2019). The oxygen level is negatively correlated with the CRISPR incidence. The most likely explanation is that oxidative-stress-associated DNA (NHEJ DNA repair pathway), which is essential for the aerobes, is inhibited by some subtype of CRISPR-Cas (type II-A) (Weissman et al., 2019). The presence of CRISPR-Cas is positively associated with temperature. At low temperature, the bacteria may face the pressure from the grazing predictors (such as planktonic organisms) and viral lysis. The grazing risk decreases precipitously when the temperature increases to over 45, and the viral lysis becomes the main resource for cell mortality ,the abundance of CRISPR-Cas increases as an essential anti-phage strategy (Lan et al., 2022).

1.2.1.2 Restriction Modification systems are abundant in bacterial genomes

The restriction-modification system (RM) is a prevalent antiviral defense system and is estimated to be present in more than 74.2% of the prokaryotic genome (Oliveira et al., 2014). It is classified into four types and the type II RM system is the most studied system as it originally received the most interest in developing tools for recombinant DNA technology (Loenen et al., 2014).

The four types of RM systems differ in architecture, sequence recognition, cleavage sites and cofactor requirements (Oliveira et al., 2014). The classical type II RM systems include methyltransferase(M) and restriction endonuclease (R) as the main components. The RM system can discriminate the foreign and self-DNA sequence via recognizing the specific restriction-site sequence methylated by the methyltransferase. The foreign DNA sequence cannot be modified and cleaved by a restriction endonuclease. The type IV restriction system acts in the different way of Type II, as it contains modification-dependent enzymes to cut the modified DNA containing glycosylated bases, or methylated on adenine or cytosine residues(Loenen et al., 2014). Type I and type III contain RE and MTase activities but more complicated architecture compared to type II (Loenen et al., 2014).

1.2.1.3 Abortive Infection systems (Abi)

Apart from protecting the hosts at the unicellular level, the bacteria can also recruit an abortive infection system (Abi) to protect bacteria at the population level. Abi is not a single defense system, but a defense strategy employed by a variety of defense systems (Lopatina et al., 2020).

During an abortive infection, the infected bacterial host cells commit suicide to interrupt life cycle of infecting phages. This ensures that few or no new viral particles are released into the environment and prevents further spread of phages (Chopin et al., 2005). The sacrifice of infected cells can limit the spread of phages and benefit the whole community.

The Abi system can be triggered by multiple factors such as phage proteins, intermediates of phage genome replications, nucleic acids, the inhibition of other defense systems, or the mitigation of host gene expression upon phage infection (Bernheim & Sorek, 2020; Lopatina et al., 2020). The Abi systems contain at least two modules, one module is to sense the infection of phages and the other is to induce the death or metabolic arrest of the host cells (Lopatina et al., 2020). The Abi system has a high diversity in the mechanism of action, including mitigating the membrane permeability, inhibiting the protein synthesis, phosphorylating cell proteins and cleaving host RNA (Lopatina et al., 2020).

As the activation of the Abi system can eventually lead to the cell death of the host, it is supposed to be the last resort of defense and is only activated at the late stage of infection cycles if other defense systems (such as CRISPR-Cas system, RM system) cannot defend well in the earlier stage (Lopatina et al., 2020). However, it was recently shown that sometimes the Abi can cause reversible cell growth arrest to allow the other defense systems to have time to work (Lopatina et al., 2020).

1.2.1.3.1 Other Abi-like systems

Some newly discovered systems act as a specific form of Abi systems such as CBASS (Cyclic Oligonucleotide-based Antiphage Signaling Systems), Thoeris depletion (Ofir et al., 2021) and Retron system (Millman et al., 2020).

- In CBASS, the cGAS proteins detect phage DNA and produce the cyclic GMP–AMP (cGAMP) signaling molecule which activate the upstream effector (phospholipase), leading the cell death due to membrane degradation (Cohen et al., 2019).
- For the Thoeris defense system, phage infection is detected by the TIR-domain of ThsB protein, and the defense system produces a molecule signal (an isomer of cyclic ADP-ribose). The molecule signal can activate ThsA as the NADase effector, eventually leading to cell death due to NAD+ depletion (Ofir et al., 2021).
- As a guard-like system, the Retron system is composed of a reverse transcriptase (RT), a non-coding RNA (ncRNA) and effector proteins. The system can sense the phage-mediated inhibition of RecBCD, which acts a central role in DNA damage repairs and anti-phage activities. Then the Retron system utilizes the effector protein to induce the abortive infection and cell death (Millman, Bernheim et al. 2020). There is some connection between

the Retron system and the CBASS system as the Retron effector gene also contains some protein domains from CBASS (Millman, Bernheim et al. 2020).

1.2.1.4 Other anti-phage defense systems

Defense systems have diverse mechanisms, for example the system can use the chemical interference mechanism as an anti-phage strategy. The host can produce DNA-intercalating molecules that can target the invaded phages and block the replication of phages (Kronheim et al., 2018). Prokaryotic viperins (antiviral protein) can suppress the phage transcription by viral RNA polymerase (Bernheim et al., 2021). A newly discovered defense system that encodes the protein, such as cytidine or guanine deaminase proteins that can degrade the dCTP or dGTP, consequently blocking the replication due to the nucleotide depletion (Tal et.al.,2022). Two plasmid -elimination Ddm systems encoded on the pathogenic island were found. The Ddm systems can degrade the small plasmids and increase the burden of carrying long-conjugative plasmids and defend the phage by an Abi-like mechanism, which explains the absence of plasmids in the pandemic strains (Jaskolska et al., 2022). However, the mechanisms of a lot of defense systems discovered through the defense island's methods remain unknown (Doron et al., 2018; Gao et al., 2020).

Some components of defense systems that we have already discussed, such as the TIR domain in Thoeris , viperins, and cGAS signaling in CBASS, are also widely distributed in eukaryotic immune systems. The study of the bacterial antiviral defense system may also reveal the evolution of the innate immune system of eukaryotes. For example, a hypothesis that antiviral defense systems first evolved in prokaryotes and then were inherited and evolved to ancient eukaryotes (Cohen et al., 2019).

Some newly discovered defense systems contain protein domains that also involve in other cellular functions, which may indicate the emergence of new defense systems (Rocha & Bikard, 2022), such as previously mentioned viperins, which are homologs of GTP cyclases with other functions (Bernheim et al., 2021). Defense systems can also exist on MGEs and transferred to close strains via HGT. The HGT can facilitate protein co-optation so that proteins acquire new functions and form new assemblies through recombination and mutation, contributing to the emergence of novel defense systems (Rocha & Bikard, 2022) .

1.2.2 counter-defense systems in phages

As part of the arms race between bacteria and phages , phages also evolve diverse strategies to count bacterial antiviral systems (Bernheim & Sorek, 2020).

1.2.2.1 Anti-Crispr proteins (Acrs)

Phages appear to have evolved a strategy, known as anti-CRISPRs (Acrs) to inhibit the activity of CRISPR-Cas activity for their successful invasion. Acrs were first discovered in Pseudomonas phages and prophages in 2013 (Bondy-Denomy et al., 2013). Acrs exhibit high specificity toward the subtypes of CRISPR-Cas they can target, and they are named based on the subtype of CRISPR-Cas that they can inhibit (Vyas & Harish, 2022). To date, more than 40 distinct non-homologous Acrs have been discovered that inhibit fives of CRISPR-Cas (type I, type II, type III, type V and type VI) (Vyas & Harish, 2022).

The main inhibitory mechanisms of Acrs include the inhibition of DNA-binding and DNA-cleavage (Marino et al., 2020). In DNA-binding inhibition, Acrs can cleave crRNA, force Cas protein to dimerize, cause conformational changes in PAM sites ,or directly occupy the PAM recognition sites to interrupt the PAM recognition (Vyas & Harish, 2022). In inhibiting DNA cleavage, Acrs can bind to the Cas protein and block nuclease activity, thereby interrupting cleavage. Other strategies include disrupting the assembly of the CRISPR-Cas complex and degrading the second message formed by type III CRISPR-Cas (Vyas & Harish, 2022).

Acrs have low sequence and structural similarity (J. Wang et al., 2021), thus the detection of Acrs based on homology-search strategy is very limited. However, it has been discovered that Acr-encoding genes are often located near to the transcription suppressor genes that can encode a more conserved protein with helix-turn-helix (HTH) DNA binding domain (Yin et al., 2019). The guilt-by-association strategy was then developed to search for the Aca loci firstly and filter the neighborhood genes as the Acr candidates(Yin et al., 2019). However, bioinformatics analysis revealed that only a small percentage of Arc genes encoded by bacteria have Aca neighbors (23%) (Yin et al., 2019). The self-targeted CRISPR-Cas is supposed to be inactivated to ensure the survival of the host. Self-targeting strategy is also used to find the Acrs in the genomes with self-targeting CRISPR-Cas (Rauch et al., 2017). Based on these strategies, the bioinformatic tool was developed to discover the Acr loci from the DNA sequences (Yi et al., 2020).

1.2.2.2 Other counter-defense systems

In the RM systems, there is discrimination between the host DNA and foreign DNA through the unmethylated recognition sites (short DNA sequence). The recognition sites of host DNA are protected by methylation. The virus can recruit a strategy of eliminating recognition sites of their genomes to avoid the cleavage by the RM system (Rusinov et al., 2018). In the toxin-antitoxin systems, the host comprises the toxin proteins which can be neutralized by host-encoding cognate antitoxin protein. Upon the infection of phage, the toxin protein is triggered to block the phage development. It was found that phages can encode the inhibitors of toxins to overcome a defensive TA system (Otsuka & Yonesaki, 2012). For example, the T4 phage encodes *dmd* genes expressed as inhibitors of RnlA toxin to suppress the defensive activity of *E. coli* K12 TA system (RnlAB) (Otsuka & Yonesaki, 2012). The P4 phages can express two proteins (RIIA and RIIB) that can mediate the activity of the Rex system which belongs to Abi systems (Lopatina et al., 2020).

1.3 Pan-immune system model

As we described above, bacteria have a high diversity of bacterial defense systems, and many systems are probably still awaiting discovery. Because many defense systems can only target several specific invading phages, bacteria may recruit many defense systems together to provide multiple layers of protection against a wide range of MGEs (Bernheim & Sorek, 2020). For example, the co-existence of the CRISPR-Cas system and RM system can provide additive protection (Dupuis et al., 2013). The phage can evolve to count bacterial anti-viral defense systems such as Acrs. Protection by multiple defense systems can make the host more resistant to the mutated phages. For instance, The RM systems act as rapid and short -term defense which are overcome by phage by methylating viral genomes (Maguin et al., 2022). The methylated phages do not impair the spacer acquisition of the CRISPR-Cas system. The CRISPR-Cas systems which acquire the spacers from the methylated phages can offer robust protection against the methylated phages (Maguin et al., 2022).

However, there are trade-offs in recruiting multiple defense systems ,such as the cost of fitness (such as energy burden and autoimmunity) (Wimmer & Beisel, 2019), strong competitive disadvantages (Bernheim & Sorek, 2020), and incompatibility between the defense systems and other mechanism (NHEJ repair mechanism is inhibited by CRISPR-Cas II) (Bernheim et al., 2017). Besides, antagonistic effects between defense systems have been discovered, such as the RM system (PcaRCI) which can protect the host against the unmethylated MGEs. The PcaRCO has an epigenetic conflict with a methylation-dependent REase (PcaRCII), as the PcaRCII can target the methylation motif (Birkholz et al., 2022).

Some communities can gain immunity through their own variation. For example, CRISPR-Cas can enhance the immunity through acquisition of new spacers to target new enemies (Vassallo et al., 2022). The systems can be transferred via HGT, and indeed RM systems can be transferred between hosts via horizontal gene transfer(Rodic et al., 2017). The discovery of the abundance of defense systems in MGEs also suggests that the defense systems can be transferred across the closely related strains(Rocha & Bikard, 2022). And the MGEs can be gained at high rates and also lost frequently due to cost (Rocha & Bikard, 2022).

Bernheim and Sorek proposed the existence of a pan -immune system model, suggesting that the defense systems can be shared as the community sources and the individual strains can gain the immune defense mechanism from closely related strains through the HGT (Bernheim & Sorek, 2020).

In the pan-immune model, the related strains can encode diverse antiviral systems as accessory genes. The antiviral systems encoded by all related strains serve as community resources for all strains and the systems can be transferred between related strains via HGT(Bernheim & Sorek, 2020).

2 Method and Materials

2.1 Pipeline Overview

The pipeline was built with snakemake 7.8.5 (Molder et al., 2021). The pipeline starts with collecting data from NCBI- Refseq. After that, the CRISPR-Cas detection and other defense systems detection is performed on collected genomes. For the genomes with CRISPR-Cas system, the anti-CRISPR loci and spacer targets are also identified. Finally, all data are merged, analyzed and visualized.

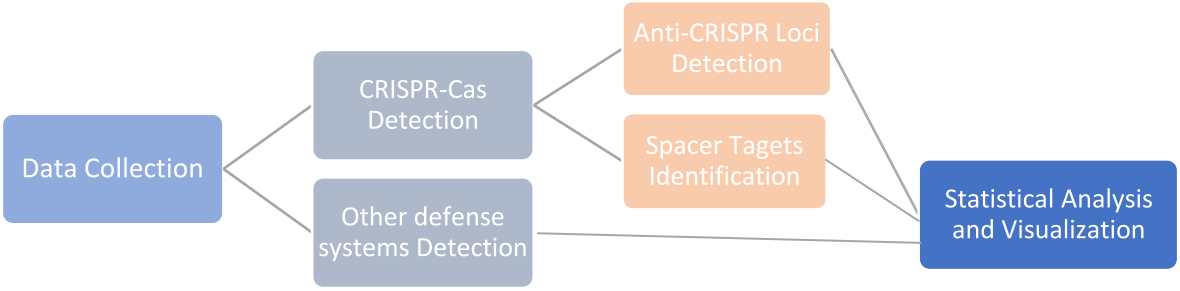

Data Collection
CRISPR-Cas Detection
Other defense systems Detection
Anti-CRISPR Loci Detection
Spacer Tagets Identification
Statistical Analysis and Visualization

Figure 2.1 The pipeline overview. The pipeline starts from data collection from NCBI RefSeq, then the CRISPR-Cas and other known defense systems of each collected genome are identified. For the genome encodes the CRISPR-Cas, the anti-CRISPR loci and targets of spacers are identified for the next step. And finally, all data are merged, analyzed, and visualized.

2.2 Data collection

2.2.1 Species selection

The list of all available bacterial genomes (N=27015) was collected from NCBI genome/prokaryotes (https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/ (last accessed Feb 14th). The top 10 bacterial species with the most abundant number of complete genomes were selected.

Only genomes with the complete assembly level were extracted. In the genomes with complete assembly level, the chromosomes are without gaps and there are no non-localized scaffolds between them. The issues of fragmentation can be avoided, and the plasmid sequences can also be provided.

The details of selected species can be seen in Table 3.1 and also Appendix 1. The FTP downloading links, genome size, and GC contents of their genomes were extracted. The outliers were removed if they had extremely high or low value under manual inspection. All genomes were downloaded from their FTP sites. The plasmids of each plasmid-equipped species were extracted into a new fasta file. To make the species trees of the selected 10 species, the Phylip tree file of taxonomy information was downloaded from NCBI taxonomy. The Phylip tree file was uploaded and visualized by NCBI Tree Viewer 1.19.2 with midpoint rooting.

2.2.2 Description of selected species

The top 10 species (table 2.1) with the most available genomes also represent the species that obtain the most research interest as they are strongly associated with human health and economic importance. Some strains were duplicated in the analysis and kept for subsequent analysis due to the small proportion of duplicated strains (Table 2.1), for example, only five *P. aeruginosa* strains are duplicated, each with two genomes.

Table 2.1. The total number of genomes and the number of strains of selected top 10 species. (From top to bottom: The largest number of complete genomes to smallest complete genomes).

Species	Total genomes	Strains number
Escherichia coli	1996	1943
Salmonella enterica	1170	1154
Klebsiella pneumoniae	1068	1064
Staphylococcus aureus	703	692
Bordetella pertussis	572	570
Pseudomonas aeruginosa	396	391
Acinetobacter baumannii	306	293
Mycobacterium tuberculosis	302	298
Listeria monocytogenes	268	266
Streptococcus pyogenes	251	249

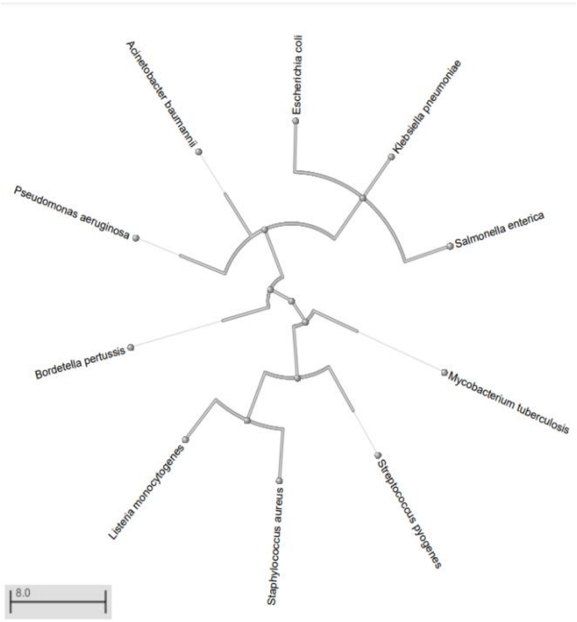


Figure 2.2 The species trees of selected 10 species with midpoint rooting.

Half of the selected species are the members of “ESKAPE” group pathogens (*Staphylococcus aureus*, *Klebsiella pneumonia*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Escherichia coli*.), the group of bacterial pathogens contributing to most nosocomial infections and associated with high capacity of development of antimicrobial resistance and high risk of mortality (Mulani et al., 2019; Rice, 2008). Over the past few decades, ESKAPE pathogens developed antibiotic resistance through natural selection of resistant strains or horizontal gene transfer, threatening global human health (Mulani et al., 2019; Rice, 2008).

Streptococcus pyogenes is a virulent pathogen responsible for a series of diseases ranging from mild, superficial infections to life-threatening diseases, contributing to 700 million cases of pharyngitis annually globally (Castro & Dorfmueller, 2021). *Bordetella pertussis* is an infectious agent of Pertussis (whooping cough,) which was estimated at 16 million cases and 195,000 deaths globally in 2008, and the cases increased considerably year by year (Kilgore et al., 2016). *Mycobacterium tuberculosis* is the etiological agent of human tuberculosis, and 10.4 million cases were reported in 2016 (Chai et al., 2018). *Listeria monocytogenes* are adaptable foodborne pathogens that can cause listeriosis. They are widely distributed as the contamination of food manufacturing and retail or food service setting environments, causing a heavy burden on the food industry and human health (Lopes-Luz et al., 2021). *Salmonella enterica* is the most common foodborne pathogen, causing zoonotic infection in animals and humans, contributing to thousands of deaths and substantial economic losses (Jajere, 2019).

The species tree (figure 2.2) of these ten species illustrates the phylogenetic relationship of these ten species. Some species have close distances, for instance, *E. coli*, *K. pneumoniae*, and *S. enterica*, three of them belong to the Enterobacteriaceae family.

2.2.3 Potential bias of extracted data

The extracted genomes only represent part of the population. From the NCBI genome, the 250687 bacteria individuals were sequenced (last access 23rd Feb 2022), and 27015 individuals with complete assembly level were available. 7030 individuals (2.8%) were used for this analysis to discover the common patterns of the bacteria kingdom. For *E. coli*, we obtained 2000 genomes with complete assembly levels, and the total number of genomes of *E. coli* is 27498 (13 times larger than the number we selected).

In the following sections, we want to discover the common pattern in the bacteria kingdom using the selected ten bacteria species. The bias of data availability should be realized since most of them are related to human activities and represent only the tip of the iceberg of the bacteria kingdom. And in a species, the strains that are most likely to be studied are those that are important to humans, for example, the pathogenic strains or the strains that are found in human hosts.

2.3 The detection of CRISPR-Cas

2.3.1 The selection of CRISPR-Cas detection tools

The identification of the CRISPR-Cas loci from genomic sequence includes the identification of CRISPR arrays and *cas* operons. The available tools for CRISPR arrays detection includes the CRISPRDetect (Biswas et al., 2016), CRISPRClassify (Nethery et al., 2021), and CRISPRIdentify (Mitrofanov et al., 2021). And *cas* operons can be detected using MacSyFinder (Abby et al., 2014). The available tools to identify the complete CRISPR-Cas system and identify subtypes include CRISPRCasFinder (Couvin et al., 2018) and CRISPRCasTyper (Russel et al., 2020).

CRISPRCasFinder

CRISPRCasFinder is the most cited tool for CRISPR-Cas loci detection(Couvin et al., 2018). It uses CRISPRFinder v4.2 to detect CRISPR repeats (Couvin et al., 2018). The detected CRISPR arrays can be classified into four confidence levels, from low-confidence level one to high-confidence level four, based on several criteria such as spacers number, conservation, and overall percentage identity of spacers. After using Prodigal v2.6.3 (Hyatt et al., 2010) to predict coding sequences (CDSs), Cas proteins and types are detected based on the similarity using HMM protein files using CasFinder based on MacSyFinder (Abby et al., 2014).

CRISPRCasTyper

CRISPRCasTyper identifies and annotates CRISPR and *cas* loci based on the latest classification schemes (Russel et al., 2020). It can identify 44 subtypes/variants and 31 subtypes with a median accuracy of 98.6%. CRISPRCasTyper detects the *cas* loci and other related genes using HMMERE3 v3.2.1 (Eddy, 2009) based on HHMs profiles merged from multiple sources. Then, the type of *cas* operon is classified based on their scoring scheme. The machine learning method (XGboost) is used to detect the CRISPR array based on a combination of tetramers of arrays.

Compared to CRISPRCasFinder, it has higher or equal accuracy of 19 subtypes, equal false positives (0.4%) and detection of additional 12 subtypes missing in CRISPR-Cas.

2.3.2 CRISPR-Cas loci identification using CRISPRCasTyper

The CRISPRCasTyper (v1.6.1) was selected to detect CRISPR Cas systems of all extracted genomes because it can detect more CRISPR-Cas subtypes. CRISPRCasTyper is more convenient to install and use. The outputs of CRISPRCasTyper are easier to parse.

The output of CRISPRCasTyper contains information, such as the position, subtypes, and distance between the CRISPR array and the *cas* operon of CRISPR-Cas systems. The details of all discovered CRISPR arrays and *cas* operons are also available. The information also includes the orphan CRISPR arrays (the CRISPR arrays that are not associated with *cas* operon) and isolated *cas* operons (no CRISPR arrays nearby). The spacers from all CRISPR arrays are also generated as .fa files. Outputs of CRISPRCasTyper also list some low-quality results which is marked as *._putative.tab. To make the detection more rigorous, the low-quality detections were excluded from our analysis.

For each genome, we extracted the presence of CRISPR-Cas systems, the subtypes of CRISPR-Cas systems, the presence of orphan CRISPRs, the number of spacers of orphan CRISPR arrays, and the number of spacers of *cas* operon adjacent CRISPR arrays.

Detection of CRISPR-Cas on plasmid sequences was also performed to investigate the abundance of CRISPR-Cas with CRISPRCasTyper.

2.4 The detection of Anti-CRISPR loci

2.4.1 Comparison of Acrs-detection tool

After a literature search, several bioinformatics tools for detecting Acrs were selected and compared (Table 2.2). The main approaches for detecting Acrs include homology-based method, machine learning-based methods, and guilt by association (GBA) method. For the homology-based method, the query protein is compared with already known Acrs, e.g., the Blast and HMM-based methods. However, the newly discovered Acrs have low sequence similarity between Acrs, indicating the homology-based is not a perfect method (Wang et al., 2020). The machine learning-based method, e.g., AcRanker and PaCRISPR, uses intrinsic patterns of protein sequence and have a higher capacity to capture the Acrs.

Table 2.2 List of current bioinformatics tools for detecting anti-CRISPR.

Tool	Description	Last updated	Reference
AcrFinder	It is a workflow combining homology-based detection and guilt-by association (GBA) detection methods and the presence of self-targeting spacers. Online server and standalone program are both available.	May, 2020	(Yi et al., 2020)
ArcHub	It is an online hub provided to investigate, predict, and map the Acrs. It integrates three predictors such as HMM based predictor, AcRanker and PaCRISPR.	Jan, 2021	(J. Wang et al., 2021)
AcrCatalog	A database of Acrs predicted by random forest model, which uses eight sequence features, including self-targeting and direction. The python script for the prediction is also available on the GitHub repository.	June, 2020	(Gussow et al., 2020)
AcRanker	Gradient boosting based model to rank the phage proteomes for potential Acrs. The python script is available.	Dec, 2020	(Eitzinger et al., 2020)

PaCRISPR	An online server to predict the Acrs using the machine learning based method (Support Vector Machine) based on features such as position-specific scoring matrix (PSSM).	April, 2020	(Wang et al., 2020)

2.4.1.1 The detection based on homology and GBA routes

AcrFinder was developed both as a standalone software and web server to discover the Acrs via two routes: the homology-based and the guilt-by-associated methods (Yi et al., 2020). For the homology-based search, DIAMOND ((Buchfink et al., 2015) is used based on 56 experimentally validated Acrs. Using DIAMOND, Acrs are used as query protein to blast against the genome to find Acrs homology.

For the guilt by association route, the Acrs database is used to build the conserved Acr-associated (Aca) regulator database via Acrs homologous gene neighborhood. The Aca are genes that encode the helix-turn-helix (HTH) domain and sit adjacent to the Acrs (Yi et al., 2020). The Aca homology is detected by DIAMOND. And then, the CRISPRCasFinder is used to evaluate the presence of complete CRISPR-Cas systems and to find the self-targeting spacers. By integrating the results of Aca and self-targeted spacers, three confidence levels are generated for the inferred CRISPR-Cas subtypes. At the high confidence level, the Aca is near (within 5000 bp) to a self-targeted protospacer. At the medium confidence level, the Aca loci has a self-targeted protospacer, but the spacers are not nearby (5000 bp away). At the low confidence level, the Aca loci is only detected alone without a self-targeted protospacer.

2.4.1.2 Machine learning based method

The machine learning-based method can predict some novel Acrs candidates, which can be further experimentally validated (Eitzinger et al., 2020; Wang et al., 2020). PaCRISPR makes predictions based on the position-specific scoring matrix (PSSM), and AcRanker uses the same features such as amino acid composition and the frequency of dimer and trimer to make the prediction based on extreme gradient boosting (XGboost) (Eitzinger et al., 2020; Wang et al., 2020). PaCRISPR is a time-consuming method, as it is costly to calculate the PSSM. AcrHub is the latest method that combines three predictors (AcRanker, PaCRISPR, and HMM-based model) and makes predictions based on the most updated Acrs database (J. Wang et al., 2021).

Although several strengths of the machine learning-based method were demonstrated, AcrFinder was ultimately used. The main reason is that AcrHub and PaCRISPR only provide online servers without local standalone applications or API. The detection of Acrs was performed on more than 7000 genomes, which is very time-consuming for the large-scale analysis with the online server. AcRanker and AcrCatalog do make the script available on GitHub as well. However, they cannot work because they are not maintained. And AcrFinder can be supported by Docker, without any issues of updated dependencies.

Some machine learning methods only predict the presence of Acrs, but not the subtypes of Acrs. For example, AcRanker does not provide any information about the subtypes of Acrs, but the Acrs are highly specific to CRISPR-Cas (Eitzinger et al., 2020). The subtypes of Acrs are available from AcrFinder. And AcrFinder does not lose too much compared to some machine learning methods such as AcRanker. The performance of AcrFinder is close to that of AcRanker (Yi et al., 2020).

The main limitation of AcrFinder is that AcrFinder can only identify Acrs via the GBA route when there is at least one complete CRISPR-Cas system present on the same genomes. However, in this analysis, we detect Acrs to evaluate the activities of detected CRISPR-Cas. This limitation can be ignored in this analysis. Another limitation of AcrFinder is the lack of updates for the Acrs database of AcrFinder. The database was last updated in May 2020, and contains 64 experimentally validated Acrs (Yi et al., 2020). But for AcrHub, which was released in January 2021, contains 339 nonredundant experimentally validated Acrs. New families of Acrs continue to be discovered, but their database has not been updated since January 2021 (J. Wang et al., 2021).

The comparison between the different tools suggests there is a need for a self-updated standalone software for the large-scale analysis.

2.4.2 The detection of anti-CRISPR loci from AcrFinder

The output of AcrFinder contains two output files representing the results of two routes, the homology-based route and the GBA route. For the outputs of two routes, the Acrs and subtypes of Acrs were extracted. For the output of the GBA route, the inferred acrs and confidence levels were extracted. To determine whether CRISPR-Cas is inhibited or not, the subtypes of Acrs encoded on the same genomes were screened to see whether they can target the CRISPR-Cas.

The Acrs from the GBA contain three confidence levels: low confidence level, medium confidence level, and high confidence level. The low confidence level means that the Acr-Aca loci has been detected but no self-targeted spacers are present. The self-targeted spacers are detected at the medium and high confidence levels, but at the high confidence level, the Acr-Aca loci is near (within 5000bp) self-targeted spacers. At the medium confidence level, the Acr-Aca loci has no self-targeted spacers nearby. The CRISPR-Cas system is thought to be inactivated by the presence of self-targeted spacers. Since the Acrs were detected to reassess the activity of the encoded CRISPR-Cas encoded, the inferred Acrs from GBA with medium and high confidence levels were used.

We realized some genomes might encode two CRISPR-Cas systems, the complete inhibition or partial inhibition of genomes encoding two CRISPR-Cas were also determined manually. Some Acrs with dual inhibition effects were collected from different sources (Table 2.3) (Yi et al., 2020; Yu & Marchisio, 2020). The inhibition effects of Acrs were often invalidated against the preselected types of CRISPR-Cas and the dual inhibition of other Acrs is unknown.

Table 2.3 The list of Acrs can detect at least two Types of CRISPR-Cas ,the data are collected from two sources (Yi et al., 2020; Yu & Marchisio, 2020).

Inhibited type	Acr type
TypeIIA + TypeIB	AcrIIA1
TypeIIA + TypeIB	AcrIIA2
TypeIIA + TypeIB	AcrIIA3
TypeIIA + TypeIB	AcrIIA4
TypeIF + TypeIE	AcrIE4-F7
TypeIB + TypeIIA	AcrIIA15

2.5 Detection of defense systems loci

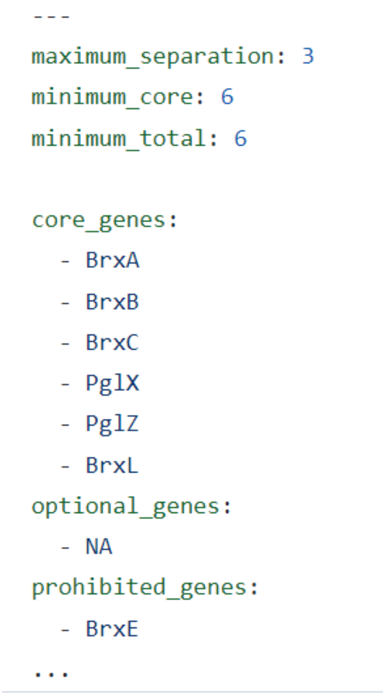
To detect the locus of all available defense systems, there are two tools available both as a standalone package and as a web server: DefenseFinder (Tesson et al., 2022) and PADLOC (Payne et al., 2022; Payne et al., 2021).

2.5.1 Comparison of available tools

2.5.1.1 PADLOC (Prokaryotic Antiviral Defense LOator)

The first version of PADLOC (Prokaryotic Antiviral Defense LOator) was released in November 2021, which can detect only about 11 defense systems. The latest version for the standalone package was released in February 2022 and can detect about 58 defense systems with 206 subtypes (last check on July 31, 2022). The new web server with expanded database and functionality was released in July 2022 (Payne et al., 2022).

The PALODC database was built by collecting profile HMMs representing different defense system proteins from different sources, such as other published papers and databases. When the profile HMMs of some defense system proteins are not available, they are generated using HMMER3 (Eddy, 2011). These profile HMMs are applied to identify the defense system proteins using HMMER3 (Eddy, 2011). After the defense genes are identified, the decision on the existence of a complete defense system is made based on the criteria specified in the YAML file. The defense system genes are designated as core genes, optional genes, or prohibited genes. In the classification process, the core genes are strictly required, and the prohibited proteins are strictly forbidden. To successfully classify one system, criteria such as the minimum number of genes should be met. On average, the detection accuracy is 97% for multi-gene systems and 89% for single-gene systems.



```
---
maximum_separation: 3
minimum_core: 6
minimum_total: 6

core_genes:
  - BrxA
  - BrxB
  - BrxC
  - PglX
  - PglZ
  - BrxL
optional_genes:
  - NA
prohibited_genes:
  - BrxE
...
```

Figure 2.3 The classification criteria for the Type I Brex system described in the brex_type_I.yaml file(Payne et al., 2021). The classification criteria include the core genes, optional_genes and prohibited_genes. The maximum separation of defense genes from Type I Brex is 3 genes, the minimum core genes are 6 genes and the minimul_total genes are 6 genes.

An example of a YAML file is brex_type_I.yaml. To meet the criteria for the type I Brex system, the number of core genes should be at least 6, the BrxE gene is strictly prohibited, and the maximum number of unrelated genes allowed should be less than 3.

The nucleotide FASTA file can be provided as input, and Prodigal (Hyatt et al., 2010) is used to predict the open reading frame. The thresholds for E-value and alignment coverage are different for multi-gene defense systems and single-gene systems. For multi-gene systems, an E value of 1×10^{-5} and an alignment coverage of 30% are required. And the 1×10^{-25} E-value and 50% alignment coverage are required for a single gene system.

A limitation of PADLOC is that it can only detect the coding genes and classify based on the coding genes. The non-coding sequences such as CRISPR arrays and retron-associated ncRNAs cannot be detected and used for classification. However, in the latest version of the web server, a new functionality has been added and the limitation of non-coding sequences has been solved (Payne et al., 2022).

The DMS family defined by the author includes the [system]_other subtypes of some defense systems such as disarm, Dpd, BREX, GAO 29, phophorothioation, and RM systems. The reason for the construction of this family is that several defense systems are very similar to non-defense molecular systems. The DMS family allows for relaxed system classification.

2.5.1.2 DefenseFinder

It was first released in Sep 2021 as a preprint and officially published in May 2022 (Tesson et al., 2022), which can detect 60 antiviral families and 151 subtypes (Figure 1.2).

DefenseFinder uses MacSyFinder (Abby et al., 2014) to detect defense systems in steps that are very close to PADLOC. The HMM profiles of DefenseFinder are collected from an exhaustive search of established HMM profiles of known defense systems. The HMM profiles are built by HMMER3(Eddy, 2011) when they are not available. The customized rules are made based on the genetic architecture of the defense systems. After identifying defense proteins using HMMs, rules are used to determine whether the complete defense systems are present. The decision rules also include the mandatory, accessory, and forbidden defense proteins. It can achieve high specificity (97.4% to 99.4%) and high sensitivity (96.7% to 99.97%) for the defense systems that can be invalidated (Tesson et al., 2022)

Comparing the PADLOC and DefenseFinder databases, DefenseFinder is able to detect many more antiviral families (60 versus 11) than the initial version of PADLOC. However, PADLOC constantly updates the database, the number of detected families is close (DefenseFinder (v.1.0.0) contains 60 defense systems, and PADLOC-DB (v.1.4.1) contains 58 families). PADLOC has more HMM profiles (3824 for PADLOC and 845 for DefenseFinder) and more detected subtypes (206 vs 151).

The performance of both tools can only be evaluated if the ground truth of defense systems is available. Unfortunately, most of them are not available. When comparing 11 defense systems that can be detected both by PADLOC and DefenseFinder using 18 683 genomes, the detection result is close for most genomes (Tesson et al., 2022).

Although the PADLOC method is slightly slower, it was finally chosen because it contains more HMMs profiles and detects more subtypes compared to DefenseFinder. The performance of the two tools is close, and PADLOC is much easier to implement (the nucleotide sequences can be used as input). PADLOC is more informative, as the loci of the defense system genes, the strand and E value, and the hmm coverage of each hit can be specified only in PADLOC. PADLOC can also generate the .gff file with the annotations of defense genes, which can be visualized by other software.

2.5.2 Identification of defense system loci with PADLOC

The standalone package PADLOC (V.1.1.0) and its database PADLOC -DB (1.4.0) were finally used. The sys_meta.txt downloaded from the PADLOC-DB GitHub repository contains the information about the subtypes, ID of the defense systems. Currently PADLOC can detect 58 systems and 206 subtypes.

For each genome, the number of each system and the start and end positions of each system were collected. And the subtypes of the defense systems belonging to the same defense family were counted together. The identification of the antiviral defense systems on the plasmids of each genome was also performed. Due to the requirements of Prodigal (Hyatt et al., 2010). PADLOC cannot be used for FASTA files with < 100 kbp, so the defense systems cannot be identified in the plasmid less than 100 kbp cannot be identified.

2.6 The spacers target identification

2.6.1 Spacer extraction

The spacers of the individual CRISPR-Cas encoded genomes were extracted from the output of CRISPRCasTyper. The spacers of the individual genomes were merged into a fa file. Since we do not know whether the isolated CRISPR array away from *cas* genes can function, the spacers of the CRISPR arrays and the spacers of the CRISPR-near *cas* operons were extracted separately.

After checking the output of CRISPRCasTyper, we found that the detected spacers also included the spacers of the low-quality CRISPR (most of them are false positives). To make the results more stringent, the low-quality CRISPRs were removed. In addition, the results of spacers from orphan CRISPRs and spacers from CRISPRs near *cas* operons were extracted and analyzed separately.

2.6.2 The establishment of the BLAST database

To find the targets of the spacers, the three types of candidates (the integrative and conjugative elements (ICEs), the viral genome, and the plasmid) were selected. The ICEs were downloaded from the ICEeberg database (version 2.0) (Liu et al., 2019). The viral genomes were downloaded from NCBI CBI (ftp://ftp.ncbi.nih.gov/refseq/release/viral/) (last accessed in May), and the plasmid sequences were downloaded from PLSDB (Schmartz et al., 2022). The database was processed, and the 14743 viral genomes and 34512 complete genomes and 552 ICE sequences were eventually selected. For the ICE sequences, the ID was converted to a shorter ID that met the blast requirements of the blast. The three blast databases were created based on these processed nucleotide sequences.

To investigate whether CRISPR-Cas can directly target the defense systems, we first selected candidate defense systems based on the correlation plot of CRISPR-Cas and other defense systems. The defense systems that have a correlation of more than 0.2 with CRISPR-Cas were selected. As the PADLOC (Payne et al., 2021) and DefenseFinder (Tesson et al., 2022) collected or created the HMMs profiles for each defense system and the original sequences of defense genes were not provided. The sequence of defense systems RM, Abi, KIWA, LAASSU, SEPTU, SHEDU, THOERIS, ZORYA, DND from PADS Arsenal (Zhang et al., 2020). And the sequences of retron, qatABCD, ietAs, and ppl were also collected (Gao et al., 2020).

2.6.3 BLAST to find the best hits

The blastn with blastn -short task (optimized for sequences with less than 30 nucleotides) was performed and the spacer sequence (from both the orphan arrays and *cas* operon adjacent to the CRISPR arrays) as the query sequence was blasted to each database. The threshold used for blast was 95% sequence identity and 95% sequence coverage should be achieved. For the spacer with multiple qualified hits, the hit with the lowest e value was selected.

2.7 Statistical analysis and visualization

Data from all sections were merged and analyzed. Data analysis and visualization was mainly performed using Rstudio (R version 3.6.3 (2020-02-29)). The R package ggplot2 was used to generate all plots.

Most significance tests (e.g., relationship between the presence of CRISPR-Cas and genome size) were performed using the Wilcoxon rank sum test (with p=0.05 as the cutoff value) after checking for non-normality using the Shapiro-Wilk test. Some correlation tests (e.g., the correlation between the number of CRISPR-Cas and the number of other defense systems) use the Kendall rank correlation test, a nonparametric correlation test for capturing the non-linear relationship. The method of the correlation test is indicated in the results of each analysis.

The genome file and .gff file from the PADLOC output were collected and imported into Geneious (2022.2) to visualize the loci of detected defense genes.

The code and data for the analysis can be available through GitHub (https://github.com/Raminmian/masterthesis).

3 Results and discussion

3.1 The distribution of genome size of all species

The bimodal distribution of genome size in *P. aeruginosa* has been observed previously and the bimodal distribution is linked to the presence of CRISPR-Cas system (Wheatley & MacLean, 2021). We plotted the distribution of genome sizes of the 10 most abundant species (Figure 3.1) to check whether the bimodal distribution of genome size is common in other species.

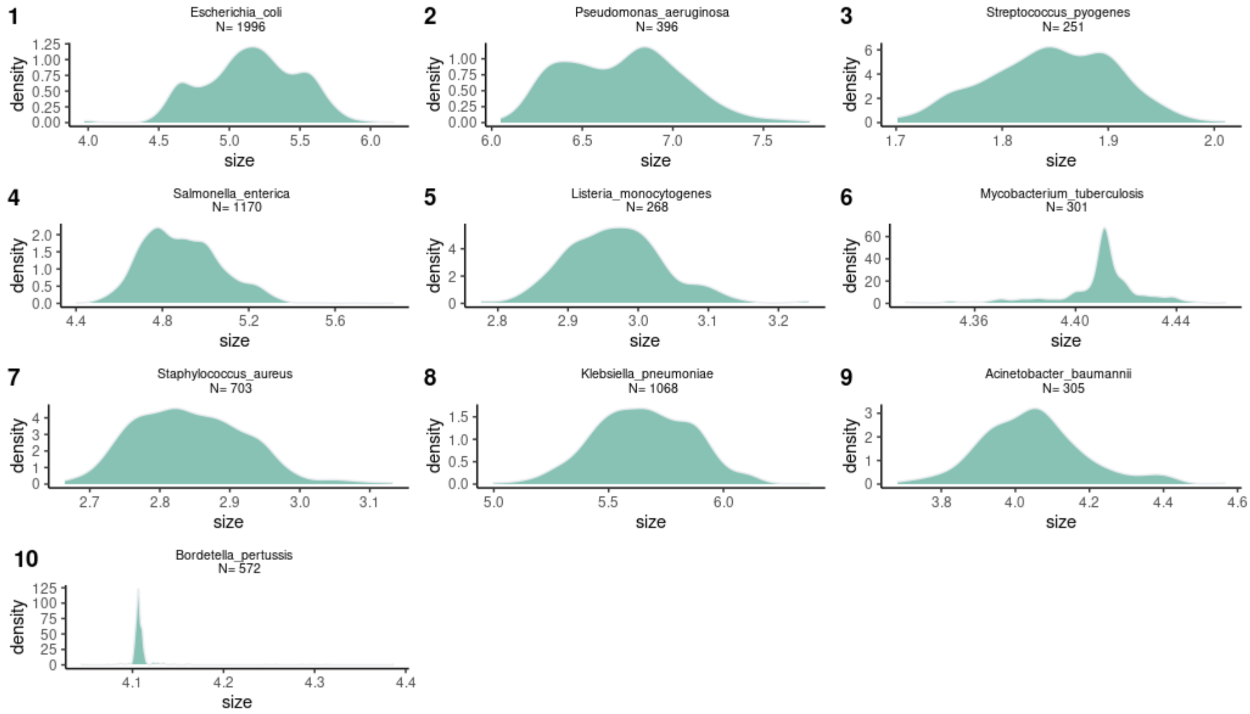


Figure 3.1 The distribution of genome size of each species. The x-axis is the genome size (Mb) and the y-axis is the density. The mean distribution and the total number of individuals are also displayed.

We observe the distinct bimodal distribution of genome size only in *P. aeruginosa* and *S. pyogenes* shows an ambiguous bimodal pattern. Most species show unimodal distribution of genome size, and *E. coli* shows trimodal distribution of genome size. The *P. aeruginosa* strains have the largest genome size (over 6Mb) and the *S. pyogenes* strains have the smallest genome size. Some species have larger genome size ranges, such as *E. coli*, *P. aeruginosa*, and *K. pneumoniae*. In contrast, *M. tuberculosis*, *S. pyogenes*, *L. monocytogenes,* and *B. pertussis* have a much narrower range of genome size. *B. pertussis* has less diversity in genome size, and almost all strains have a genome size very close to 4.1 Mb.

Genome size may be a potential indicator of genome diversity. The species with extremely low diversity of genome size may be associated with higher similarity of the genome. From the information in the Genome Neighbor section of the NCBI Genome, the *B. pertussis* strain VS401 has a symmetrical identity (genome similarity) of over 98% with more than 500 genomes. For *M. tuberculosis*, strain MTB1 has a symmetrical identity of over 98% with more than 274 *M. tuberculosis* genomes. These data from the Genome Neighbor section of the NCBI Genome indicate that many genomes of *B. pertussis* and *M. tuberculosis* share a high degree of similarity, which explains why the genome size distribution of these two species is so narrow.

The exact mechanisms that determine genome size are still unknown, but bacterial genome size is linked to several factors, such as resource availability and environmental stability (Gweon et al., 2017). The species with larger genomes are most likely to survive in an environment with scarce but diverse resources as they have more genes to adapt to new environments (Gweon et al., 2017). The bimodal distribution of genome size shows some forces may drive the evolution of genome size. In *P. aeruginosa*, the bimodal distribution of genome size is linked to the presence of CRISPR-Cas systems, the *P. aeruginosa* strains with CRISPR-Cas are associated with smaller genome sizes (Wheatley & MacLean, 2021). The CRISPR-Cas systems in these *P. aeruginosa* act as a constraint to the HGT (Wheatley & MacLean, 2021). The barrier to HGT prevents the acquisition of beneficial genes to adapt to new environments. And the HGT is one essential source for genome expansion, genomes with active CRISPR-Cas systems gain less genes transferred via HGT, resulting in a smaller size (Wheatley & MacLean, 2021). We wondered whether that would be the case for other species.

Next, we detected the CRISPR-Cas systems and investigated whether there is a relationship between the presence of CRISPR-Cas and genome size in other species.

3.2 Detection of CRISPR-Cas systems from CRISPRCasTyper

3.2.1 The abundance of CRISPR-Cas systems

The percentage of genomes encoding at least one complete CRISPR-Cas system in 10 species is shown in Table 3.1. Globally, 54.35% of all genomes encode at least one complete CRISPR-Cas system. The abundance of CRISPR-Cas varies among the 10 species, *S. enterica*, and *M. tuberculosis* have the highest percentage of genomes encoding CRISPR-Cas, and *B. pertussis* has no strains with the CRISPR-Cas system. Our results are close to the research conducted before, such as 0.8 % in *S. aureus* (Cruz-Lopez et al., 2021), 41.2% in *K. pneumoniae* (Li et al., 2018), 50% in *P. aeruginosa*(Wheatley & MacLean, 2021) and 42% in all bacteria(Makarova et al., 2020). Since no CRISPR-Cas has been detected in *B. pertussis* (even no orphan CRISPR array or Cas proteins detected), it is excluded from the following analysis.

After checking the CRISPR-Cas loci, most CRISPR-Cas systems are encoded on chromosomes and some of them are encoded on plasmids. Only 0.26% (N=4) of *E. coli* and 10.55% (N=103) of *K. pneumoniae* encode CRISPR-Cas on their plasmids. The 4 genomes of *E. coli* with plasmid-encoded CRISPR-Cas all have chromosome-encoded CRISPR-Cas and the CRISPR-Cas encoded on the plasmids are subtype IV-A3. In *K. pneumoniae*, all plasmid-encoded CRISPR-Cas are subtype IV-A3. And 52 of the *K. pneumoniae* genomes have both plasmid-encoded CRISPR-Cas (subtype IV-A3) and chromosome-encoded CRISPR-Cas (type I-E).

To better understand and compare the CRISPR-Cas of each species, we also investigated the subtypes and spacer numbers of the CRISPR-Cas of each species.

Table 3.1 The results of CRISPR-Cas detection by CRISPRCasTyper. The total number of genomes of each species, and percentage of CRISPR-Cas + (the genomes encode CRISPR-Cas) and the types of CRISPR-Cas are included.

Species	Total number	CRISPR-Cas+ percentage	CRISPR-Cas types
Escherichia coli	1996	77.10%	I-E, IV-A3, I-F, I-E&IV-A3, I-3&I-F
Salmonella enterica	1170	92.80%	I-E
Klebsiella pneumoniae	1068	32.00%	IV-A3, I-E, I-E&IV-A3
Staphylococcus aureus	703	1.28%	II-A
Bordetella pertussis	572	0%	none
Pseudomonas aeruginosa	396	50.10%	IV-A2, I-F, I-C,I-E,IV-A1
Acinetobacter baumannii	306	19.90%	I-F
Mycobacterium tuberculosis	302	100%	III-A
Listeria monocytogenes	268	19.90%	I-B, II-A, I-B&II-A
Streptococcus pyogenes	251	73.30%	I-C, II-A, I-C&II-A

3.2.2 The subtype of CRISPR-Cas possessed by nine species

We observe 8 subtypes (I-F, I-C, I-E, I-B, IIA, III -A, IV -A1, and IV -A3) of CRISPR-Cas from four types (type I, type II, type III, and type IV) and two classes (class 1 and class 2) of CRISPR-Cas (Figure 3.2). Type I is the most abundant type among the nine species.

Some species have more than one subtype of CRISPR-Cas, such as *P. aeruginosa* with 5 subtypes and *E. coli* with three subtypes. Some species have a considerable proportion of genomes encoding two types of CRISPR-Cas systems. For example, 4.85% of *L. monocytogenes* encode I-B and II -A and 4.96% of *K. pneumoniae* encode I-E and IV -A3.

Most CRISPR-Cas types are encoded on bacterial chromosomes, and only one plasmid-encoding CRISPR-Cas subtype (type IV -A3) is enriched in *K. pneumoniae* (9.83%). It is mainly encoded on plasmids and lacks the *cas* gene for adaptation and interference, which may be involved in plasmid competition (Kamruzzaman & Iredell, 2019). We then tested if the genomes with plasmid-encoding CRISPR-Cas have more plasmids. Using the Wilcoxon test (p =3.7e-07), we don't observe that the genomes with type IV-A3 tend to have a lower number of plasmids (Figure A.1). Based on the hypothesis that the plasmids encoding the IV -A3 type compete only with the plasmids with the same properties and lifestyles (Moya-Beltran et al., 2021), the information about whether the plasmids have the overlapping niche or function is not available. Genomes with a larger number of plasmids are more likely to have a complete CRISPR-Cas detected. In addition, more than half of the type IV-A3 co-exist with type I-E, suggesting the crosstalk between type IV-A3 and other types of CRISPR-Cas. The type IV-A3 may compensate for the lack of adaptation by using the Cas1-Cas2 adaption machinery from type I-E encoded on the chromosomes (Moya-Beltran et al., 2021). And the plasmid encoding CRISPR-Cas tends to have a larger size and high conjugative transmissibility (Pinilla-Redondo et al., 2022).

Overall, we observe a high diversity of CRISPR-Cas, suggesting an arms race between bacteria and viruses, such as coevolution between the bacterial CRISPR-Cas system and anti-CRISPR proteins (Makarova et al., 2020). The different mechanisms, various Cas proteins, and diverse architecture of different types of the CRISPR-Cas make the bacteria more resistant to different MGEs (Makarova et al., 2020). If a virus carries Acrs that can inhibit certain types of CRISPR-Cas, the bacteria can benefit from encoding two subtypes of CRISPR-Cas systems. However, the fitness burden also increases when multiple CRISPR-Cas systems are encoded, which may explain why only a small proportion of individuals carry two subtypes of CRISPR-Cas. According to the pan-immune system model, the bacteria can share the defense systems as community resources (Bernheim & Sorek, 2020). The presence of CRISPR-Cas on MGEs suggests that CRISPR-Cas systems can be transferred between related strains through HGT (Bernheim & Sorek, 2020). It is not necessary for one strain to carry all types of CRISPR-Cas as it can gain the CRISPR-Cas from closely related strains.

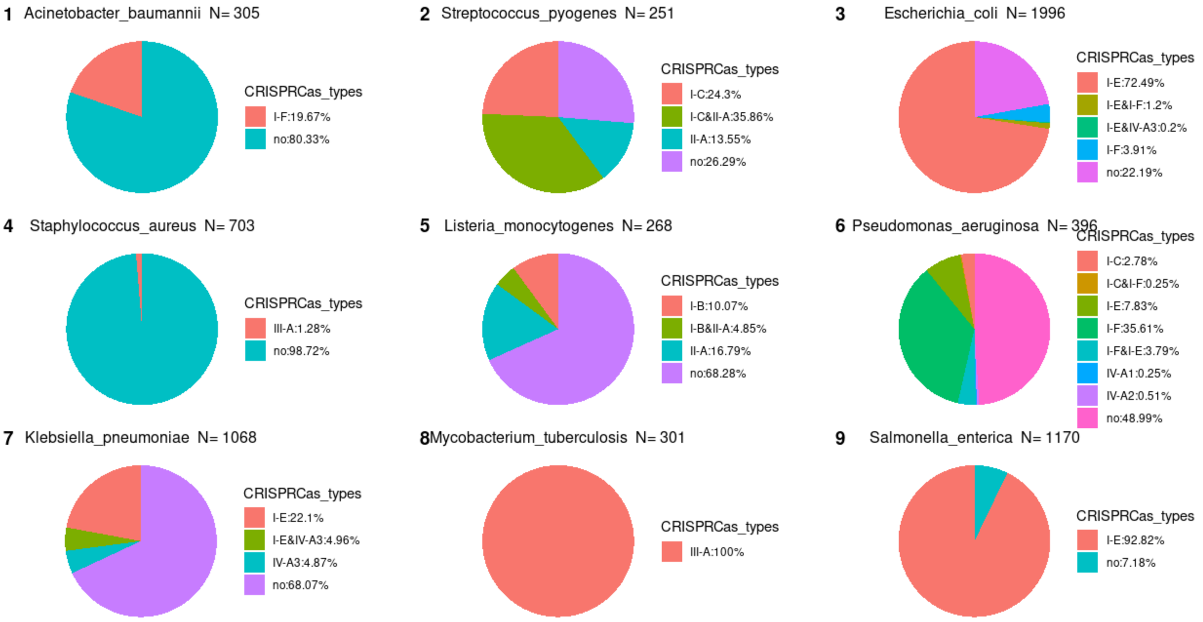


Figure 3.2 The subtype of CRISPR-Cas of each species. Color represents the proportion of genomes encoding certain CRISPR-Cas types. The N is the total number of genomes with CRISPR-Cas systems.

3.2.3 The distribution of the number of spacers from CRISPR-Cas arrays

The spacer repository can serve as a list of potential targets that the CRISPR complex can target. With more spacers, the CRISPR-Cas can target more MGEs. The distribution of spacer number in each species is quite diverse (Figure 3.3), ranging from 1 to 100 in most species, with some strains of *A. baumannii* having more than 200 spacers. The genome with the most spacers is GCF_019703285.1, which has two type I-F CRISPR-Cas systems encoded on its genome. It is reasonable that genomes encoding two complete CRISPR-Cas have a greater number of spacers. The presence of two identical types of CRISPR-Cas is not common in *A. baumannii* (2/65), and both have a vast number of spacers (185 and 294). In addition, two *E. coli* genomes encoding two I-E have 36 and 24 spacers, and one genome of *S. enterica* encoding two I-E has 108 spacers. According to a statistical test, the genomes encoding two complete CRISPR-Cas systems are associated with more spacers (Figure A.3).

Figure 3.3 includes the spacers both from orphan CRISPR arrays and Cas-adjacent CRISPR arrays. The distribution of spacers from orphan CRISPR arrays and the spacers from the Cas-adjacent CRISPR arrays can be seen in Figure A.2. For the orphan array, *E. coli* (72%), *L. monocytogenes* (63.5%), *P. aeruginosa* (56.4%), and *M. tuberculosis* (24.9%) have a relatively larger proportion of genomes encoding the orphan arrays, and the proportion is less than 5% for the other species (see Figure A.2). Comparing the number of spacers of the orphan CRISPR arrays with the spacers of the Cas- adjacent CRISPR arrays, the spacers of the Cas-adjacent CRISPR arrays have more spacers, suggesting that the adjacent adaptation module can contribute to the expansion and maintenance of the array.

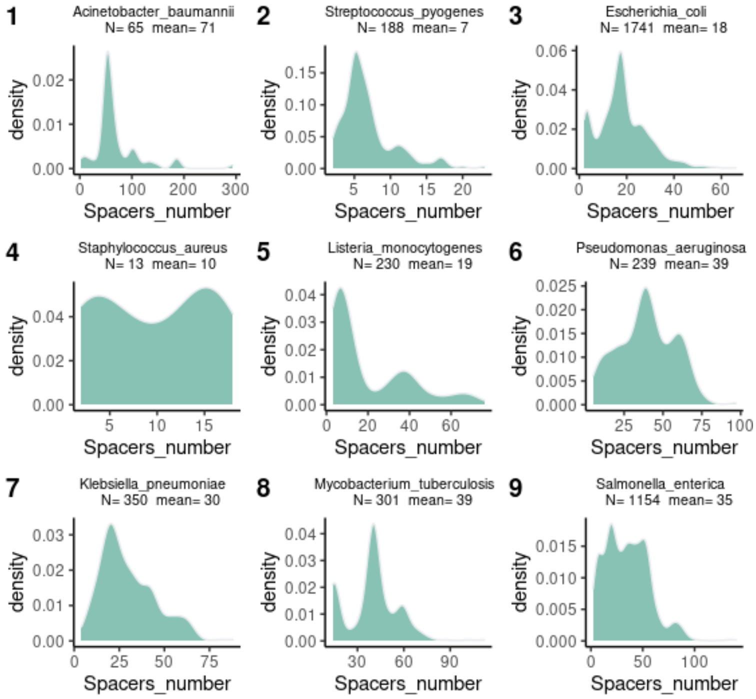


Figure 3.3 The distribution of spacer number in nine species. The N is a total number of genomes with CRISPR arrays. The mean is the mean number of spacers possessed by genomes with spacers. The spacers are from orphan CRISPR arrays and Cas operons adjacent CRISPR array.

The number of spacers possessed by CRISPR-Cas varies among different subtypes of CRISPR-Cas. Figure S 6.3.5 shows that type I-F tends to have more spacers than type I-F in *P. aeruginosa* and *E. coli*. Type I-B tends to have more spacers than type II -A in *L. monocytogenes* and type I-E tends to have more spacers than IV -A3.

The number of spacers is commonly accepted as a trade-off between the high physiological burden of maintaining long spacers and better protection against broad enemies(Levin et al., 2013). However, one model suggests that longer arrays cause few physiological burdens but have a dilution effect on the CRISPR complex, which is armed with young spacers that carry the sequence information of the most recently evolved (Martynov et al., 2017). With few spacers, the host can be protected against only a few types of viruses, but the young spacers can be used more effectively than the host with long arrays. CRISPR-Cas favors short arrays in the environment with highly mutated viruses and it favors long arrays in the environment with a more diverse virus (Martynov et al., 2017). The strains that have more spacers may have more spacers for targeting HGT-related genes, but the efficiency may be quite low due to the dilution effects.

3.3 The detection of Acr loci with AcrFinder

As Acrs can inhibit CRISPR-Cas activity, the Acr loci are detected to find genomes with active CRISPR-Cas systems (Figure 3.4).

As we described in section 2, AcrFinder can detect Acrs via two routes, the homology-based search route and guilt-by-association route. The homology-based search route is more convincing, but the Acrs of different families are quite diverse and the homology-based search route is quite limited (Yi et al., 2020). The GBA route can identify conserved Aca loci and combine other information such as the presence of nearby self-targeted spacers to infer the presence of Acrs with three confidence levels (Yi et al., 2020).

S. aureus has the largest percentage of Acrs detected via homology-based search route (Figure 3.4). For the other 8 species, the majority of Acrs are inferred from the GBA route. For some species (*P. aeruginosa*, *L. monocytogenes, K. pneumoniae*, *A. baumanni*), the large proportions of the medium- and high-confidence levels to low-confidence level results show that the genomes with Aca are common to have self-targeted spacers. In *S. pyogenes*, *S. enterica*, *P. aeruginosa*, *E. coli* and *M. tuberculosis*, most genomes have Aca loci.

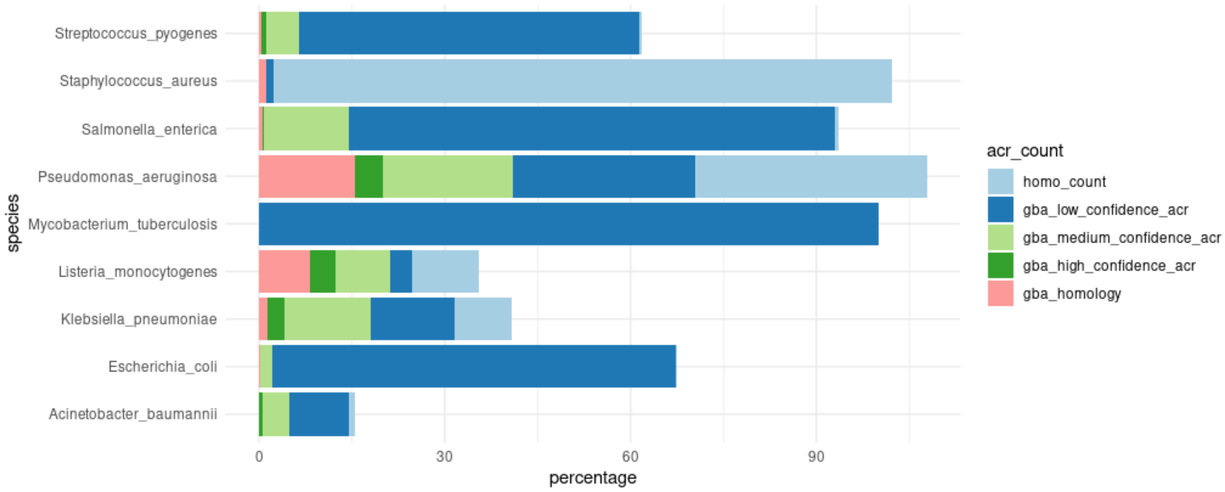


Figure 3.4 The detection results of each species via two routes. The percentage is calculated as the count of genomes with detected Acrs divided by the total number of genomes of each species. The homo_count is the number of Acrs only detected from homology-based detection routes. The gba_low condifence _acr, the gba_medium_confidence_acr and the gba_high_confidence_acr are the Acrs of three confidence levels only detected from GBA route. The gba_homology represents the Acrs detected by both routes.

In the specie of *S. aureus*, we observed a large percentage (99.96%) of Acrs detected via the homology-based route (most convincing results), but only 1.28% (N=9) have a complete CRISPR-Cas system and all detected CRISPR-Cas systems are subtype III-A. The Acrs show high specificity of targeted subtypes of CRISPR-Cas, and they are named based on the subtypes which they can inhibit. It is unexpected that the abundance of Acrs existed on the genomes without active CRISPR-Cas. Yin *et al*. suggest this phenomenon may be caused by the absence of CRISPR-Cas which can make the invasion of phage with Acrs much easier (Yin et al., 2019). Next, we suggest two other possible reasons for the unexpected abundance of Acrs.

The Acrs found in *S. aureus* are AcrIIA14 and AcrIIA15, which can inhibit the *Staphylococcus aureus* Cas9 (Ran et al., 2015). *Staphylococcus aureus* Cas9 (SauCas9) which is widely applied in biotechnology (Yourik et al., 2019), belongs to type II CRISPR-Cas. However, type III-A CRISPR-Cas is only identified in *S. aureus* in our results, which does not contain Cas 9 (Makarova et al., 2020). We suggest that the first possible reason is the unsuccessful detection and classification of the type II-A CRISPR-Cas system of *S. aureus* using CRISPRCasTyper. CRISPRCasTyper can reach about 89% of accuracy of detecting CRISPR-Cas subtype III-A (Russel et al., 2020). Our detection is close to the results of Cruz-Lopez *et al.* (Cruz-Lopez et al., 2021), who used CRISPRCasFinder to identify 0.83% of 716 *S. aureus* with the CRISPR-Cas subtype III-A. In the literature, we found that *SauCas 9* is a less well-characterized homolog with only 17% similarity (Yourik et al., 2019). CRISPRCasTyper uses HMMs-based homology search to find the Cas protein based on translated genomic data. The sensitivity of detecting Cas 9 from *S. aureus* might be low. To validate it, we found that SauCas is extracted from *Staphylococcus aureus subsp*. *Aureus* (Ran et al., 2015). The reference genome of *Staphylococcus aureus subsp. Aureus* is strain NCTC 8325 (GCA_000013425.1 ASM1342v1), which is already included in our data. The detection of type II-A CRISPR-Cas from this genome with CRISPRCasTyper is unsuccessful, only a low -quality CRISPR array is detected.

Acrs detected based on homology are also abundant in the strains of *P. aeruginosa* without a complete CRISPR-Cas. The presence of orphan arrays suggests the evolution loss of CRISPR-Cas when the CRISPR-Cas systems are inactivated for long-term. The orphan CRISPR arrays are likely the remnants of decaying CRISPR-Cas (Shmakov et al., 2020). CRISPRCasTyper results indicate that about 30% of *S. aureus* have an orphan CRISPR array (including low-quality detection). The abundance of CRISPR arrays suggests the possible evolution loss of the CRISPR-Cas system.

Next, the inhibitory activity of Acrs is checked to obtain the genomes with Arcs that can inhibit CRISPR-Cas encoded on the same genome. The dual inhibitory activity of some Acrs (Table 2.3) is also considered. 148 genomes encode two different types of chromosomal CRISPR-Cas systems and 14 of them (from three species, *L. monocytogenes, S. pyogenes*, *P. aeruginosa*) have detected Acrs. After manual inspection, we found that 4 of them are fully inhibited and 10 of them are supposed to be partial inhibited. The genomes encoding two different types of CRISPR-Cas systems are still considered as being active when the Acrs can only target one type of CRISPR-Cas.

Finally, 120 cells have both CRISPR-Cas and Acrs detected via the homology-based route, and 55.8% (n=67) of them possessing Acrs can target the subtype of CRISPR-Cas encoded on the same genome. From the GBA route, the inferred Acrs with low confidence level are excluded and the inferred Acrs with medium and high confidence level are extracted (N=482). At medium and high confidence level, the self-targeted CRISPR-Cas is present in the genome, indicating that CRISPR-Cas cleavage activity is most likely inhibited to avoid self-immunity. After merging the results from two routes, 499 genomes are considered to have inhibited CRISPR-Cas.

The equipment of two types of CRISPR-Cas in their genome could be a strategy of bacteria to evolve against the inhibition of Acrs. Only 2.7% (4/148) of strains encoding two types of CRISPR-Cas are expected to lose their CRISPR-Cas-mediated immunity. Compared with the overall situation ,499/3822 (13%) of the strains encoding CRISPR-Cas are expected to lose their CRISPR-Cas activity. The strains that have two types of CRISPR-Cas have at least some CRISPR-Cas activities to defend the invaded MGEs when an Acr is present. And the diversity and dual inhibition of Acrs also show the dynamics of the army race between phages and bacteria.

3.4 The relationship between CRISPR-Cas, Acrs, genome size and HGT

Next, we merged all data about CRISPR-Cas systems, genome size, and Acrs from 8 species (*B. pertussis* and *M. tuberculosis* are excluded because they either have no CRISPR-Cas encoding genome or all genomes have CRISPR-Cas) to explore whether there is a consistent relationship between the active CRISPR-Cas and genome size in 8 species.

3.4.1 The relationship between CRISPR-Cas and genome size

3.4.1.1 The presence of CRISPR-Cas and genome size

After adding the information related to CRISPR-Cas into plots of the distribution of genome size, we can observe two separate peaks related to the presence of CRISPR-Cas in *P. aeruginosa*. The blue region representing the strains with complete CRISPR-Cas systems have smaller size than the red region representing the strains without complete CRISPR-Cas systems (Figure 3.5). Combined with the plots of distribution of genome size, we can see that bimodal distribution of genome size of *P. aeruginosa* is linked to presence of CRISPR-Cas systems.

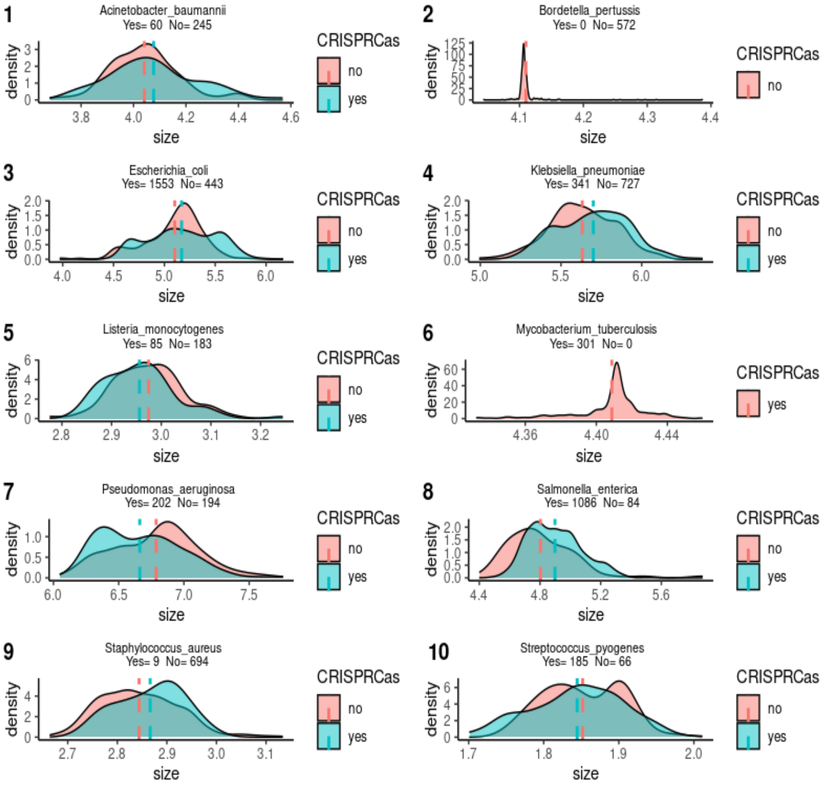


Figure 3.5 The distribution of genome size associated with the presence of the CRISPR-Cas system. The red region represents the genomes without CRISPR-Cas and the blue region represents the genome without the CRISPR-Cas encoded. The dash lines are the mean genome size of two groups.

And the bimodal distribution of genome size must meet at least two conditions. The first condition is that the total number of strains without CRISPR-Cas should be close to the total number of strains with CRISPR-Cas systems (50% for *P. aeruginosa*). And the second condition is that the strains with the CRISPR-Cas are associated with smaller genome sizes. The percentage of genomes encoding the

CRISPR-Cas system for each species is already shown in Table 3.1, most species do not meet the first condition. Next, the second condition is tested for each species.

The relationship between the presence of CRISPR-Cas and genome size varies in 10 species. In *P. aeruginosa* and *L. monocytogenes,* the presence of CRISPR-Cas is associated with smaller genome size (Figure 3.6). The pattern that the presence of CRISPR-Cas is associated with smaller genome size in *P. aeruginosa* is consistent to previously conducted research (Wheatley & MacLean, 2021). Wheatley and MacLean suggest that the presence of CRISPR-Cas is most likely due to the restriction of HGT caused CRISPR-Cas. Since the HGT is the key source of genome expansion (Wheatley & MacLean, 2021), the restriction of HGT could result in a significantly shorter genome. However, species such as *S. enterica*, *K. pneumoniae* and *E. coli* show a reverse pattern, that the presence of CRISPR-Cas is associated with larger size (Figure 3.6). No significant difference is observed in the remaining species, suggesting that there is no relationship between genome size and the presence of CRISPR-Cas for these species.

It seems that the hypothesis that CRISPR-Cas restricts HGT and affects genome size is not true for the other 6 species.

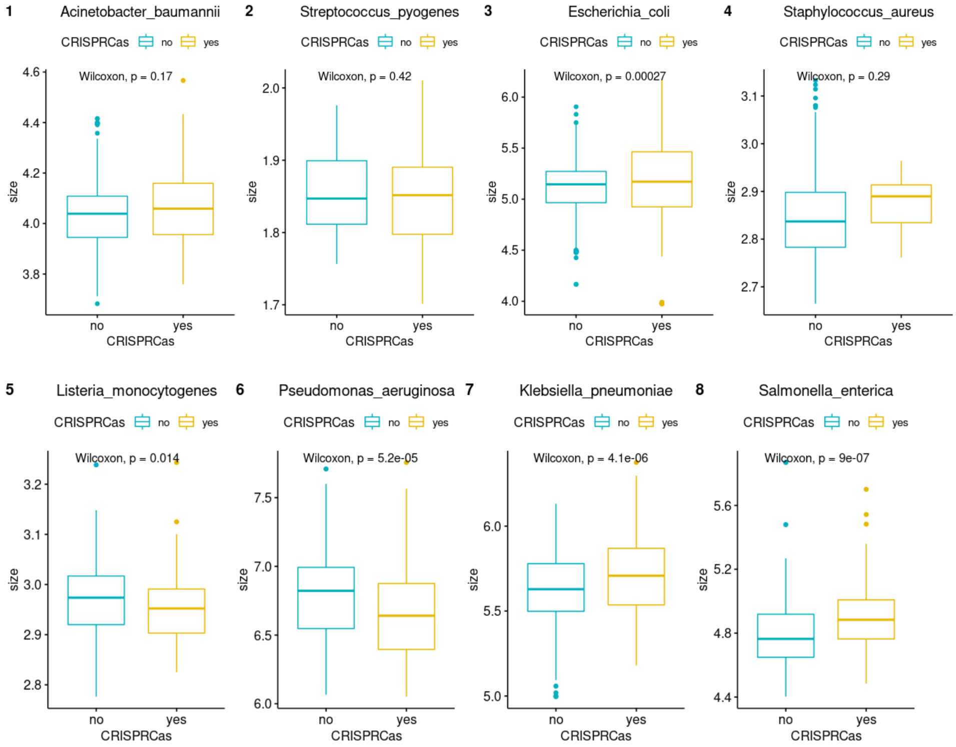


Figure 3.6 The boxplot of distribution of genome size vs the presence of CRISPR-Cas system. The p-value (0.05 is selected as the threshold) from the Wilcoxon test is calculated to show the significance. Only in *P. aeruginosa* and *L. monocytogenes*, the presence of CRISPR-Cas is associated with smaller genomes. In *S. enterica*, *K. pneumoniae* and *E. coli,* the presence of CRISPR-Cas is associated with larger genomes. For the rest of the species, no association is found.

3.4.1.2 The relationship between the subtype of CRISPR-Cas and the genome size

As five species encode multiple types of CRISPR-Cas and each type of CRISPR-Cas has different mechanisms and genetic architectures (Payne et al., 2021), which may have different relationships with genome size. The relationship between subtype of CRISPR-Cas and genome size in these species is also investigated here (Figure 3.6).

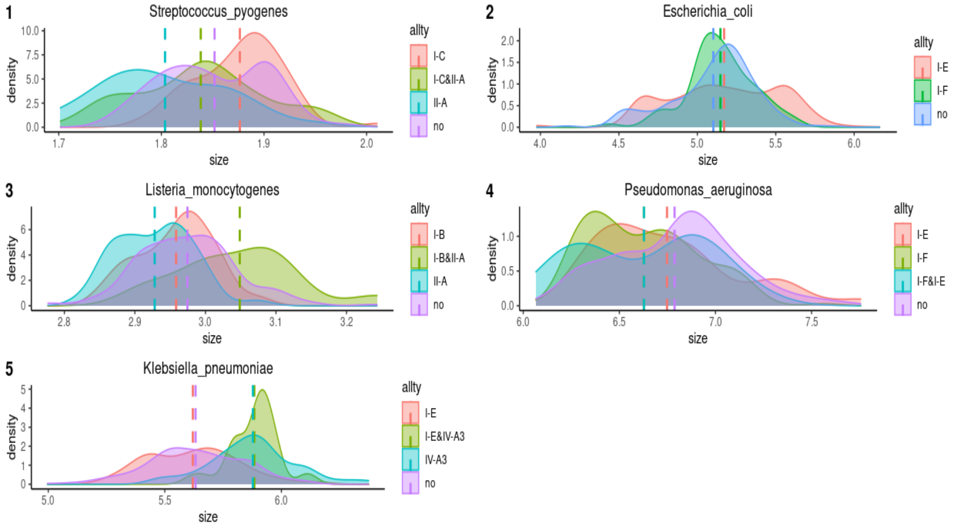


Figure 3.7 The distribution of genome size of different subtypes of CRISPR-Cas in five species.

The exact subtype of CRISPR-Cas of different species may have different relationships, and no consistent pattern is observed (Figure 3.7, 3.8). The presence of type I-F shows a significant negative correlation with genome size in *P. aeruginosa*, but the correlation between type I-F and genome size is reversed in *E. coli*. In *S. pyogenes*, the presence of type II-A is associated with the smaller genome, and this pattern is also observed in *L. monocytogenes.*

The different subtypes of CRISPR-Cas in the same species can also have different patterns (Figure 3.7, 3.8). In *S. pyogenes*, the type II-A is associated with a smaller genome, however, the subtype I-C is associated with a larger genome. The presence of both II-A and I-C has no significant correlation with genome size, suggesting that the combination of two subtypes may release or neutralize the opposite patterns. In *L. monocytogenes,* the genome carrying II-A tends to have a smaller genome size, but the genome encoding both II-A and I-B tends to have a significantly larger size. In *K. pneumoniae*, with the exception of type I-E, the presence of other subtypes is associated with larger genome size. IV-A3 is the plasmid-encoded type of CRISPR-Cas, and we show that the presence of IV-A3 is also associated with more plasmids. Plasmids are also essential sources of genome size but less stable than chromosomes. Some systemic analysis studying plasmids encoding CRISPR-Cas indicate the plasmids encoding CRISPR-Cas tend to have larger genome size (Pinilla-Redondo et al., 2022).

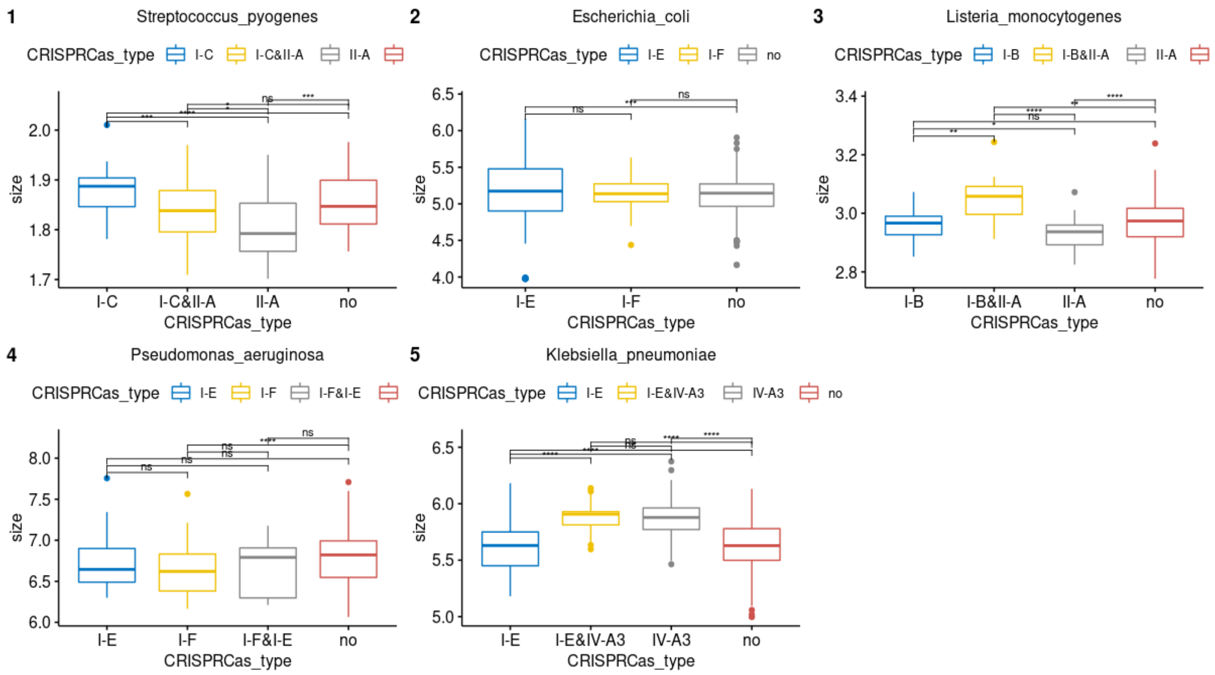


Figure 3.8 The significance test of the presence of different subtypes of CRISPR-Cas and genome size of five species. The subtype that presents less than 3% of total genomes is excluded as it is too small to have statistical meaning. The p value of the Wilcoxon test is calculated and displayed as "ns", ".","…" and "…". The ns is not significant, "." represents the p value ranging from 0.05 to 0.005, ".." represents the p value ranging from 0.005 to 0.001, "…" represents the p value smaller than 0.001.

3.4.2 The relationship between the genome size, Acrs and CRISPR-Cas

The presence of specific Acrs is expected to inactivate the CRISPR-Cas and abolish the restriction of genome expansion caused by CRISPR-Cas. Acrs have been shown to enhance conjugation of the target plasmid and thus can facilitate horizontal gene transfer (Mahendra et al., 2020). We hypothesize that genomes encoding both CRISPR-Cas and functional Acrs (CRISPRCas+Acr+) will have a larger genome size than genomes encoding only CRISPR-Cas (CRISPRCas+Acr-) because the presence of Acrs can inhibit CRISPR-Cas activity and enhance HGT. The Wilcoxon paired test is performed with p = 0.05 as a criterion to test it, we can observe that the genome size of CRISPRCas+Acr+ genomes is larger than that of CRISPRCas+Acr- genomes in 5 species (*A. baumannii, S. pyogenes*, *L. monocytogenes*, *P. aeruginosa* and *S. enterica*) (Figure 3.9). Most species (5/7) meet our hypothesis that the Acr can enhance the HGT.

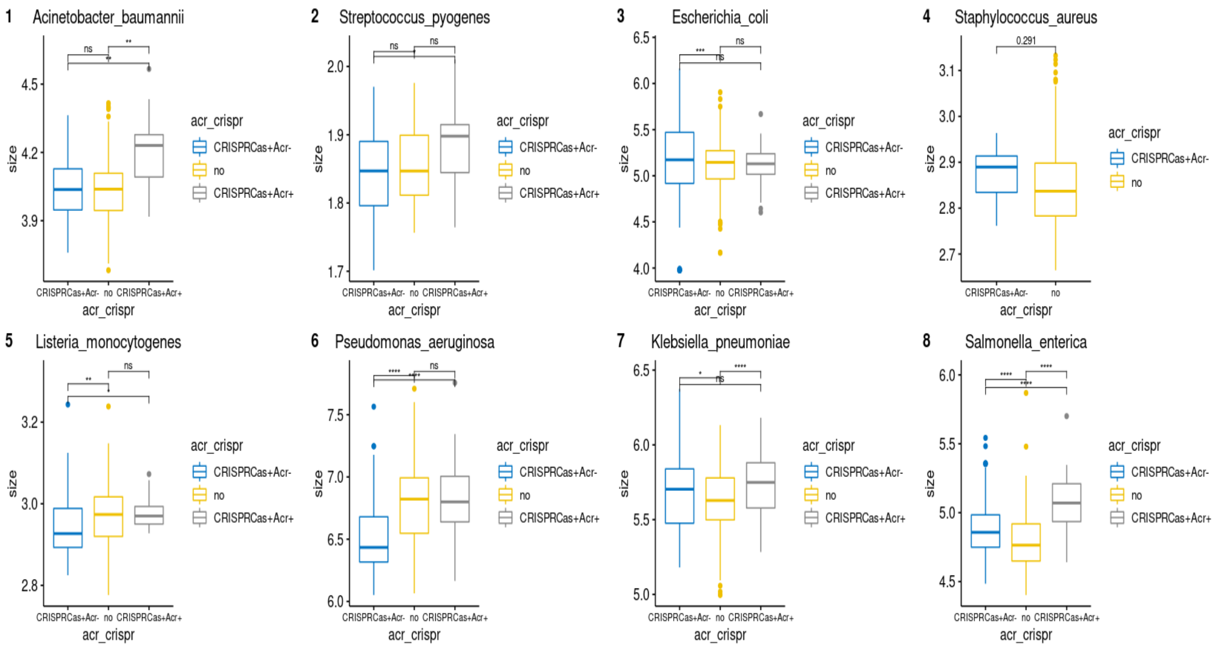


Figure 3.9 The genome size (Mb) of three groups of population. "no" represents the strain without CRISPR-Cas, and the CRISRPCas+Acr- is the group encoding CRISPR-Cas and no targeting inhibiting Acrs. The CRISPRCas+Acr+ is the group encoding CRISPR-Cas but have inhibiting Acrs. The *S. aureus* has no Acrs detected, so it has only two groups.

We investigate why *E. coli* and *K. pneumoniae* are not consistent with other species. *K. penumoniae* and *E. coli* have plasmids in their genomes. A specific subtype of CRISPR-Cas (type IV -A3) without cleavage activity is abundant in *K. penumoniae* plasmids. The presence of the IV -A3 is associated with a larger genome size due to the presence of plasmids in their genome. The IV -A3 lacks the adaptation module and targeting function and it often coexists with other chromosome-encoding CRISPR-Cas (Kamruzzaman & Iredell, 2019). We excluded the *K. penumoniae* strains that encode only IV -A3 in their plasmids without chromosomal CRISPR-Cas and performed the analysis again. The difference becomes significant ($p < 0.05$) (Figure 3.10 A) and the genome size of CRISPRCas+Acr+ genomes is larger than that of CRISPRCas+Acr- in *K. penumoniae*. The CRISPR-Cas is not abundant on the plasmids of *E. coli*. And *E. coli* has the lowest percentage of strains encoding Acrs (41/1996 of strains), only 2 Acrs are detected based on homology search route, and 39 Acrs are inferred via GBA route.

Apart from the inhibition of CRISPR-Cas caused by Acrs, another possible explanation is that the larger genome size of the genome encoded by Acrs may be caused by the large uptake of the lysogenic phage into the host genome (Wheatley & MacLean, 2021). The host gains the Acrs by integrating the lysogenic phage genome encoding Acrs into the host genome. Acrs are vectored by mobile genetic elements, suggesting that the strains with high promiscuity toward mobile genetic elements have a high chance of encoding Acrs and greater genome expansion. We can observe the CRISPR-Cas+Acr+ is the group with the largest genomes compared to other two groups in some species (*A. baumanni*, *S. pyogenes*, *L. penumoniae* and *S. enterica*) (Figure 3.9). If the presence of Acrs is an indicator of high promiscuity toward mobile genetic elements, and CRISPR-Cas can restrict the promiscuity by targeting and cleaving some MGEs, leading to a smaller genome, this can also explain why the presence of Acrs is associated with larger genomes.

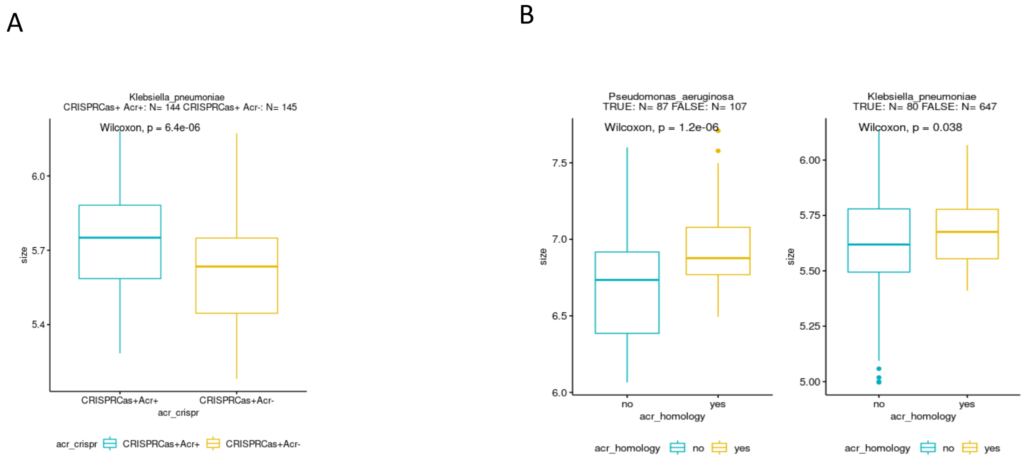


Figure 3.10 Significance test of genome size and other factors.

A: The significance test genome size of CRISPRCas+Acr+ and CRISPRCas+Acr- groups of K. *penumoniae* after excluding the strains have only Type IV-A3 encoded on their plasmids. After the exclusion, the presence of Acrs is associated with larger genomes. B: The significance test of genome size of two groups without presence of CRISPR-Cas. “no” represents the group without CRISPR-Cas and no Acrs detected. “Yes” group represents the group without CRISPR-Cas but have Acrs detected from homology-based search route using AcrFinder. The presence of Acr in CRISPR-Cas- group is also associated with larger genomes.

To test this possibility, strains not encoding CRISPR-Cas (CRISPR-Cas-) were selected and the differences in genome size between the groups with and without Acrs were compared. From the workflow of AcrFinder, we know that for the GBA route, the Acrs are only inferred when the CRISPR-Cas is present. Thus, we can only use the Acrs that were detected based on homology, resulting in a low number of detected Acrs. The species with a low number of samples in both groups (N < 10) are excluded and only two species are shown. The two selected species show that the strains encoding Acrs in their genome tend to be larger (Figure 3.10 B). The results also support the hypothesis that the larger genome size may be associated with high promiscuity to MGEs.

Comparing the group with an active CRISPR-Cas systems (CRISPR-Cas+Acr-) with the group without CRISPR-Cas systems (Figure 3.9), only *L. monocytogenes* and *P. aeruginosa* shows that the strains with active CRISPR-Cas are associated with smaller size, while *E. coli*, *K. pneumoniae* and *S. enterica* keep showing the reverse pattern. The relationship between genome size and the presence of active CRISPR-Cas is the same as the previous section (Figure 3.6). Active CRISPR-Cas systems can restrict the HGT in *L. monocytogenes* and *P. aeruginosa*, but it can have little or no restriction on genome size in other species. Interestingly, the presence of Acrs is associated with larger genome size in the CRISPRCas+ group of most species and the CRISPRCas- group of *P. aeruginosa* and *K. pneumoniae.* This indicates that Acrs may enhance HGT through other mechanisms or that the presence of Acrs is an indicator of high promiscuity to mobile genetic elements.

The relationship between the Acrs, genome size and subtypes of CRISPR-Cas is also investigated (Figure A.5 and Figure 3.10.1). Most of the relationships are same as Figure 3.8, the only difference is that the relationship between active I-E and genome size becomes significant in *P. aeruginosa*.

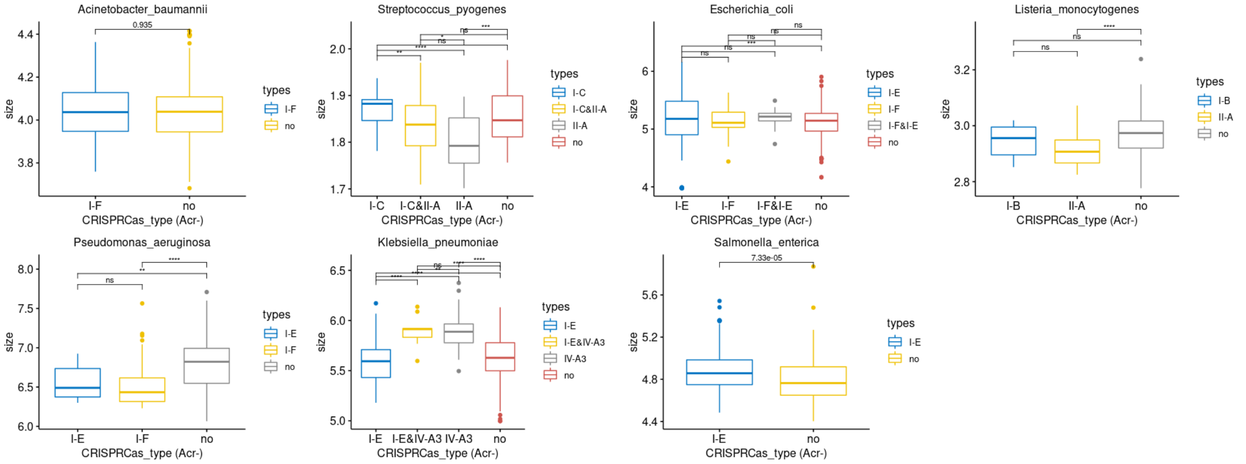


Figure 3.10.1 The relationship between the genome size and various subtypes of active CRISPR-Cas. The subtypes with less than 20 genomes are removed.

3.5 The investigation of targets of spacers of CRISPR arrays

3.5.1 The targets of total spacers of each species.

Next, we identify the spacer targets for the direct evidence that CRISPR-Cas systems can restrict HGT.

We observed some genomes (N=1342) encoding a complete CRISPR-Cas system and a CRISPR array. The orphan arrays are CRISPR arrays that are far from *cas* operons. It is unclear whether orphan arrays co-existing with a complete CRISPR-Cas system can function or not, we extracted the spacers from the orphan CRISPR arrays and the spacers from the complete CRISPR-Cas system and studied them separately (Figure 3.11).

Comparing the targets of CRISPR arrays from two different sources (Figure 3.11), the composition of spacer targets is very similar in most species (except *L. monocytogenes*). The close composition of the arrays from two sources suggests that the orphan arrays were recently functionally active (Shmakov, Utkina, et al., 2020). In *L. monocytogenes*, the proportion of spacers from orphan arrays targeting plasmids is increasing. To make the results more precise, the spacers from cas operon adjacent CRISPR arrays are investigated.

In *K. penumoniae*, plasmids occupy a large proportion of spacers due to the presence of the plasmid-encoded type IV -A3, which contains a large proportion of spacers that target plasmids and can participate in competition between plasmids (Kamruzzaman & Iredell, 2019). Plasmids also take a considerable proportion of targets in other species except *S. pyogenes*. Only a smaller proportion of spacers can target ICE. *P. aeruginosa* has the largest proportion (5.24%) and other species have less than 2%.

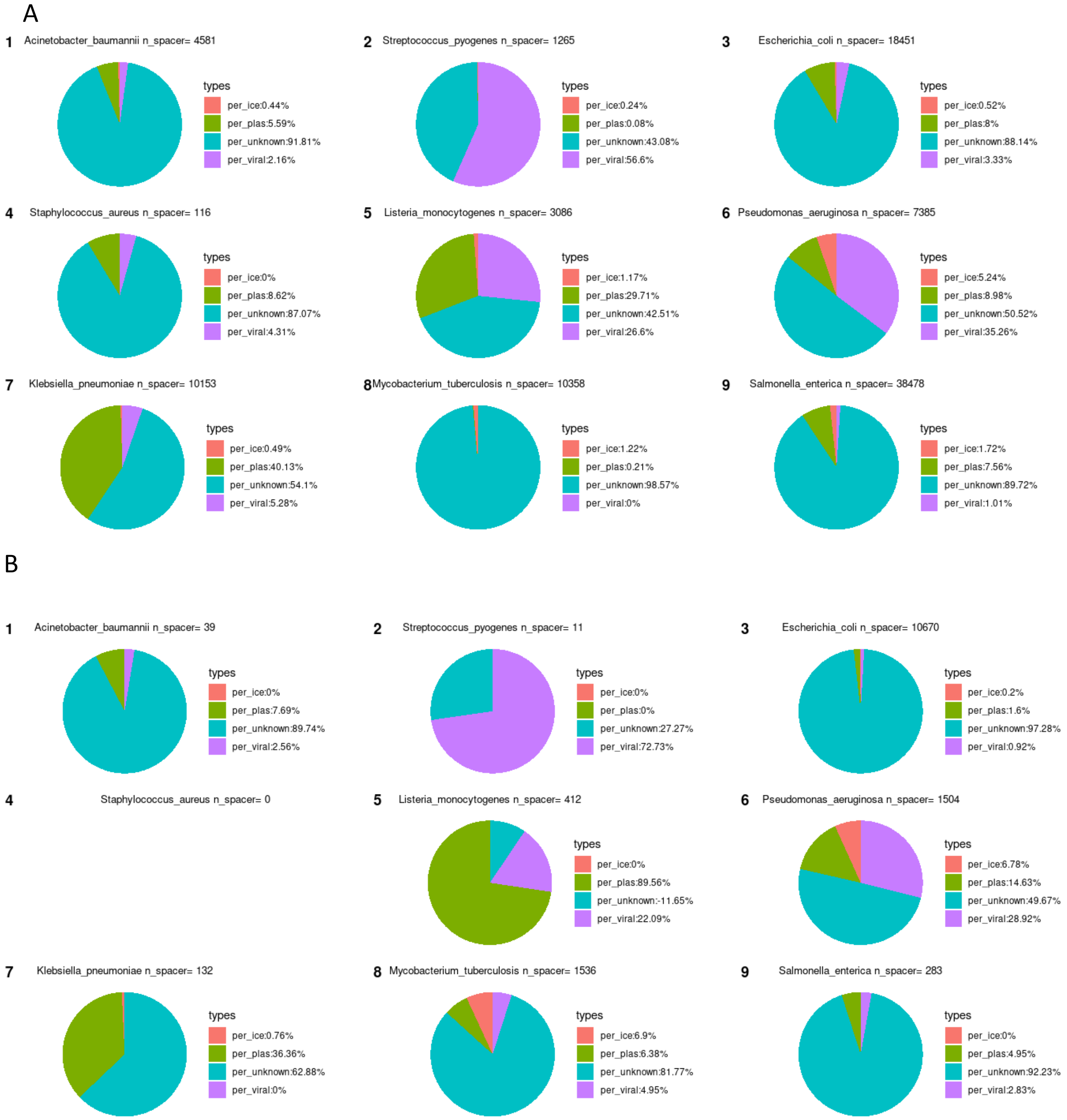


Figure 3.11 The proportion of targets of total spacers of each species. A: The spacer from the complete CRISPR-Cas system. B: The spacers from orphan CRISPR arrays. There is no orphan arry in S. aureus, so the field is blank.

In some species (*S. pyogenes*, *L. monocytogenes, P. aeruginosa*), viral genomes are the common targets for a large proportion of spacers because the evolutionary pressure of encountering phages force bacteria to have CRISPR-Cas with anti-phage activities. Spacers targeting the viral genetic sequence are thought to originate from host-specific viromes (Shmakov, Wolf, et al., 2020). Among these spacers, temperate phages are the most common targets in *P. aeruginosa*(Wheatley & MacLean, 2021).

In many species, more than half of the spacers have unidentified targets, and over 98% of spacers of *M. tuberculosis* have unknown targets. The dark matter of CRISPR may represent the unidentified mobile gene elements (Shmakov et al., 2017). And over 80% of the dark matter of the spacerome likely originates from the unsampled viromes (Shmakov, Wolf, et al., 2020).

We observed the distinct difference between the *P. aeruginosa* and other species is that *P. aeruginosa* has the most spacers targeting ICEs (5.24%) and for the rest of species, the proportion of spacers targeting ICEs is less than 2%. The highest proportion of spacer targeting ICEs is strong evidence that the CRISPR-Cas can restrict the HGT. Next, we calculate how many genomes in this species have at least one spacer targeting plasmids, ICEs and viral sequences (Table 3.2).

Table 3.2 The percentage of genomes with active CRISPR-Cas have at least one spacer targeting plasmid, ICE and viral sequences in 8 species.

Species	Plasmid	ICE	Viral sequences
Acinetobacter baumannii N= 47	93.62%	17.02%	38.30%
Streptococcus pyogenes N= 171	0.58%	1.75%	94.74%
Escherichia coli N= 1512	38.56%	6.22%	32.61%
Staphylococcus aureus N= 9	100.00%	0.00%	22.22%
Listeria monocytogenes N= 59	91.53%	25.42%	98.31%
Pseudomonas aeruginosa N= 105	95.24%	81.90%	98.10%
Klebsiella pneumoniae N= 195	93.85%	20.00%	48.72%
Salmonella enterica N= 925	49.51%	34.49%	17.19%

The species *P. aeruginosa* and *L. monocytogenes,* where there is a negative correlation between the presence of active CRISPR-Cas and genome size, both have a high percentage of genomes with spacers that can target plasmids (Table 3.2). CRISPR-Cas has been found to target conjugative plasmids and restrict HGT (Westra et al., 2013). In *P. aeruginosa*, the genomes with the spacers targeting ICEs are common (81.9%). *L. monocytogenes* has a lower percentage of ICE-targeting spacers (25.42%) but still higher than most species. *K. pneumoniae* has a high percentage of genomes with spacers that can target plasmids (93.85%) and ICEs (20%). But more than 50 genomes (25% of them) encode only type IV-A3 without cleavage activity, the abundance of effective spacers may be lower. *S. enterica* has the second largest percentage of genomes with ICE-targeting spacers (34.49%) and it only has 49.51% of the genome with plasmid-targeting spacers.

The longer CRISPR-Cas arrays may have a dilution effect on other spacers (Martynov et al., 2017), making the ICEs-targeting spacers less effective. And *A. baumannii* has the largest mean number of 75 (Figure A.2). The dilution effects of long CRISPR arrays may suggest that CRISPR-Cas in *A. baumannii* may be less effective to target plasmids and ICEs. The abundance of plasmid-targeting spacers and ICEs-targeting spacers in *L. monocytogenes* may suggest the association between smaller genome and presence of CRISPR-Cas in *L. monocytogenes.*

The abundance of plasmid-targeting and ICE-targeting spacers in *P. aeruginosa* could explain the distinct difference in genome size distribution between the CRISPR-Cas encoded group and the group without CRISPR-Cas group (Figure 3.4). The conservation of these spacers targeting ICEs among genomes of *P. aeruginosa* provides the evidence that CRISPR-Cas most likely prevents the acquisition of conjugative elements and puts large constraints on HGT. The much larger abundance of genomes with ICE-targeting spacers may mainly suggest the CRISPR-Cas can restrict the HGT in *P. aeruginosa*. The low abundance of plasmid-targeting spacers or ICE-targeting spacers in other species suggests the active CRISPR-Cas puts less or no constraint on HGT and makes little impact on genome size.

From section 3.1 to section 3.5, we note that the bimodal distribution of genome size is present only in *P. aeruginosa*. Because CRISPR-Cas can potentially restrict HGT and affect genome size, we investigate whether the presence of CRISPR-Cas is associated with genome size in other species. However, the association between the presence of CRISPR-Cas and lower genome size is only observed in *P. aeruginosa* and *L. monocytogenes*. The Acrs loci are discovered to detect the CRISPR-Cas with active activity. The association between active CRISPR-Cas systems and smaller genome size is still observed in *P. aeruginosa* and *L. monocytogenes*. Interesting, we found that the presence of Acrs is associated with larger genome size in the CRISPR-Cas+ group (6/7) and the CRISPR-Cas- group (2/2). We then identify the targets of the spacers. The much higher proportion of the genome with at least one ICE-targeting space can explain why *P. aeruginosa* is so unique.

3.6 The Detection of Anti-phage Defense systems using PADLOC

As more defense systems have been discovered recently (Tesson et al., 2022), we used PADLOC to detect most discovered defense systems and gave a systemic description of pan-immune systems of 10 species selected before.

The PADLOC database contains 57 families of defense systems and one self-defined family (DMS family). The DMS family is a collection of ambiguously classified subtypes of several defense systems, including disarm, Dpd, BREX, GAO 29, phophorothioation and RM systems. The DMS family is created to allow a flexible classification because some defense systems comprise the same domains with other proteins involved in other functions (Payne et al., 2021).

PADLOC can localize the defense genes of each defense system on prokaryotic genomes. A visualized example of PADLOC output can be seen from Figure 3.13.D. The example is from *P. aeruginosa* strain Carb01 63 (accession: NG_CIP1131.1), which encodes 12 defense families and 24 defense systems on its chromosome. We can see five distinct clusters of defense genes on its genome (Figure 3.13 D), with three clusters enriched in defense genes, supporting the defense island theory that the defense genes are located together as an island in the genome (Makarova et al., 2011).

When we compare the detection of CRISPR-Cas between PADLOC and CRISPRCasTyper, we find over 99.9% of consistency of presence of CRISPR-Cas. The main difference is that the PADLOC we used only determines CRISPR-Cas based on Cas proteins without detecting CRISPR arrays (the new version of PADLOC can provide the CRISPR array detected by CRISPRDetect (Biswas et al., 2016) as an input).

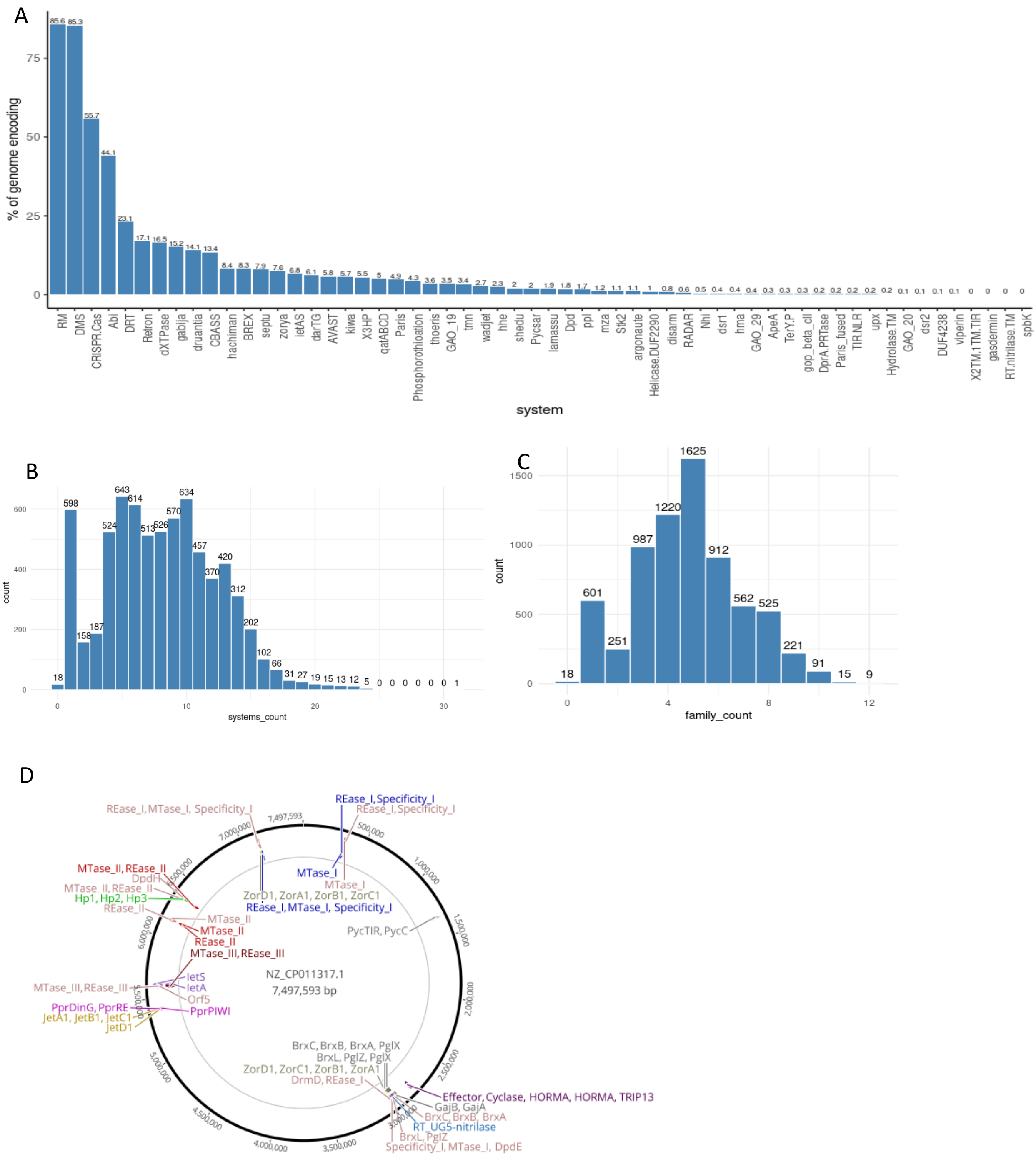


Figure 3.12 . The distribution of defense systems from 7030 genomes. A: The defense systems and the percentage of total genomes encoding them. B: The distribution of number of defense systems encoded on the genome. C. The distribution of number of defense families encoded on one genome. D: The annotation of defense system genes on one example genome of *P. aeruginosa* (ID: NG_CIP1131.1).

3.6.1 The abundance of defense systems of in bacterial genomes

Within 58 families, we detected 36 defense systems present in more than 1% of genomes. RM (85.6%), DMS (85.3%), CRISPR-Cas (56.7%), and Abi (23.1%) are the most common defense families. The abundance of the RM system is close to the estimated abundance in prokaryotic genomes (over 74%) (Oliveira et al. 2014). The distribution of the total number of defense systems and the number of defense system families is also shown in Figure 3.12. The majority of genomes encode 1-15 systems and 1-9 families (Figure 3.12 B and C). On average, about 8.1 defense systems and 4.8 families of defense systems are encoded in a genome.

In most defense families, a complete system is encoded on a single genome, but in some families, such as the RM families, on average two complete RM systems are encoded on a single genome. An extreme case is a strain of *K pneumoniae* (accession number: GCF_016864395.1) in which 31 defense systems from 12 families are detected.

3.6.2 The antiviral arsenal of bacteria varies among species

The heatmap (Figure 3.14) illustrates that the distribution of defense systems varies among species. Some anti-phage systems are widespread in most species and some anti-phage systems are species-specific. For example, RM, CRISPR-Cas, and Abi are widely distributed in most species. Systems such as darTG (a toxin-antitoxin mechanism system) are enriched only in *M. tuberculosis*.

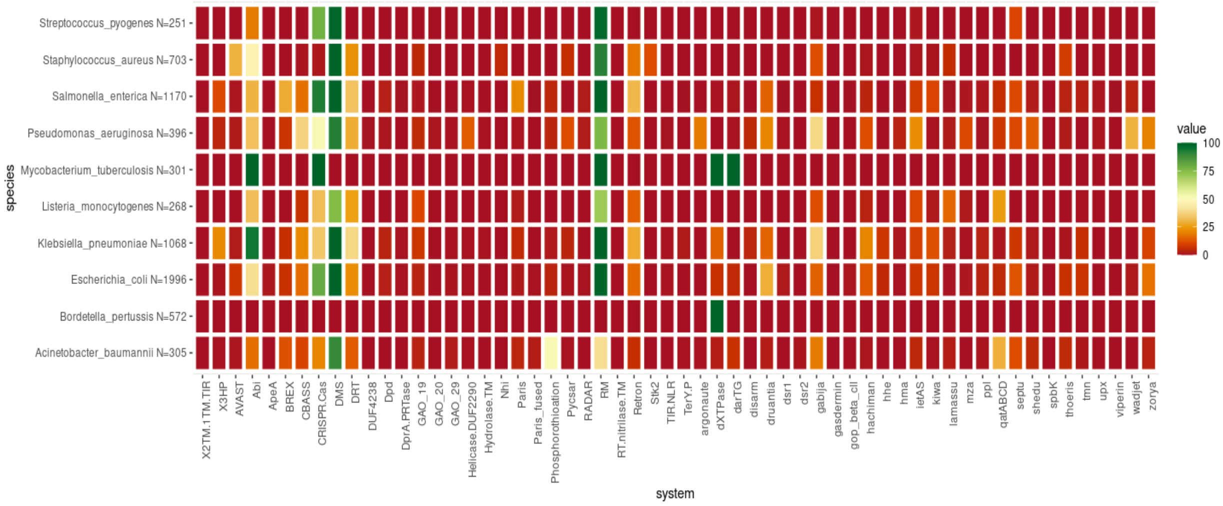


Figure 3.13 The distribution of defense systems in 10 species. The color represents the percentage of genomes in that species that encodes that system.

The diversity and abundance of defense systems also vary among species (Figure A.6), *P. aeruginosa* (average families= 6.2) and *K. pneumoniae* (average families= 6), *S. enterica* (average families= 5.5), *E. coli* (average families= 5.5) have more encoded families but *S. pyogenes* (average families= 3.1) and *L. monocytogenes* (average families= 3.4) and *S. aureus* (average families= 3) have fewer defense families. *B. pertussis* (average families= 1) has the least number of families. The species with a large diversity of defense systems also have a large number of defense systems. For example, *S. enterica*, *E. coli* and *K. pneumoniae* have on average more than 10 defense systems encoded in their genomes. In contrast, *S. pyogenes*, *L. monocytogenes, S. aureus* and *B. pertussis* have fewer than 5 systems equipped. Only one defense system is detected in *B. pertussis* suggests more defense systems await discovery.

Combining with the distribution of genome size (Figure 3.1), we found the species with large genome size have high diversity and abundance of defense systems, such as *P. aeruginosa*, *K. pneumoniae*, *S. enterica* and *E. coli* (genome size > 4.4Mb). And the species with small genome size such as *B. pertussis*, *S. pyogenes* and *L. monocytogenes* (genome size < 3.2 Mb) have low diversity and abundance of defense systems.

The intraspecific antiviral arsenal may be quite variable in some species, but in some species it may be uniform. In species such as *P. aeruginosa*, there are many families of defense systems, and strains can form different combinations of defense systems, such as strain Carb01 63 (GCF_000981825.1), which encodes 24 systems from 12 families, and strain DVT413 (GCF_013343475.1), which encodes only one Abi system. Surprisingly, two species (*B. pertussis* and *M. tuberculosis*) have a fairly uniform combination of defense systems. All *B. pertussis* strains encode only dxtTPase and almost all *M. tuberculosis* strains encode five systems (dxtPase, CRISPRCas, darTG, RM, and Abi). Both strains also have extremely low genome size diversity and high genome similarity, suggesting that fewer HGT events have occurred in both species and that all strains share a close phylogenetic relationship.

Strains encode more systems as they have more chances to survive under the invasion of phage but have a competitive disadvantage when the enemies are absent (Bernheim & Sorek, 2020) Even within a species, different strains may encode different combinations of defense systems. According to the pan-immune model, species share defense systems as community resources and defense genes are considered accessory genes (Bernheim & Sorek, 2020). The presence of more defense systems in the community can increase the likelihood that some individuals can be protected from MGEs. However, the environment can also influence the distribution of defense systems, such as time, density, the type of MGEs existing in the environments (Rocha & Bikard, 2022). As we discussed before, viral diversity and temperature can influence the presence of CRISPR-Cas (Lan et al., 2022).

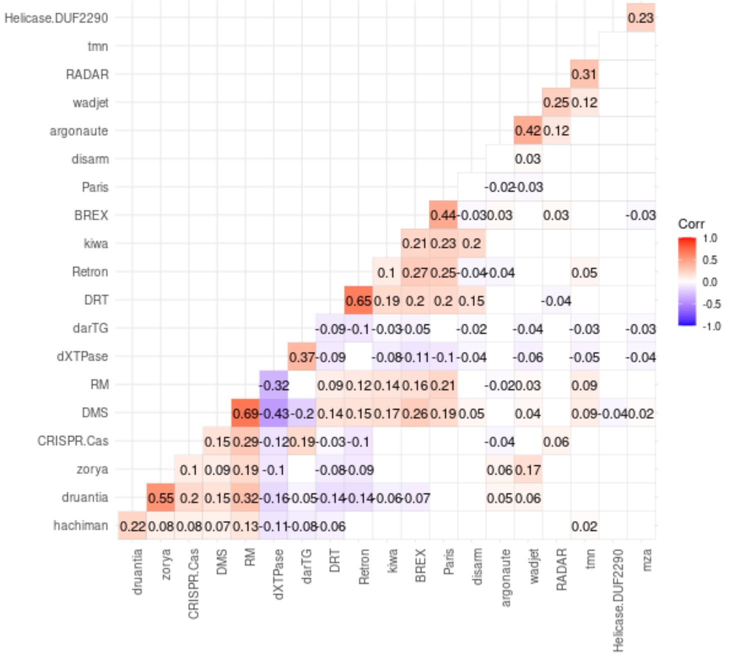
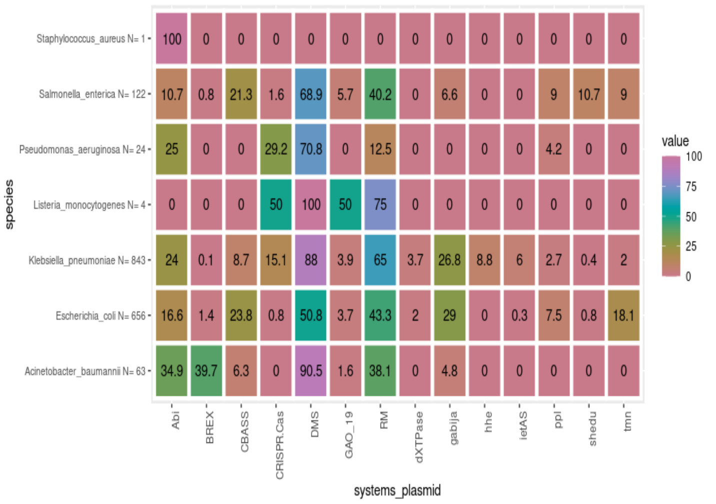


Figure 3.14 The correlation of defense systems (Kendall rank correlation test). The systems are selected if they have at least one system with correlation value larger than 0.2. The non-significant pair is set to blank.

To understand the compatibility of some defense systems, the correlation of some defense systems is plotted (Figure 3.14). Some defense systems are highly positively correlated. The high positive correlation of defense systems indicates that they can be located together, e.g., in defense islands. For example, Darian and Zorya are highly correlated because they were discovered together in defense islands of genomes (Doron et al., 2018). DRT (defense-associated reverse transcriptase) is also highly correlated with Retron (a genetic element consisting of reverse transcriptase (RT) and a non-coding RNA (ncRNA)). The high correlation may due to close properties of these two systems. The negative correlation between species may indicate the incompatibility of some defense systems. And dxtPase is negatively associated with most defense systems because it is enriched only in *B. pertussis*, a species that has only a single defense system.



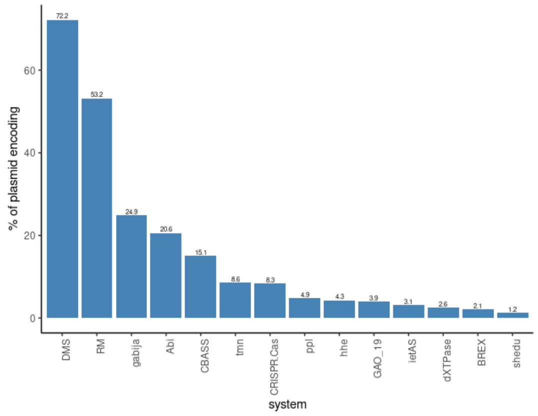


Figure 3.15 The distribution of plasmid-encoded defense systems at species level. The color and values represent that percentage of certain defense systems for each species. the percentage is calculated as (number of individuals with that defense systems encoded on the plasmid) / (total number of individuals with at least one detected plasmid-encoded defense systems). B: The frequency of each system detected.

The abundance of several plasmid-encoded defense systems is also investigated. 1650 individuals have the plasmids with defense systems encoded. The DMS, RM, Gabija and Abi are the most common systems occurring in the plasmids. Most systems are detected on *K. pneumoniae* and *E. coli*, which have the largest number of defense system encoding plasmid. These two species have the most diversity of defense systems. The presence of defense systems on the plasmids provides the evidence that the defense systems can be transferred via HGT.

3.7 The relationship between CRISPR-Cas, genome size and defense systems.

Next, we explore the factors which may influence the abundance and diversity of defense systems, such as CRISPR-Cas and genome size.

3.7.1 The relationship between genome size and defense genes

The genome size is assumed to play a critical role in the abundance and diversity of anti-phage systems. Indeed, genome size is associated with more abundant (Kendall correlation: 0.62, p<2.2e-16) (Figure A.6) and diverse defense systems (Kendall correlation 0.58, p< 2.2e-16) overall. Furthermore, we investigate if this association also occurs at the intraspecific level (Figure 3.16). Except for *B. pertussis* and *M. tuberculosis*, which both have almost unified defense systems equipped among different strains, other species show a significant positive association between the genome size and diversity and abundance of defense systems.

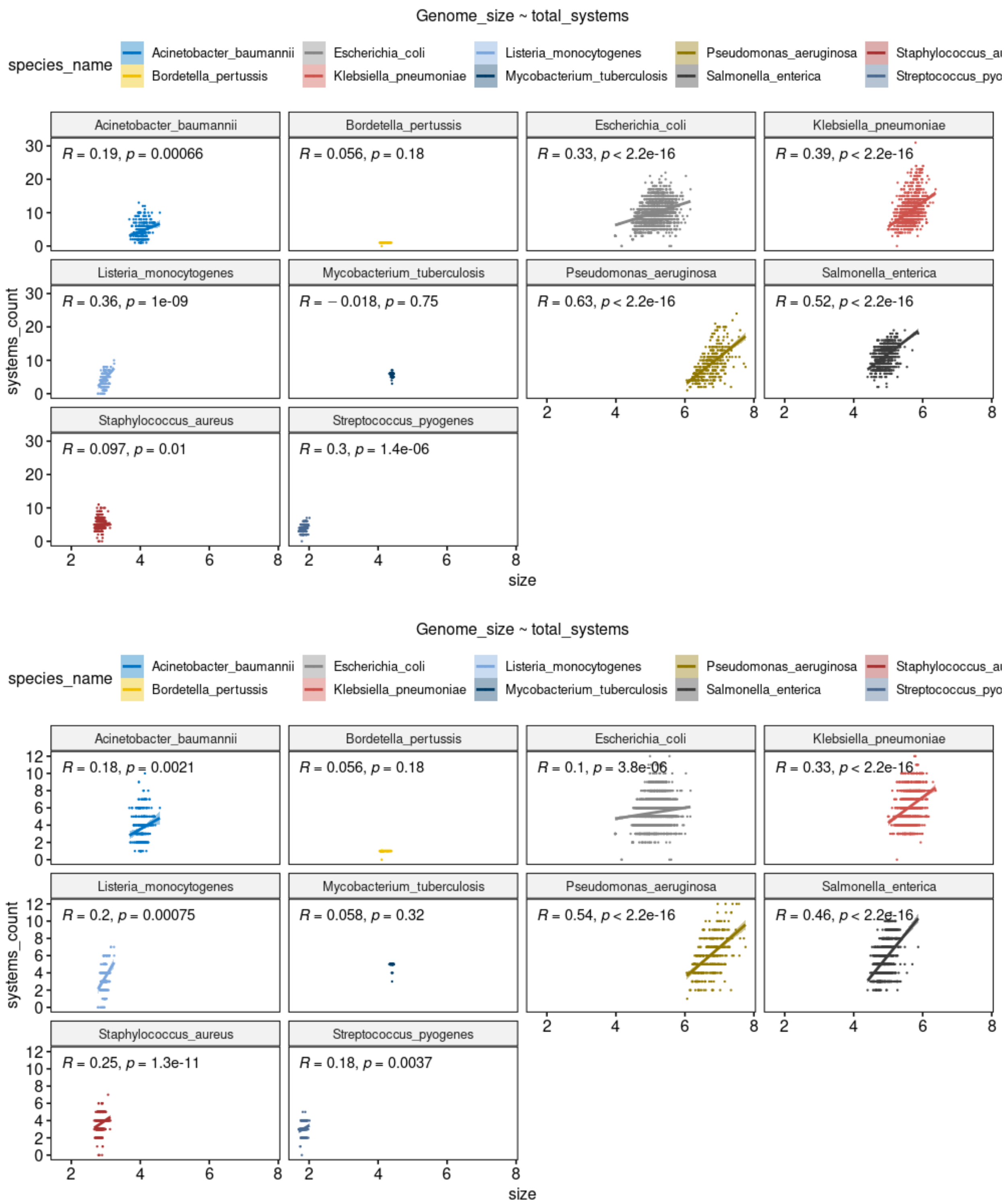


Figure 3.16 The genome size vs the number of families and systems. Upper: The Kendall test for the defense systems counts and genome size in ten species. Lower: The Kendall rank test for the defense families counts and genome size in ten species.

The proportion of all defense genes in the total genome is calculated to show what percentage of the genome is defense genes. Overall, the average percentage for 7030 genomes is 0.7%. The largest proportion (2.9%) is held by a *K. pneumoniae* strain with the greatest abundance and diversity of defense systems. In addition, the proportion also varies between species. Two species, *S. pyogenes* (1.15%) and *S. enterica* (1.05%), have the largest average percentages. *M. tuberculosis* (0.365%) and *B. pertussis* (0.028%) have the smallest mean percentages.

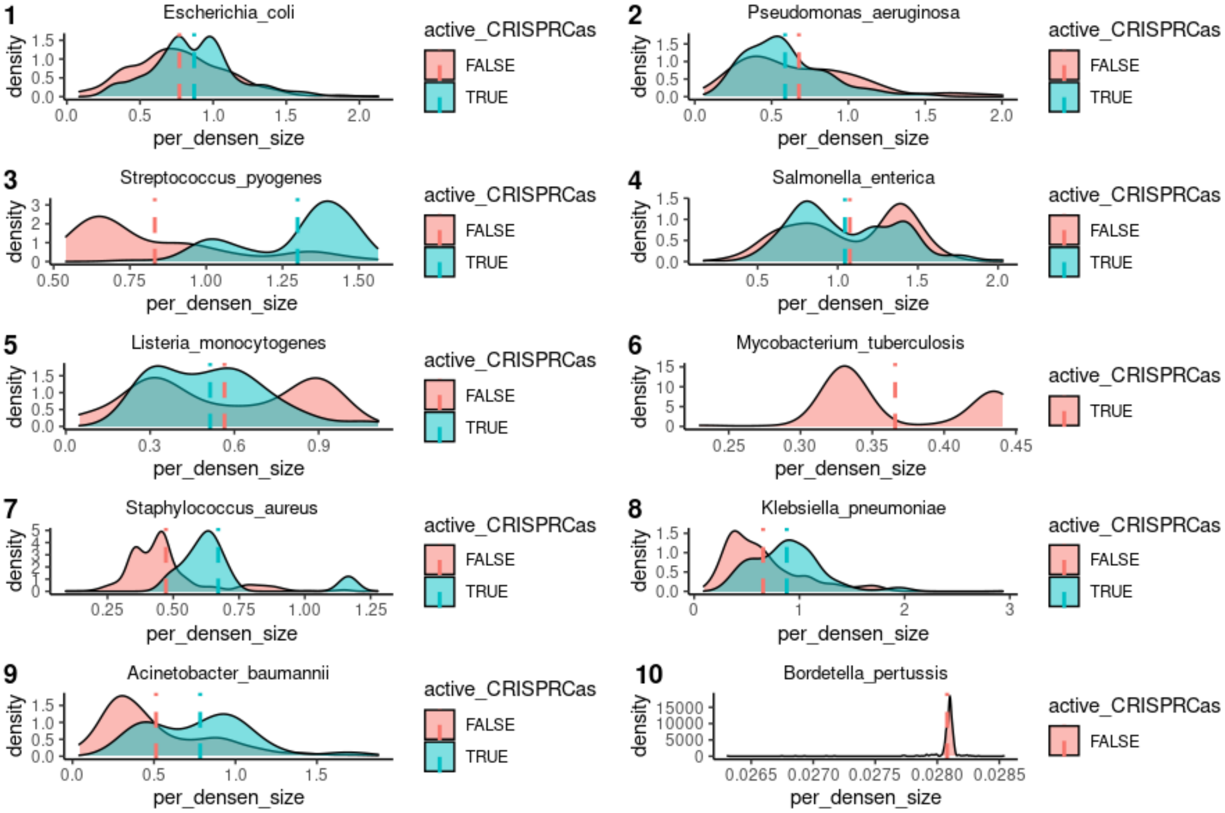


Figure 3.17 The distribution of the proportion of all defense genes in the total genome impacted by the presence of an active CRISPR-Cas systems.

And association between the presence of CRISPR-Cas and the defense genes percentage is also investigated (Figure 3.17). In *A. baumannii*, *K. pneumoniaa*, *S. aureus*, *S. pyogenes*, *E. coli,* the presence of active CRISPR-Cas (all Wilcoxon p values are smaller than 0.005) is associated with larger percentages. In *S. enterica* (Wilcoxon p= 0.14) and *P. aeruginosa*, *L. monocytogenes* (Wilcoxon p-value =0.46), the associations are not significant (Wilcoxon p= 0.14) (Figure A.9).

The association between active CRISPR-Cas and diversity and abundance of defense systems is also investigated. For the species having a negative association between the active CRISPR-Cas and genome size, such as *P. aeruginosa* and *L. monocytogenes,* the presence of active CRISPR-Cas is also associated with lower number of defense systems. However, it is not associated with diversity (Figure A.10). For the species with the reverse association between the active CRISPR-Cas presence and genome size, such as *K. pneumoniae*, *A. baumannii*, *S. pyogenes*, *E. coli*, they also have the opposite pattern that the presence of active CRISPR-Cas is associated with larger abundance and diversity of defense systems (Figure A.10). Surprisingly, the *S. enterica* that have a positive association between the presence of active CRISPR-Cas and genome size, showing a negative association between the CRISPR-Cas and abundance (Wilcoxon p-value 3.9e-06) and diversity (3.8e-06) of defense systems (Figure A.10).

3.7.2 The relationship between CRISPR-Cas and other defense systems

In *P. aeruginosa*, the CRISPR-Cas can potentially restrict the HGT, from the pan-immune systems model, the defense genes can be horizontal transmitted between closely related strains. The CRISPR-Cas may potentially restrict the transmission of defense genes. The correlation (Kendall rank test) between active CRISPR-Cas and other defense systems of seven species was calculated (Figure 3.18).

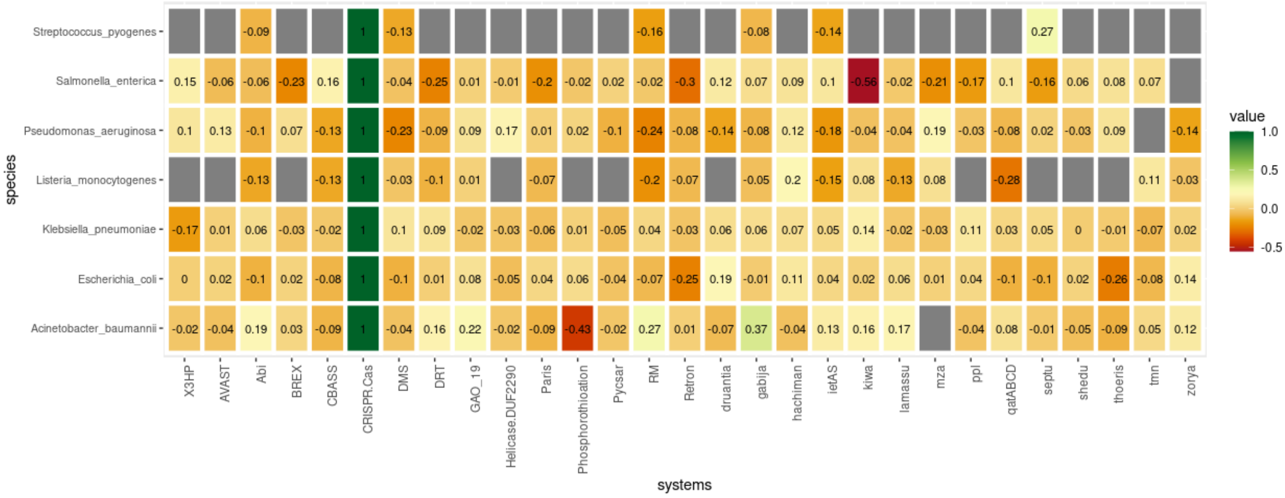


Figure 3.18 The correlation (kendall rank test) of active CRISPR-Cas and other defense systems of seven species. The defense systems occur less than 3 species are removed. And *B. pertussis* and *S. aureus* are not considered in this analysis as the percentage of encoding CRISPR-Cas is low. *M. tuberculosis* is removed as most strains have the same defense systems.

The correlation between CRISPR-Cas and other defense systems is different in different species, we cannot observe one defense system that is negatively associated with CRISPR-Cas in all species.

Some defense systems, such as the RM system and CBASS show a negative association with the CRISPR-Cas in most species. Most of the systems show a weak correlation with CRISPR-Cas, and the correlations vary among species. Some systems can only show a strong negative association in certain species, for example in *S. enterica*, the CRISPR-Cas has a strong negative association with kiwa (Kendall rank test: p-value <2e-16, correlation: -0.6). However, the association becomes quite weak, if the inhibition activity of Acrs is not considered (Kendall rank correlation = 0.031 and p = 0.3), suggesting the CRISPR-Cas may directly or indirectly restrict the presence of kiwa and Acrs may mediate the restriction through inhibiting CRISPR-Cas activity. The mechanism of kiwa is not well known ,and it is membrane-associated, suggesting it may involve in a membrane depolarization Abi like defense pathways. (Doron et al., 2018). Phosphorothioation is a family of phosphorothioate-based defense systems, including SspABCD, SspE, SspFGH and Dnd systems(He et al., 2015; S. Wang et al., 2021). The Phosphorothioation defense family is strongly negatively associated with CRISPR-Cas (Kendall rank correlation=-0.43, p=3e-14) in *A. baumannii*.

In the majority of species, most systems have a weak negative association with CRISPR-Cas and the positive strong association is less common. But in *A. baumannii* we observe several positive associations, such as Gabija (Kendall test: correlation=0.37, p=7e-11) and RM (Kendall test: correlation=0.27, p= 7e-07). 1/3 (16/48) of *A. baumannii* with active CRISPR-Cas both have Gabija and qatABCD. After investigating the loci, they do not sit closely in the genomes.

3.7.3 Can CRISPR-Cas directly target other defense systems?

To determine if the CRISPR-Cas can restrict other defense systems directly. We selected several candidate defense systems based on the correlation plot and investigated if the spacers from the complete CRISPR-Cas system can target the defense systems. The defense genes were collected from different sources and some defense genes (RM,Abi, Dnd, Zorya, LAMASSU, kiwa) are abundant and some defense genes are not (Retron, qatABCD, ietAs).

Table 3.3 The number of spacers of each targeted system. The number of spacers of each targeted system. N is the number of genomes. Per of CRISPR+ is the percentage of genomes with CRISPR-Cas that have spacers can target on the defense systems. The total spacers are the total spacers of each species. The diversity is the diversity of defense genes.

species	Per of CRISPRCas+	total_spacers	ABI	LAMASSU	ZORYA	DND	RM	Diversity
Streptococcus_pyogenes N= 48	25.95%	366	0	0	15	0	52	9
Mycobacterium_tuberculosis N= 203	67.22%	8218	0	0	0	0	203	1
Escherichia_coli N= 230	14.80%	4084	0	0	32	0	202	13
Acinetobacter_baumannii N= 31	50.82%	2211	0	0	7	0	29	8
Listeria_monocytogenes N= 39	45.88%	1558	0	0	9	2	62	20
Salmonella_enterica N= 159	14.65%	8769	0	1	23	0	170	37
Pseudomonas_aeruginosa N= 123	60.89%	4821	27	0	20	5	181	35
Klebsiella_pneumoniae N= 84	24.56%	3435	0	0	6	0	96	13

It is not surprising that RM is the most common target as the RM system is the most common defense systems present in bacterial genomes and also most studied (Oliveira, Touchon et al. 2014). Some new discovered systems (such as qatABCD) have less study and limited defense genes. The Zorya is the second most common gene, the Abi, LAMASSU and DND defense genes are only found to be targeted by few spacers and no hits are found for other defense genes.

P. aeruginosa (correlation -0.24 and p =7e-08) and *L. monocytogenes* (correlation -0.2 and p=3e-04) have a significant negative relationship between RM systems. And others show no significant relationship and *A. baumannii* shows a positive relationship. The strains with defense genes targeting spacers are more prevalent (60.89%, 45.88%) and highly diverse (35/123, 20/39) in *P. aeruginosa* and *L. monocytogenes,* which may explain why only *P. aeruginosa* and *L. monocytogenes* are negatively associated with RM.

However, the evidence is not strong enough. Zorya is the second largest target of spacers; however, the negative association between Zorya and CRISPR-Cas is observed only in *P. aeruginosa*. And *A.*

baumannii has a high percentage of genomes targeting defense genes, but the CRISPR-Cas has a weak positive association (Kendall rank correlation: 0.27) with RM.

4 Conclusions and perspectives

In this dissertation, a large-scale comparative analysis of 7030 genomes of the top 10 bacterial species with the most complete genomes was performed. A pipeline was built starting with data collection, CRISPR-Cas and anti-CRISPR annotation, spacer targets identification, detection of other anti-phage defense systems and ending with statistical analysis. The developed pipelines can be applied to other strains and species from the updated database in the future.

The genome size of *P. aeruginosa* exhibits a bimodal distribution (Figure 3.1), suggesting that there is a force causing the bimodal distribution pattern. Wheatley & MacLean pointed out a relationship between the presence of CRISPR-Cas and larger genomes in *P. aeruginosa* (Wheatley & MacLean, 2021). And the author suggested that CRISPR-Cas may restrict HGT and prevent the acquisition of new genetic elements and finally restrict the genome expansion. Here, we investigated whether the bimodal distribution of genome size is prevalent in other bacterial species and whether the presence of CRISPR-Cas is also associated with a smaller genome in other species.

We observed the distinct bimodal distribution of genome size only in *P. aeruginosa*. And the association between the presence of CRISPR-Cas and a smaller genome is only in *P. aeruginosa* and *L. monocytogenes*. For three species in the Enterobacteriaceae family (Figure 2.1), (*S. enterica*, *K. pneumoniae*, and *E. coli)*, the pattern is reversed, as the presence of CRISPR-Cas is associated with a larger genome. No association was found in the remaining species.

To further investigate why the patterns differ, we investigate the influence of subtypes of CRISPR-Cas (Figure 3.7 and Figure 3.8) as well as anti-CRISPR (Figure 3.9) and spacer targets (Figure 3.11) of genomes with CRISPR-Cas. We found that some subtypes are associated with smaller genomes in some species, such as the type II- A, which is associated with smaller size in *S. pyogenes* and *L. monocytogenes*. Extending the analysis to more species with the type II - A may further reveal whether the type II -A is associated with smaller genome size in other species.

Since Acrs have been found to enhance HGT by inhibiting CRISPR-Cas activity (Mahendra et al., 2020), we detected and investigated the impacts of Acrs on genome size (Figure 3.9). After using Acrs to correct CRISPR-Cas activity, the relationship between CRISPR-Cas and genome size does not change in any species. Interestingly, we found that in most species, the presence of anti-CRISPR is associated with larger genome size, both in the CRISPR-Cas+ (Figure 3.9) and CRISPR-Cas- (Figure 3.10) group. The association between the presence of Acrs and larger genomes in CRISPR-Cas+ group suggests that the Acrs might enhance HGT by inhibiting CRISPR-Cas activity. However, the association between the presence of Acrs and larger genome in CRISPR-Cas- suggests that the presence of Acrs may be caused by high promiscuity of MGEs. However, data about Acrs of CRISPR-Cas groups of was lacking in some species due to the limitation of AcrFinder. To our knowledge, AcrFinder is the only bioinformatics tool that provides a standalone software for large-scale analysis of Acr genes. And AcrFinder can infer Acrs only on genomes with CRISPR-Cas systems via GBA routes. The database of AcrFinder is not updated thus the homology-based route can only detect limited Acrs. The tool with updated database can improve our analysis.

Next, we investigated the targets of spacers (Table 3.2). We show that only *P. aeruginosa* has a larger percentage of genomes with plasmid-targeting (93.85%) and ICEs-targeting spacers (81.9%), suggesting that a high percentage of genomes with CRISPR-Cas in *P. aeruginosa* can restrict the HGT. The much higher percentage of genomes with ICEs-targeting indicate HGT is restricted more in *P. aeruginosa* than in other species. And we conclude that high percentage of genomes with ICEs-targeting spacers can explain why *P. aeruginosa* is the only species with bimodal distribution of genome size. And the CRISPR-Cas is supposed to be a strong force to drive the formation of bimodal distribution of genome size.

L. monocytogenes also has a high percentage of plasmids-targeting spacer and a substantial percentage of ICEs-targeting spacers (25.42%) and a shorter CRISPR array may reflect the CRISPR-Cas can also put constraints on the genome size but less than *P. aeruginosa*. As the pipeline has been built, more species can be analyzed in the future to determine if the abundance of plasmid-targeting and ICEs-targeting spacers is associated with smaller genomes.

The bacterial immune system of a species is the sum of all immune systems that can be horizontally transmitted in that species. We performed a large-scale analysis of bacterial anti-phage defense systems using genomic data with the latest version of PADLOC and provided a quantitative description of the pan-immune systems of 10 species. On average, a genome contains 8.1 defense systems and 4.8 families, and 0.7% of genomes are defense genes. The abundance and diversity of defense systems varies at the species level. And the lack of detected defense systems in some species, such as *B. pertussis*, suggests that some anti-phage systems have not yet been discovered. And the study of the co-occurrence of some defense systems may indicate the compatibility of some defense systems. The highly negatively correlated defense systems can be validated experimentally to explore their mechanisms. We also investigated the factors can influence the abundance and diversity of defense systems and found the larger genome size is associated with more defense families and defense systems in most of species.

Some defense systems have been detected on mobile genetic elements (Bernheim & Sorek, 2020; Pinilla-Redondo et al., 2022), we also investigate the abundance of some defense systems on plasmids, providing some evidence that the defense systems can be transmitted via HGT. The evidence can be stronger if we also measure the conjugative transmissibility of plasmids with defense systems. Pinilla-Redondo *et al*. found that plasmids with CRISPR-Cas have higher conjugative transmissibility(Pinilla-Redondo et al., 2022). The defense systems on the plasmids with high conjugative transmissibility are more likely to be horizontal transferred.

As the CRISPR-Cas can target and cleave the MGEs (Mahendra et al., 2020; Marraffini & Sontheimer, 2008; Wheatley & MacLean, 2021) and it can potentially act as a barrier to the transmission of defense systems, we investigated if the presence of active CRISPR-Cas can influence the bacterial pan-immune systems. We found the active CRISPR- Cas system is associated with lower abundance of other defense systems and observed some defense systems are negative correlated with active

CRISPR-Cas in some species. We also investigated if the CRISPR-Cas can direct target other defense systems and found the RM is common target of the CRISPR-Cas in many species.

The study can be further improved by adding some analysis. We hypothesize that CRISPR-Cas may limit HGT, and several studies directly measure HGT rates with different strategies (Gophna et al., 2015; Lehtinen et al., 2020). Measuring of HGT rates may provide more related results to see if CRISPR-Cas can directly affect HGT. The phylogenetic analysis can be performed to know the population structure of each species which may find difference of patterns in certain species clades. Environment factors also be included as the ecological factors can influence the distribution of bacterial defense systems. For example, temperature, oxygen, viral density, and viral abundance can influence the CRISPR-Cas systems (Lan et al., 2022; Weissman et al., 2019). The ecological data can provide more information about the interaction between bacterial pan-immune systems and environments.

Overall, our works can improve the comprehensive understanding of the bacterial pan-immune systems and the evolutionary role of CRISPR-Cas on bacterial genome size .

Reference

Abby, S. S., Neron, B., Menager, H., Touchon, M., & Rocha, E. P. (2014). MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. *PLoS One*, *9*(10), e110726. https://doi.org/10.1371/journal.pone.0110726

Al-Shayeb, B., Sachdeva, R., Chen, L. X., Ward, F., Munk, P., Devoto, A., Castelle, C. J., Olm, M. R., Bouma-Gregson, K., Amano, Y., He, C., Meheust, R., Brooks, B., Thomas, A., Lavy, A., Matheus-Carnevali, P., Sun, C., Goltsman, D. S. A., Borton, M. A., . . . Banfield, J. F. (2020). Clades of huge phages from across Earth's ecosystems. *Nature*, *578*(7795), 425-431. https://doi.org/10.1038/s41586-020-2007-4

Benmayor, R., Hodgson, D. J., Perron, G. G., & Buckling, A. (2009). Host mixing and disease emergence. *Curr Biol*, *19*(9), 764-767. https://doi.org/10.1016/j.cub.2009.03.023

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

Bernheim, A., Millman, A., Ofir, G., Meitav, G., Avraham, C., Shomar, H., Rosenberg, M. M., Tal, N., Melamed, S., Amitai, G., & Sorek, R. (2021). Prokaryotic viperins produce diverse antiviral molecules. *Nature*, *589*(7840), 120-124. https://doi.org/10.1038/s41586-020-2762-2

Bernheim, A., & Sorek, R. (2020). The pan-immune system of bacteria: antiviral defence as a community resource. *Nat Rev Microbiol*, *18*(2), 113-119. https://doi.org/10.1038/s41579-019-0278-2

Bikard, D., Hatoum-Aslan, A., Mucida, D., & Marraffini, L. A. (2012). CRISPR interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. *Cell Host Microbe*, *12*(2), 177-186. https://doi.org/10.1016/j.chom.2012.06.003

Birkholz, N., Jackson, S. A., Fagerlund, R. D., & Fineran, P. C. (2022). A mobile restriction-modification system provides phage defence and resolves an epigenetic conflict with an antagonistic endonuclease. *Nucleic Acids Res*, *50*(6), 3348-3361. https://doi.org/10.1093/nar/gkac147

Biswas, A., Staals, R. H., Morales, S. E., Fineran, P. C., & Brown, C. M. (2016). CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics*, *17*, 356. https://doi.org/10.1186/s12864-016-2627-0

Bondy-Denomy, J., Pawluk, A., Maxwell, K. L., & Davidson, A. R. (2013). Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature*, *493*(7432), 429-432. https://doi.org/10.1038/nature11723

Bondy-Denomy, J., Qian, J., Westra, E. R., Buckling, A., Guttman, D. S., Davidson, A. R., & Maxwell, K. L. (2016). Prophages mediate defense against phage infection through diverse mechanisms. *ISME J*, *10*(12), 2854-2866. https://doi.org/10.1038/ismej.2016.79

Braz, V. S., Melchior, K., & Moreira, C. G. (2020). Escherichia coli as a Multifaceted Pathogenic and Versatile Bacterium. *Front Cell Infect Microbiol*, *10*, 548492. https://doi.org/10.3389/fcimb.2020.548492

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat Methods*, *12*(1), 59-60. https://doi.org/10.1038/nmeth.3176

Castro, S. A., & Dorfmueller, H. C. (2021). A brief review on Group A Streptococcus pathogenesis and vaccine development. *R Soc Open Sci*, *8*(3), 201991. https://doi.org/10.1098/rsos.201991

Chai, Q., Zhang, Y., & Liu, C. H. (2018). Mycobacterium tuberculosis: An Adaptable Pathogen Associated With Multiple Human Diseases. *Front Cell Infect Microbiol*, *8*, 158. https://doi.org/10.3389/fcimb.2018.00158

Chevallereau, A., Pons, B. J., van Houte, S., & Westra, E. R. (2022). Interactions between bacterial and phage communities in natural environments. *Nat Rev Microbiol*, *20*(1), 49-62. https://doi.org/10.1038/s41579-021-00602-y

Chopin, M. C., Chopin, A., & Bidnenko, E. (2005). Phage abortive infection in lactococci: variations on a theme. *Curr Opin Microbiol*, *8*(4), 473-479. https://doi.org/10.1016/j.mib.2005.06.006

Clokie, M. R., Millard, A. D., Letarov, A. V., & Heaphy, S. (2011). Phages in nature. *Bacteriophage*, *1*(1), 31-45. https://doi.org/10.4161/bact.1.1.14942

Cohen, D., Melamed, S., Millman, A., Shulman, G., Oppenheimer-Shaanan, Y., Kacen, A., Doron, S., Amitai, G., & Sorek, R. (2019). Cyclic GMP-AMP signalling protects bacteria against viral infection. *Nature*, *574*(7780), 691-695. https://doi.org/10.1038/s41586-019-1605-5

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Neron, B., Rocha, E. P. C., Vergnaud, G., Gautheret, D., & Pourcel, C. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. *Nucleic Acids Res*, *46*(W1), W246-W251. https://doi.org/10.1093/nar/gky425

Cruz-Lopez, E. A., Rivera, G., Cruz-Hernandez, M. A., Martinez-Vazquez, A. V., Castro-Escarpulli, G., Flores-Magallon, R., Vazquez, K., Cruz-Pulido, W. L., & Bocanegra-Garcia, V. (2021). Identification and Characterization of the CRISPR/Cas System in Staphylococcus aureus Strains From Diverse Sources. *Front Microbiol*, *12*, 656996. https://doi.org/10.3389/fmicb.2021.656996

de Jonge, P. A., Nobrega, F. L., Brouns, S. J. J., & Dutilh, B. E. (2019). Molecular and Evolutionary Determinants of Bacteriophage Host Range. *Trends Microbiol*, *27*(1), 51-63. https://doi.org/10.1016/j.tim.2018.08.006

Dion, M. B., Oechslin, F., & Moineau, S. (2020). Phage diversity, genomics and phylogeny. *Nat Rev Microbiol*, *18*(3), 125-138. https://doi.org/10.1038/s41579-019-0311-5

Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G., & Sorek, R. (2018). Systematic discovery of antiphage defense systems in the microbial pangenome. *Science*, *359*(6379). https://doi.org/10.1126/science.aar4120

Du Toit, A. (2018). Viral infection: CRISPR-Cas enhances HGT by transduction. *Nat Rev Microbiol*, *16*(4), 186. https://doi.org/10.1038/nrmicro.2018.28

Dupuis, M. E., Villion, M., Magadan, A. H., & Moineau, S. (2013). CRISPR-Cas and restriction-modification systems are compatible and increase phage resistance. *Nat Commun*, *4*, 2087. https://doi.org/10.1038/ncomms3087

Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Inform*, *23*(1), 205-211. https://www.ncbi.nlm.nih.gov/pubmed/20180275

Eitzinger, S., Asif, A., Watters, K. E., Iavarone, A. T., Knott, G. J., Doudna, J. A., & Minhas, F. (2020). Machine learning predicts new anti-CRISPR proteins. *Nucleic Acids Res*, *48*(9), 4698-4708. https://doi.org/10.1093/nar/gkaa219

Gao, L., Altae-Tran, H., Bohning, F., Makarova, K. S., Segel, M., Schmid-Burgk, J. L., Koob, J., Wolf, Y. I., Koonin, E. V., & Zhang, F. (2020). Diverse enzymatic activities mediate antiviral immunity in prokaryotes. *Science*, *369*(6507), 1077-1084. https://doi.org/10.1126/science.aba0372

Gil, J. F., Mesa, V., Estrada-Ortiz, N., Lopez-Obando, M., Gomez, A., & Placido, J. (2021). Viruses in Extreme Environments, Current Overview, and Biotechnological Potential. *Viruses*, *13*(1). https://doi.org/10.3390/v13010081

Gophna, U., Kristensen, D. M., Wolf, Y. I., Popa, O., Drevet, C., & Koonin, E. V. (2015). No evidence of inhibition of horizontal gene transfer by CRISPR-Cas on evolutionary timescales. *ISME J*, *9*(9), 2021-2027. https://doi.org/10.1038/ismej.2015.20

Gussow, A. B., Park, A. E., Borges, A. L., Shmakov, S. A., Makarova, K. S., Wolf, Y. I., Bondy-Denomy, J., & Koonin, E. V. (2020). Machine-learning approach expands the repertoire of anti-CRISPR protein families. *Nat Commun*, *11*(1), 3784. https://doi.org/10.1038/s41467-020-17652-0

Gweon, H. S., Bailey, M. J., & Read, D. S. (2017). Assessment of the bimodality in the distribution of bacterial genome sizes. *ISME J*, *11*(3), 821-824. https://doi.org/10.1038/ismej.2016.142

Hampton, H. G., Watson, B. N. J., & Fineran, P. C. (2020). The arms race between bacteria and their phage foes. *Nature*, *577*(7790), 327-336. https://doi.org/10.1038/s41586-019-1894-8

He, W., Huang, T., Tang, Y., Liu, Y., Wu, X., Chen, S., Chan, W., Wang, Y., Liu, X., Chen, S., & Wang, L. (2015). Regulation of DNA phosphorothioate modification in Salmonella enterica by DndB. *Sci Rep*, *5*, 12368. https://doi.org/10.1038/srep12368

Howard-Varona, C., Hargreaves, K. R., Abedon, S. T., & Sullivan, M. B. (2017). Lysogeny in nature: mechanisms, impact and ecology of temperate phages. *ISME J*, *11*(7), 1511-1520. https://doi.org/10.1038/ismej.2017.16

Hyatt, D., Chen, G. L., Locascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, *11*, 119. https://doi.org/10.1186/1471-2105-11-119

Jajere, S. M. (2019). A review of Salmonella enterica with particular focus on the pathogenicity and virulence factors, host specificity and antimicrobial resistance including multidrug resistance. *Vet World*, *12*(4), 504-521. https://doi.org/10.14202/vetworld.2019.504-521

Jaskolska, M., Adams, D. W., & Blokesch, M. (2022). Two defence systems eliminate plasmids from seventh pandemic Vibrio cholerae. *Nature*, *604*(7905), 323-329. https://doi.org/10.1038/s41586-022-04546-y

Kamruzzaman, M., & Iredell, J. R. (2019). CRISPR-Cas System in Antibiotic Resistance Plasmids in Klebsiella pneumoniae. *Front Microbiol*, *10*, 2934. https://doi.org/10.3389/fmicb.2019.02934

Kilgore, P. E., Salim, A. M., Zervos, M. J., & Schmitt, H. J. (2016). Pertussis: Microbiology, Disease, Treatment, and Prevention. *Clin Microbiol Rev*, *29*(3), 449-486. https://doi.org/10.1128/CMR.00083-15

Koonin, E. V., Makarova, K. S., Wolf, Y. I., & Krupovic, M. (2020). Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. *Nat Rev Genet*, *21*(2), 119-131. https://doi.org/10.1038/s41576-019-0172-9

Kronheim, S., Daniel-Ivad, M., Duan, Z., Hwang, S., Wong, A. I., Mantel, I., Nodwell, J. R., & Maxwell, K. L. (2018). A chemical defence against phage infection. *Nature*, *564*(7735), 283-286. https://doi.org/10.1038/s41586-018-0767-x

Lan, X. R., Liu, Z. L., & Niu, D. K. (2022). Precipitous Increase of Bacterial CRISPR-Cas Abundance at Around 45 degrees C. *Front Microbiol*, *13*, 773114. https://doi.org/10.3389/fmicb.2022.773114

Levin, B. R., Moineau, S., Bushman, M., & Barrangou, R. (2013). The population and evolutionary dynamics of phage and bacteria with CRISPR-mediated immunity. *PLoS Genet*, *9*(3), e1003312. https://doi.org/10.1371/journal.pgen.1003312

Li, H. Y., Kao, C. Y., Lin, W. H., Zheng, P. X., Yan, J. J., Wang, M. C., Teng, C. H., Tseng, C. C., & Wu, J. J. (2018). Characterization of CRISPR-Cas Systems in Clinical Klebsiella pneumoniae Isolates Uncovers Its Potential Association With Antibiotic Susceptibility. *Front Microbiol*, *9*, 1595. https://doi.org/10.3389/fmicb.2018.01595

Liu, M., Li, X., Xie, Y., Bi, D., Sun, J., Li, J., Tai, C., Deng, Z., & Ou, H. Y. (2019). ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. *Nucleic Acids Res*, *47*(D1), D660-D665. https://doi.org/10.1093/nar/gky1123

Loenen, W. A., Dryden, D. T., Raleigh, E. A., Wilson, G. G., & Murray, N. E. (2014). Highlights of the DNA cutters: a short history of the restriction enzymes. *Nucleic Acids Res*, *42*(1), 3-19. https://doi.org/10.1093/nar/gkt990

Loh, B., Kuhn, A., & Leptihn, S. (2019). The fascinating biology behind phage display: filamentous phage assembly. *Mol Microbiol*, *111*(5), 1132-1138. https://doi.org/10.1111/mmi.14187

Lopatina, A., Tal, N., & Sorek, R. (2020). Abortive Infection: Bacterial Suicide as an Antiviral Immune Strategy. *Annu Rev Virol*, *7*(1), 371-384. https://doi.org/10.1146/annurev-virology-011620-040628

Lopes-Luz, L., Mendonca, M., Bernardes Fogaca, M., Kipnis, A., Bhunia, A. K., & Buhrer-Sekula, S. (2021). Listeria monocytogenes: review of pathogenesis and virulence determinants-targeted immunological assays. *Crit Rev Microbiol*, *47*(5), 647-666. https://doi.org/10.1080/1040841X.2021.1911930

Maguin, P., Varble, A., Modell, J. W., & Marraffini, L. A. (2022). Cleavage of viral DNA by restriction endonucleases stimulates the type II CRISPR-Cas immune response. *Mol Cell*, *82*(5), 907-919 e907. https://doi.org/10.1016/j.molcel.2022.01.012

Mahendra, C., Christie, K. A., Osuna, B. A., Pinilla-Redondo, R., Kleinstiver, B. P., & Bondy-Denomy, J. (2020). Broad-spectrum anti-CRISPR proteins facilitate horizontal gene transfer. *Nat Microbiol*, *5*(4), 620-629. https://doi.org/10.1038/s41564-020-0692-2

Makarova, K. S., Wolf, Y. I., Iranzo, J., Shmakov, S. A., Alkhnbashi, O. S., Brouns, S. J. J., Charpentier, E., Cheng, D., Haft, D. H., Horvath, P., Moineau, S., Mojica, F. J. M., Scott, D., Shah, S. A., Siksnys, V., Terns, M. P., Venclovas, C., White, M. F., Yakunin, A. F., . . . Koonin, E. V. (2020). Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol*, *18*(2), 67-83. https://doi.org/10.1038/s41579-019-0299-x

Makarova, K. S., Wolf, Y. I., Snir, S., & Koonin, E. V. (2011). Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. *J Bacteriol*, *193*(21), 6039-6056. https://doi.org/10.1128/JB.05535-11

Malki, K., Kula, A., Bruder, K., Sible, E., Hatzopoulos, T., Steidel, S., Watkins, S. C., & Putonti, C. (2015). Bacteriophages isolated from Lake Michigan demonstrate broad host-range across several bacterial phyla. *Virol J*, *12*, 164. https://doi.org/10.1186/s12985-015-0395-0

Marino, N. D., Pinilla-Redondo, R., Csorgo, B., & Bondy-Denomy, J. (2020). Anti-CRISPR protein applications: natural brakes for CRISPR-Cas technologies. *Nat Methods*, *17*(5), 471-479. https://doi.org/10.1038/s41592-020-0771-6

Marraffini, L. A., & Sontheimer, E. J. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science*, *322*(5909), 1843-1845. https://doi.org/10.1126/science.1165771

Martynov, A., Severinov, K., & Ispolatov, I. (2017). Optimal number of spacers in CRISPR arrays. *PLoS Comput Biol*, *13*(12), e1005891. https://doi.org/10.1371/journal.pcbi.1005891

Medvedeva, S., Liu, Y., Koonin, E. V., Severinov, K., Prangishvili, D., & Krupovic, M. (2019). Virus-borne mini-CRISPR arrays are involved in interviral conflicts. *Nat Commun*, *10*(1), 5204. https://doi.org/10.1038/s41467-019-13205-2

Meyer, J. R., Dobias, D. T., Medina, S. J., Servilio, L., Gupta, A., & Lenski, R. E. (2016). Ecological speciation of bacteriophage lambda in allopatry and sympatry. *Science*, *354*(6317), 1301-1304. https://doi.org/10.1126/science.aai8446

Millman, A., Bernheim, A., Stokar-Avihail, A., Fedorenko, T., Voichek, M., Leavitt, A., Oppenheimer-Shaanan, Y., & Sorek, R. (2020). Bacterial Retrons Function In Anti-Phage Defense. *Cell*, *183*(6), 1551-1561 e1512. https://doi.org/10.1016/j.cell.2020.09.065

Millman, A., Melamed, S., Leavitt, A., Doron, S., Bernheim, A., Hör, J., Lopatina, A., Ofir, G., Hochhauser, D., Stokar-Avihail, A., Tal, N., Sharir, S., Voichek, M., Erez, Z., Ferrer, J. L. M., Dar, D., Kacen, A., Amitai, G., & Sorek, R. (2022). An expanding arsenal of immune systems that protect bacteria from phages. *bioRxiv*.

Mitrofanov, A., Alkhnbashi, O. S., Shmakov, S. A., Makarova, K. S., Koonin, E. V., & Backofen, R. (2021). CRISPRidentify: identification of CRISPR arrays using machine learning approach. *Nucleic Acids Res*, *49*(4), e20. https://doi.org/10.1093/nar/gkaa1158

Molder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Koster, J. (2021). Sustainable data analysis with Snakemake. *F1000Res*, *10*, 33. https://doi.org/10.12688/f1000research.29032.2

Moya-Beltran, A., Makarova, K. S., Acuna, L. G., Wolf, Y. I., Covarrubias, P. C., Shmakov, S. A., Silva, C., Tolstoy, I., Johnson, D. B., Koonin, E. V., & Quatrini, R. (2021). Evolution of Type IV CRISPR-

Cas Systems: Insights from CRISPR Loci in Integrative Conjugative Elements of Acidithiobacillia. *CRISPR J*, *4*(5), 656-672. https://doi.org/10.1089/crispr.2021.0051

Mulani, M. S., Kamble, E. E., Kumkar, S. N., Tawre, M. S., & Pardesi, K. R. (2019). Emerging Strategies to Combat ESKAPE Pathogens in the Era of Antimicrobial Resistance: A Review. *Front Microbiol*, *10*, 539. https://doi.org/10.3389/fmicb.2019.00539

Mushegian, A. R. (2020). Are There 10(31) Virus Particles on Earth, or More, or Fewer? *J Bacteriol*, *202*(9). https://doi.org/10.1128/JB.00052-20

Nethery, M. A., Korvink, M., Makarova, K. S., Wolf, Y. I., Koonin, E. V., & Barrangou, R. (2021). CRISPRclassify: Repeat-Based Classification of CRISPR Loci. *CRISPR J*, *4*(4), 558-574. https://doi.org/10.1089/crispr.2021.0021

Newsom, S., Parameshwaran, H. P., Martin, L., & Rajan, R. (2020). The CRISPR-Cas Mechanism for Adaptive Immunity and Alternate Bacterial Functions Fuels Diverse Biotechnologies. *Front Cell Infect Microbiol*, *10*, 619763. https://doi.org/10.3389/fcimb.2020.619763

Ofir, G., Herbst, E., Baroz, M., Cohen, D., Millman, A., Doron, S., Tal, N., Malheiro, D. B. A., Malitsky, S., Amitai, G., & Sorek, R. (2021). Antiviral activity of bacterial TIR domains via immune signalling molecules. *Nature*, *600*(7887), 116-120. https://doi.org/10.1038/s41586-021-04098-7

Ofir, G., & Sorek, R. (2018). Contemporary Phage Biology: From Classic Models to New Insights. *Cell*, *172*(6), 1260-1270. https://doi.org/10.1016/j.cell.2017.10.045

Oliveira, P. H., Touchon, M., & Rocha, E. P. (2014). The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. *Nucleic Acids Res*, *42*(16), 10618-10631. https://doi.org/10.1093/nar/gku734

Otsuka, Y., & Yonesaki, T. (2012). Dmd of bacteriophage T4 functions as an antitoxin against Escherichia coli LsoA and RnlA toxins. *Mol Microbiol*, *83*(4), 669-681. https://doi.org/10.1111/j.1365-2958.2012.07975.x

Payne, L. J., Meaden, S., Mestre, M. R., Palmer, C., Toro, N., Fineran, P. C., & Jackson, S. A. (2022). PADLOC: a web server for the identification of antiviral defence systems in microbial genomes. *Nucleic Acids Res*. https://doi.org/10.1093/nar/gkac400

Payne, L. J., Todeschini, T. C., Wu, Y., Perry, B. J., Ronson, C. W., Fineran, P. C., Nobrega, F. L., & Jackson, S. A. (2021). Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. *Nucleic Acids Res*, *49*(19), 10868-10878. https://doi.org/10.1093/nar/gkab883

Perry, E. B., Barrick, J. E., & Bohannan, B. J. (2015). The Molecular and Genetic Basis of Repeatable Coevolution between Escherichia coli and Bacteriophage T3 in a Laboratory Microcosm. *PLoS One*, *10*(6), e0130639. https://doi.org/10.1371/journal.pone.0130639

Pinilla-Redondo, R., Russel, J., Mayo-Munoz, D., Shah, S. A., Garrett, R. A., Nesme, J., Madsen, J. S., Fineran, P. C., & Sorensen, S. J. (2022). CRISPR-Cas systems are widespread accessory elements across bacterial and archaeal plasmids. *Nucleic Acids Res*, *50*(8), 4315-4328. https://doi.org/10.1093/nar/gkab859

Ran, F. A., Cong, L., Yan, W. X., Scott, D. A., Gootenberg, J. S., Kriz, A. J., Zetsche, B., Shalem, O., Wu, X., Makarova, K. S., Koonin, E. V., Sharp, P. A., & Zhang, F. (2015). In vivo genome editing using Staphylococcus aureus Cas9. *Nature*, *520*(7546), 186-191. https://doi.org/10.1038/nature14299

Rauch, B. J., Silvis, M. R., Hultquist, J. F., Waters, C. S., McGregor, M. J., Krogan, N. J., & Bondy-Denomy, J. (2017). Inhibition of CRISPR-Cas9 with Bacteriophage Proteins. *Cell*, *168*(1-2), 150-158 e110. https://doi.org/10.1016/j.cell.2016.12.009

Rice, L. B. (2008). Federal funding for the study of antimicrobial resistance in nosocomial pathogens: no ESKAPE. *J Infect Dis*, *197*(8), 1079-1081. https://doi.org/10.1086/533452

Rocha, E. P. C., & Bikard, D. (2022). Microbial defenses against mobile genetic elements and viruses: Who defends whom from what? *PLoS Biol*, *20*(1), e3001514. https://doi.org/10.1371/journal.pbio.3001514

Rodic, A., Blagojevic, B., Zdobnov, E., Djordjevic, M., & Djordjevic, M. (2017). Understanding key features of bacterial restriction-modification systems through quantitative modeling. *BMC Syst Biol*, *11*(Suppl 1), 377. https://doi.org/10.1186/s12918-016-0377-x

Rostol, J. T., & Marraffini, L. (2019). (Ph)ighting Phages: How Bacteria Resist Their Parasites. *Cell Host Microbe*, *25*(2), 184-194. https://doi.org/10.1016/j.chom.2019.01.009

Rousset, F., Depardieu, F., Miele, S., Dowding, J., Laval, A. L., Lieberman, E., Garry, D., Rocha, E. P. C., Bernheim, A., & Bikard, D. (2022). Phages and their satellites encode hotspots of antiviral systems. *Cell Host Microbe*, *30*(5), 740-753 e745. https://doi.org/10.1016/j.chom.2022.02.018

Rusinov, I. S., Ershova, A. S., Karyagina, A. S., Spirin, S. A., & Alexeevski, A. V. (2018). Avoidance of recognition sites of restriction-modification systems is a widespread but not universal anti-restriction strategy of prokaryotic viruses. *BMC Genomics*, *19*(1), 885. https://doi.org/10.1186/s12864-018-5324-3

Russel, J., Pinilla-Redondo, R., Mayo-Munoz, D., Shah, S. A., & Sorensen, S. J. (2020). CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. *CRISPR J*, *3*(6), 462-469. https://doi.org/10.1089/crispr.2020.0059

Salmond, G. P., & Fineran, P. C. (2015). A century of the phage: past, present and future. *Nat Rev Microbiol*, *13*(12), 777-786. https://doi.org/10.1038/nrmicro3564

Schmartz, G. P., Hartung, A., Hirsch, P., Kern, F., Fehlmann, T., Muller, R., & Keller, A. (2022). PLSDB: advancing a comprehensive database of bacterial plasmids. *Nucleic Acids Res*, *50*(D1), D273-D278. https://doi.org/10.1093/nar/gkab1111

Shmakov, S. A., Sitnik, V., Makarova, K. S., Wolf, Y. I., Severinov, K. V., & Koonin, E. V. (2017). The CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes. *mBio*, *8*(5). https://doi.org/10.1128/mBio.01397-17

Shmakov, S. A., Utkina, I., Wolf, Y. I., Makarova, K. S., Severinov, K. V., & Koonin, E. V. (2020). CRISPR Arrays Away from cas Genes. *CRISPR J*, *3*(6), 535-549. https://doi.org/10.1089/crispr.2020.0062

Shmakov, S. A., Wolf, Y. I., Savitskaya, E., Severinov, K. V., & Koonin, E. V. (2020). Mapping CRISPR spaceromes reveals vast host-specific viromes of prokaryotes. *Commun Biol*, *3*(1), 321. https://doi.org/10.1038/s42003-020-1014-1

Tesson, F., Herve, A., Mordret, E., Touchon, M., d'Humieres, C., Cury, J., & Bernheim, A. (2022). Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nat Commun*, *13*(1), 2561. https://doi.org/10.1038/s41467-022-30269-9

Vassallo, C., Doering, C., Littlehale, M. L., Teodoro, G., & Laub, M. T. (2022). Mapping the landscape of anti-phage defense mechanisms in the <em>E. coli</em> pangenome. *bioRxiv*.

Vyas, P., & Harish. (2022). Anti-CRISPR proteins as a therapeutic agent against drug-resistant bacteria. *Microbiol Res*, *257*, 126963. https://doi.org/10.1016/j.micres.2022.126963

Wang, J., Dai, W., Li, J., Li, Q., Xie, R., Zhang, Y., Stubenrauch, C., & Lithgow, T. (2021). AcrHub: an integrative hub for investigating, predicting and mapping anti-CRISPR proteins. *Nucleic Acids Res*, *49*(D1), D630-D638. https://doi.org/10.1093/nar/gkaa951

Wang, J., Dai, W., Li, J., Xie, R., Dunstan, R. A., Stubenrauch, C., Zhang, Y., & Lithgow, T. (2020). PaCRISPR: a server for predicting and visualizing anti-CRISPR proteins. *Nucleic Acids Res*, *48*(W1), W348-W357. https://doi.org/10.1093/nar/gkaa432

Wang, S., Wan, M., Huang, R., Zhang, Y., Xie, Y., Wei, Y., Ahmad, M., Wu, D., Hong, Y., Deng, Z., Chen, S., Li, Z., & Wang, L. (2021). SspABCD-SspFGH Constitutes a New Type of DNA Phosphorothioate-Based Bacterial Defense System. *mBio*, *12*(2). https://doi.org/10.1128/mBio.00613-21

Weinberger, A. D., Wolf, Y. I., Lobkovsky, A. E., Gilmore, M. S., & Koonin, E. V. (2012). Viral diversity threshold for adaptive immunity in prokaryotes. *mBio*, *3*(6), e00456-00412. https://doi.org/10.1128/mBio.00456-12

Weissman, J. L., Laljani, R. M. R., Fagan, W. F., & Johnson, P. L. F. (2019). Visualization and prediction of CRISPR incidence in microbial trait-space to identify drivers of antiviral immune strategy. *ISME J*, *13*(10), 2589-2602. https://doi.org/10.1038/s41396-019-0411-2

Westra, E. R., Staals, R. H., Gort, G., Hogh, S., Neumann, S., de la Cruz, F., Fineran, P. C., & Brouns, S. J. (2013). CRISPR-Cas systems preferentially target the leading regions of MOBF conjugative plasmids. *RNA Biol*, *10*(5), 749-761. https://doi.org/10.4161/rna.24202

Westra, E. R., van Houte, S., Oyesiku-Blakemore, S., Makin, B., Broniewski, J. M., Best, A., Bondy-Denomy, J., Davidson, A., Boots, M., & Buckling, A. (2015). Parasite Exposure Drives Selective Evolution of Constitutive versus Inducible Defense. *Curr Biol*, *25*(8), 1043-1049. https://doi.org/10.1016/j.cub.2015.01.065

Wheatley, R. M., & MacLean, R. C. (2021). CRISPR-Cas systems restrict horizontal gene transfer in Pseudomonas aeruginosa. *ISME J*, *15*(5), 1420-1433. https://doi.org/10.1038/s41396-020-00860-3

Wimmer, F., & Beisel, C. L. (2019). CRISPR-Cas Systems and the Paradox of Self-Targeting Spacers. *Front Microbiol*, *10*, 3078. https://doi.org/10.3389/fmicb.2019.03078

Yehl, K., Lemire, S., Yang, A. C., Ando, H., Mimee, M., Torres, M. T., de la Fuente-Nunez, C., & Lu, T. K. (2019). Engineering Phage Host-Range and Suppressing Bacterial Resistance through Phage Tail Fiber Mutagenesis. *Cell*, *179*(2), 459-469 e459. https://doi.org/10.1016/j.cell.2019.09.015

Yi, H., Huang, L., Yang, B., Gomez, J., Zhang, H., & Yin, Y. (2020). AcrFinder: genome mining anti-CRISPR operons in prokaryotes and their viruses. *Nucleic Acids Res*, *48*(W1), W358-W365. https://doi.org/10.1093/nar/gkaa351

Yin, Y., Yang, B., & Entwistle, S. (2019). Bioinformatics Identification of Anti-CRISPR Loci by Using Homology, Guilt-by-Association, and CRISPR Self-Targeting Spacer Approaches. *mSystems*, *4*(5). https://doi.org/10.1128/mSystems.00455-19

Yourik, P., Fuchs, R. T., Mabuchi, M., Curcuru, J. L., & Robb, G. B. (2019). Staphylococcus aureus Cas9 is a multiple-turnover enzyme. *RNA*, *25*(1), 35-44. https://doi.org/10.1261/rna.067355.118

Yu, L., & Marchisio, M. A. (2020). Types I and V Anti-CRISPR Proteins: From Phage Defense to Eukaryotic Synthetic Gene Circuits. *Front Bioeng Biotechnol*, *8*, 575393. https://doi.org/10.3389/fbioe.2020.575393

Zhang, F., Zhao, S., Ren, C., Zhu, Y., Zhou, H., Lai, Y., Zhou, F., Jia, Y., Zheng, K., & Huang, Z. (2018). CRISPRminer is a knowledge base for exploring CRISPR-Cas systems in microbe and phage interactions. *Commun Biol*, *1*, 180. https://doi.org/10.1038/s42003-018-0184-6

Zhang, Y., Zhang, Z., Zhang, H., Zhao, Y., Zhang, Z., & Xiao, J. (2020). PADS Arsenal: a database of prokaryotic defense systems related genes. *Nucleic Acids Res*, *48*(D1), D590-D598. https://doi.org/10.1093/nar/gkz916

Zheng, Z., Zhang, Y., Liu, Z., Dong, Z., Xie, C., Bravo, A., Soberon, M., Mahillon, J., Sun, M., & Peng, D. (2020). The CRISPR-Cas systems were selectively inactivated during evolution of Bacillus cereus group for adaptation to diverse environments. *ISME J*, *14*(6), 1479-1493. https://doi.org/10.1038/s41396-020-0623-5

Appendix

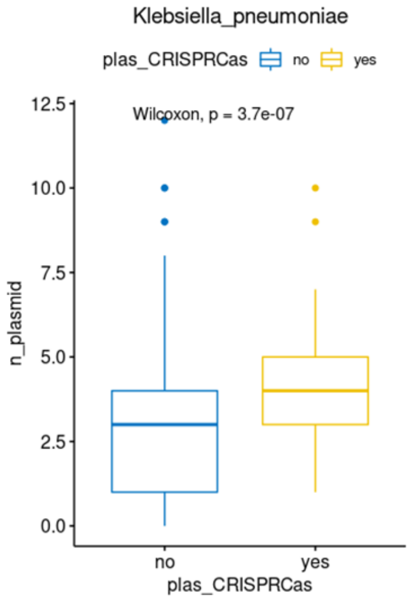


Figure A.1 The significant test between the presence of plasmid encoding

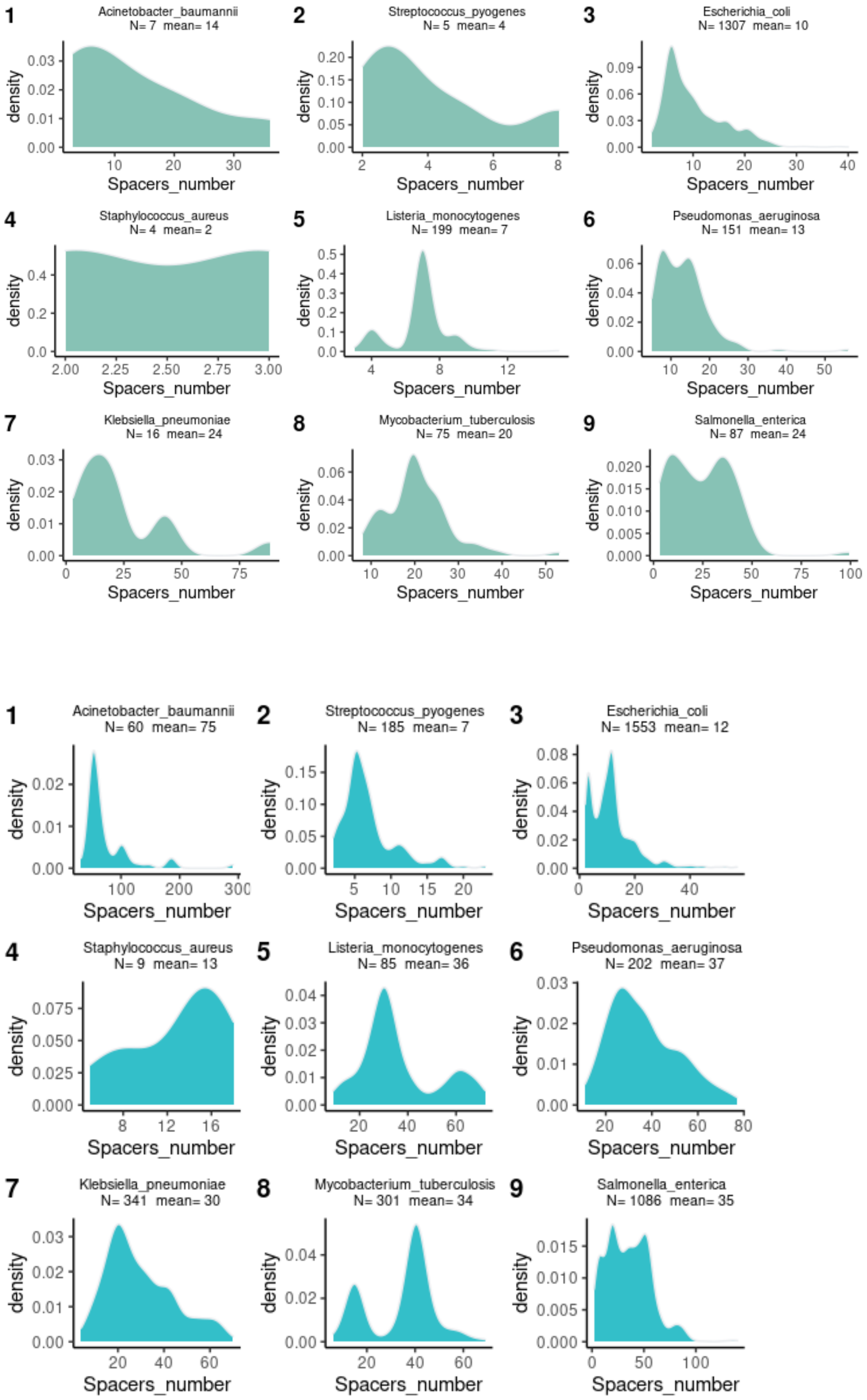


Figure A.2 The distribution of spacer number from orphan CRISPR arrays (top) and Cas operon adjacent CRISPR arrays (bottom).

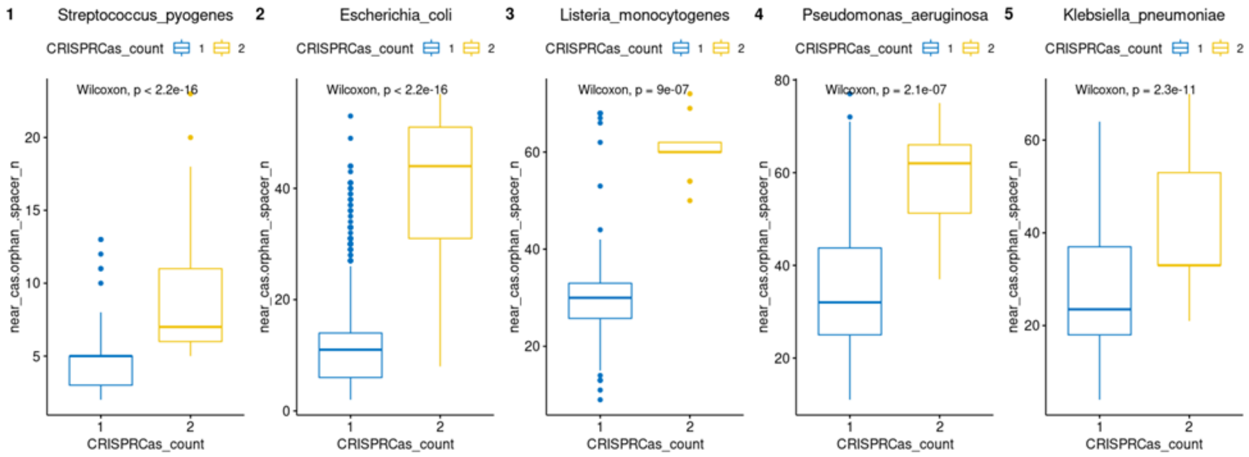


Figure A.3 The significant test between the number of CRISPR-Cas and spacer numbers.

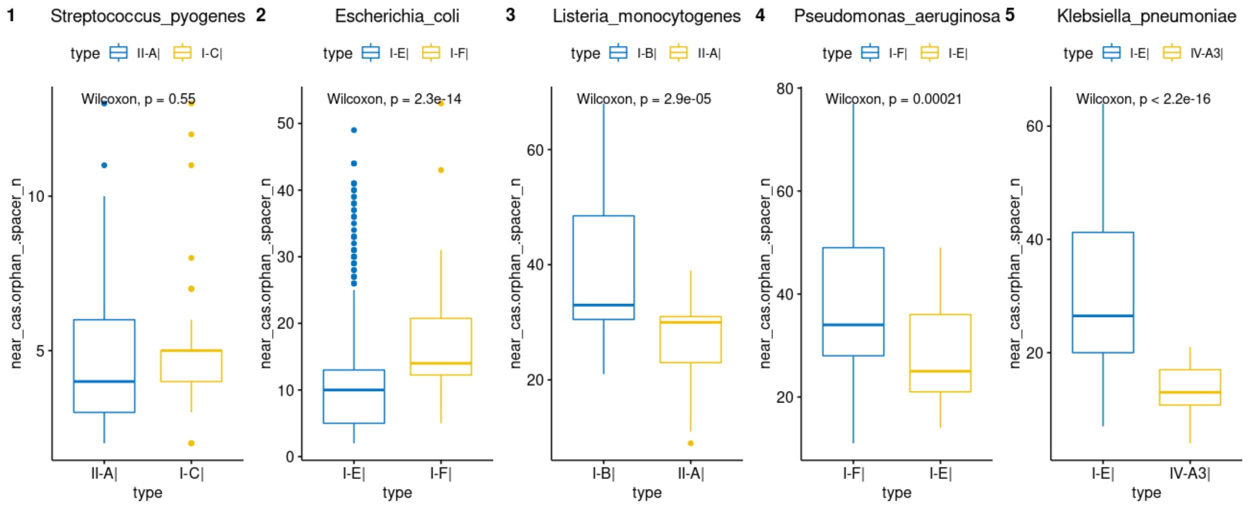


Figure A.4 The association between the subtypes of CRISPR-Cas and the spacer number of each system. Removing the subtypes smaller than 5%. The test is conducted on the genome only encoding one complete CRISPR-Cas system

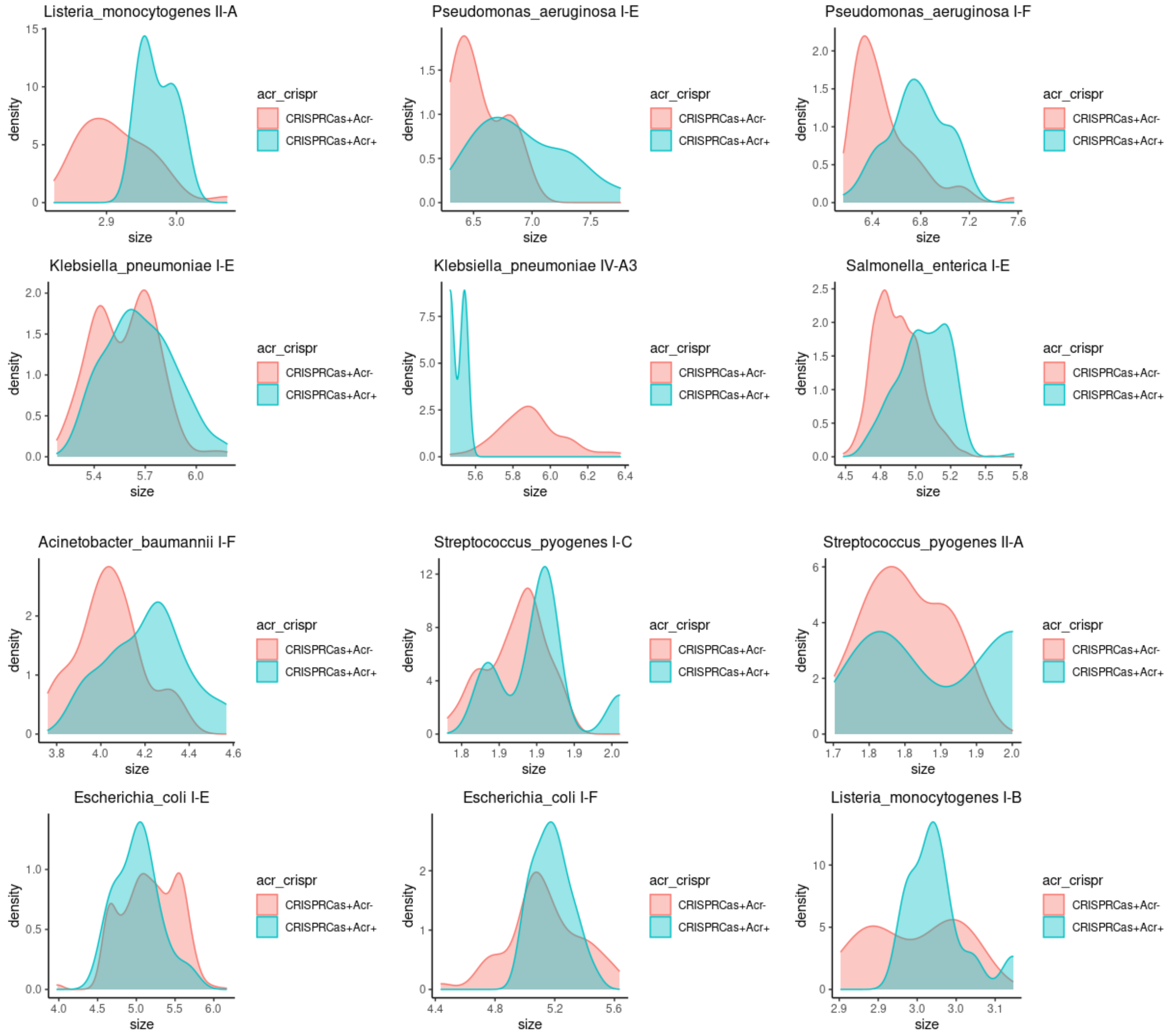


Figure A.5 The impacts of active CRISPR-Cas on genome size of each subtype of CRISPR-Cas

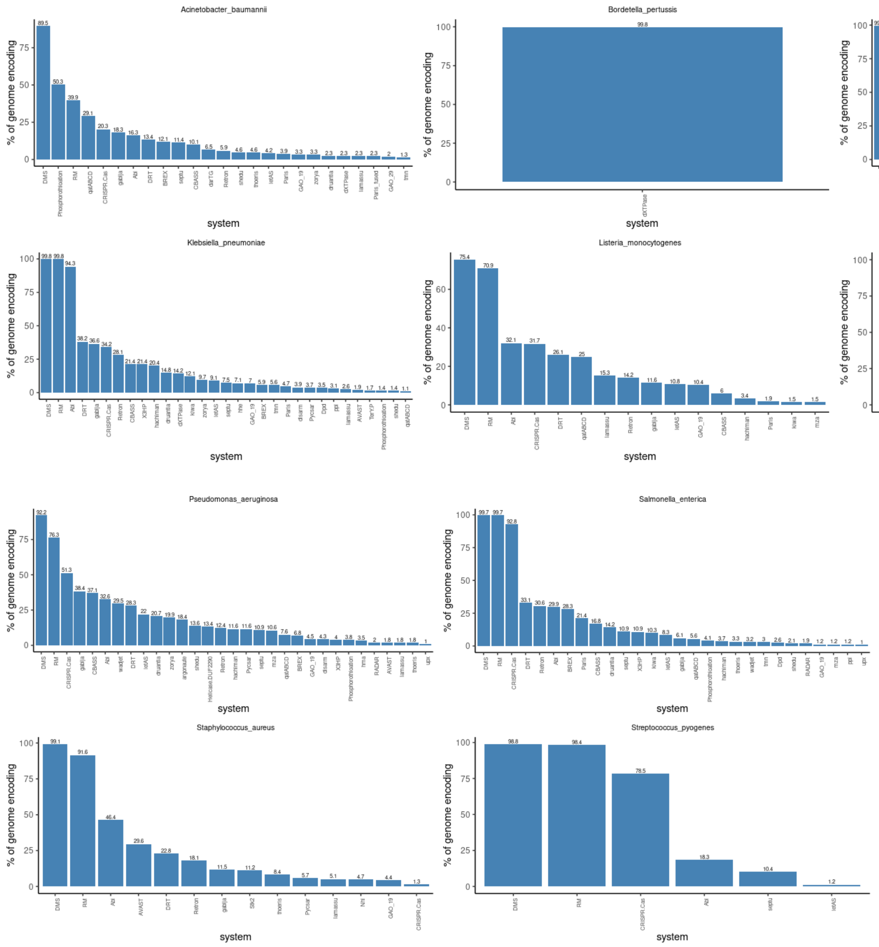


Figure A.6 A. The frequency of systems encoded on genomes of each 10 species.

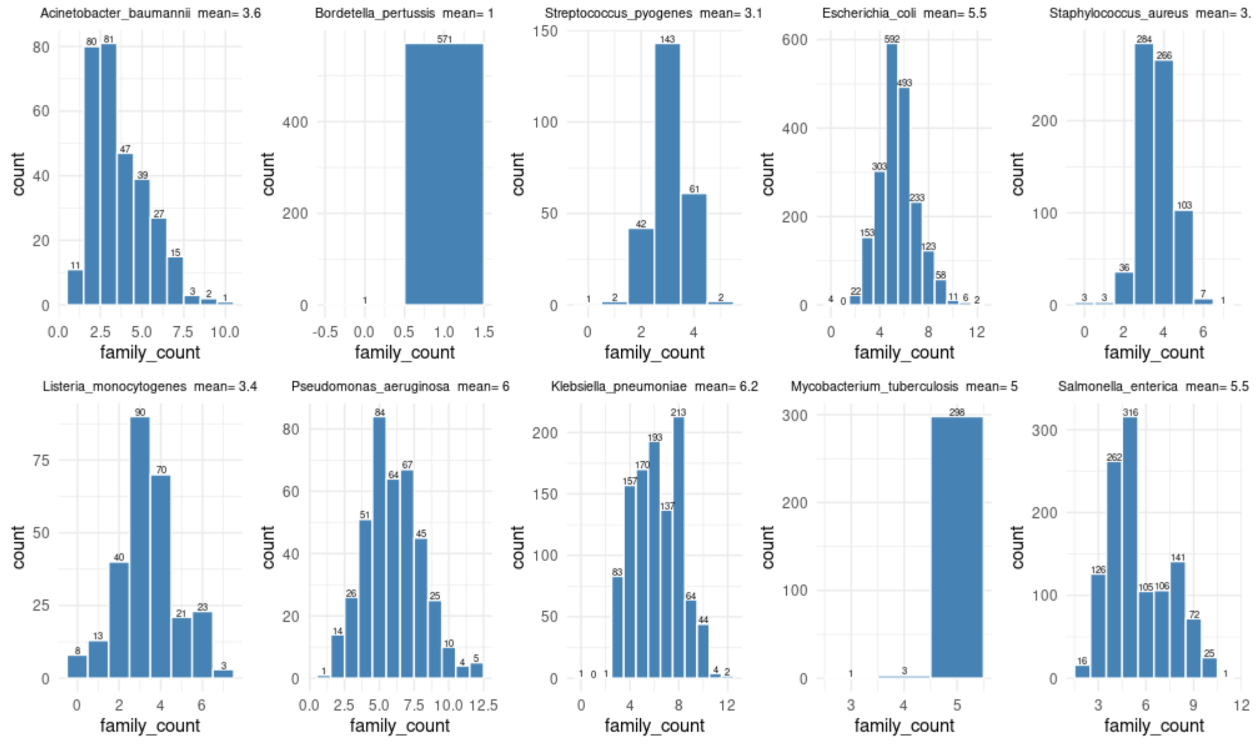


Figure A.6.B. The distribution of family count of 10 species.

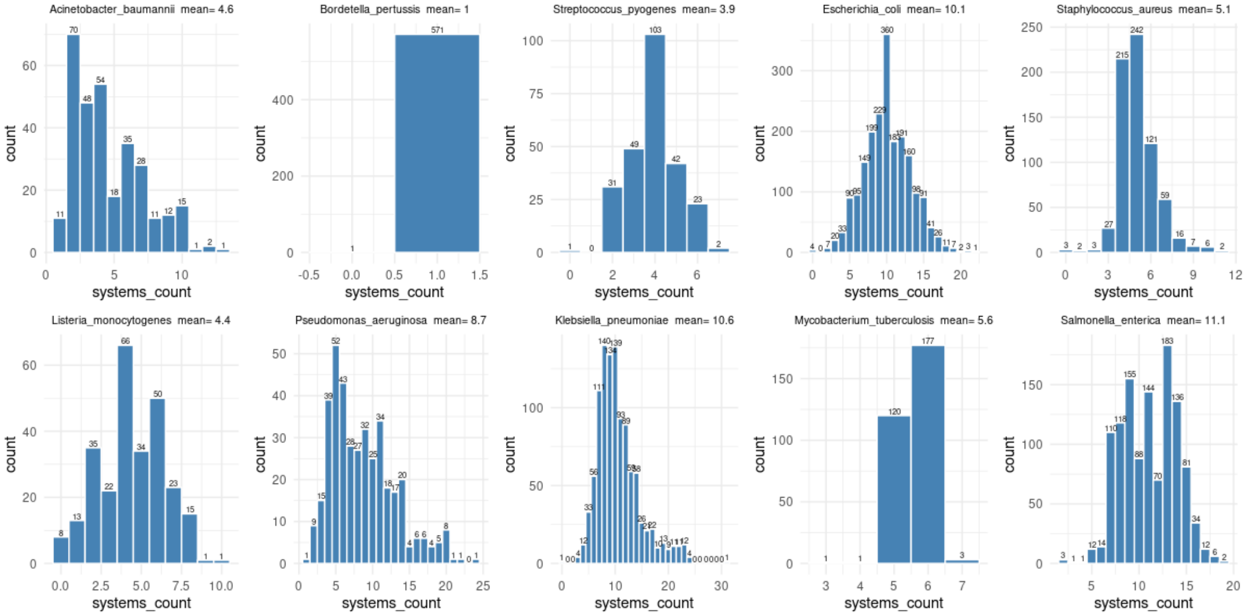


Figure A.6.C. The distribution of systems counts of 10 species

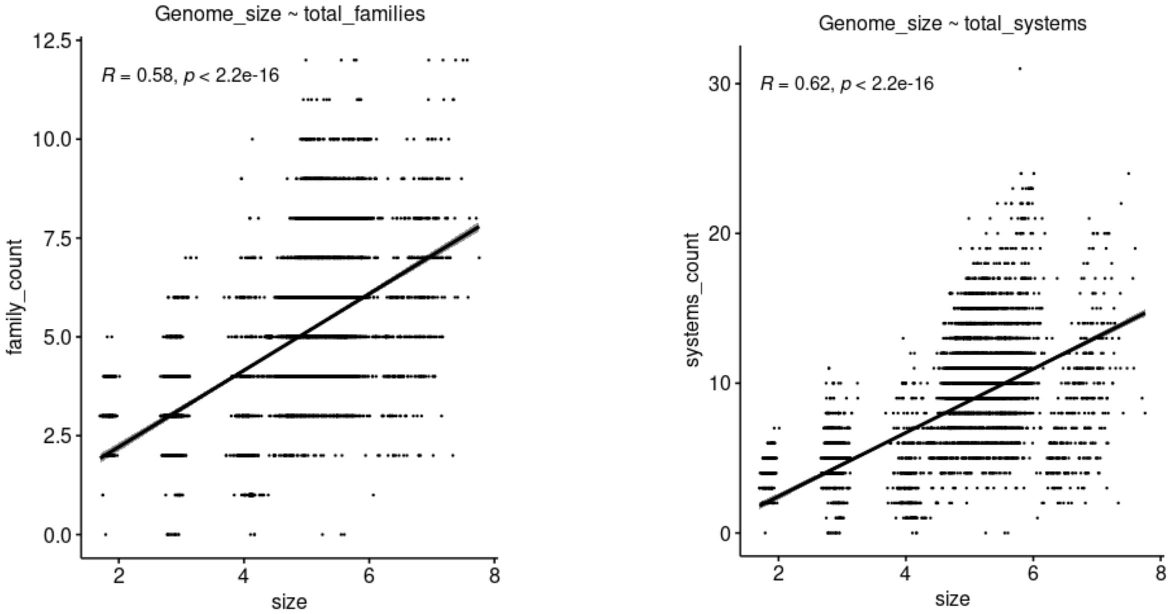


Figure A.7 The correlation test of genome size and the number of families (left) and the number of systems (right).

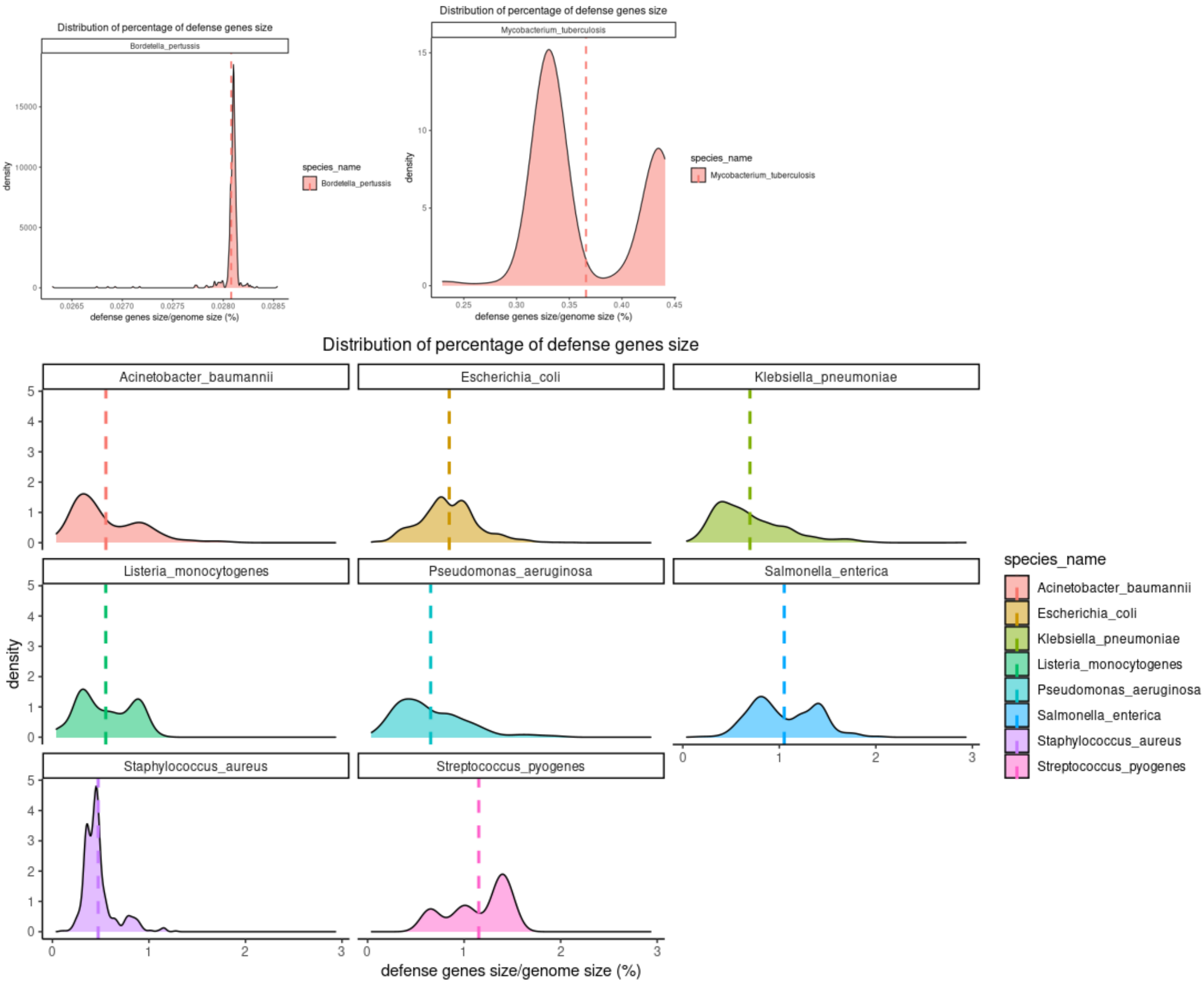


Figure A.6 The distribution of defense genes/genome size of each specie.

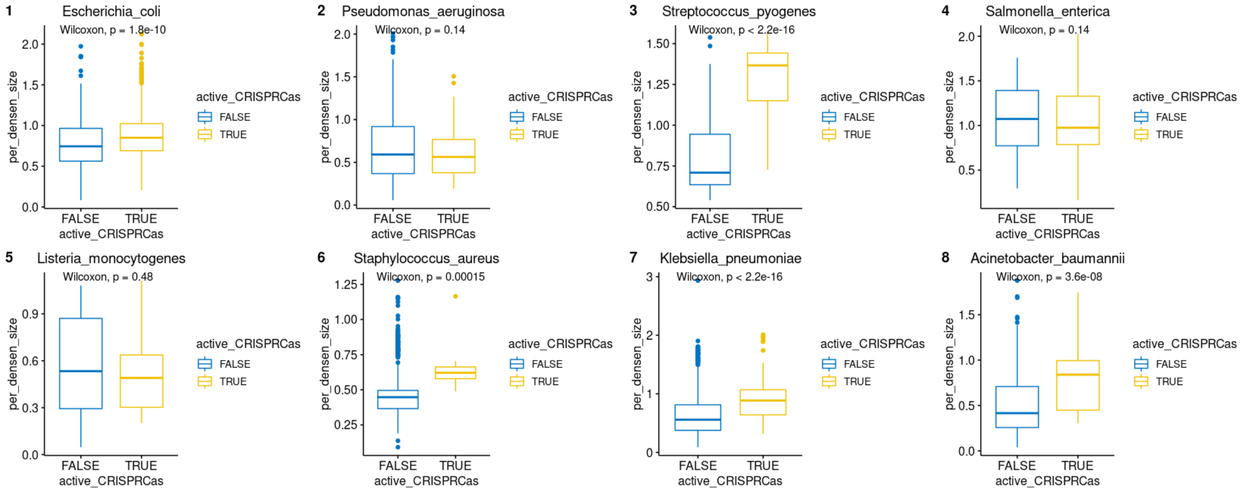


Figure A.7 The siginificane test of presence of active CRISPRCas and the distribution of defense genes/genome size of each specie.

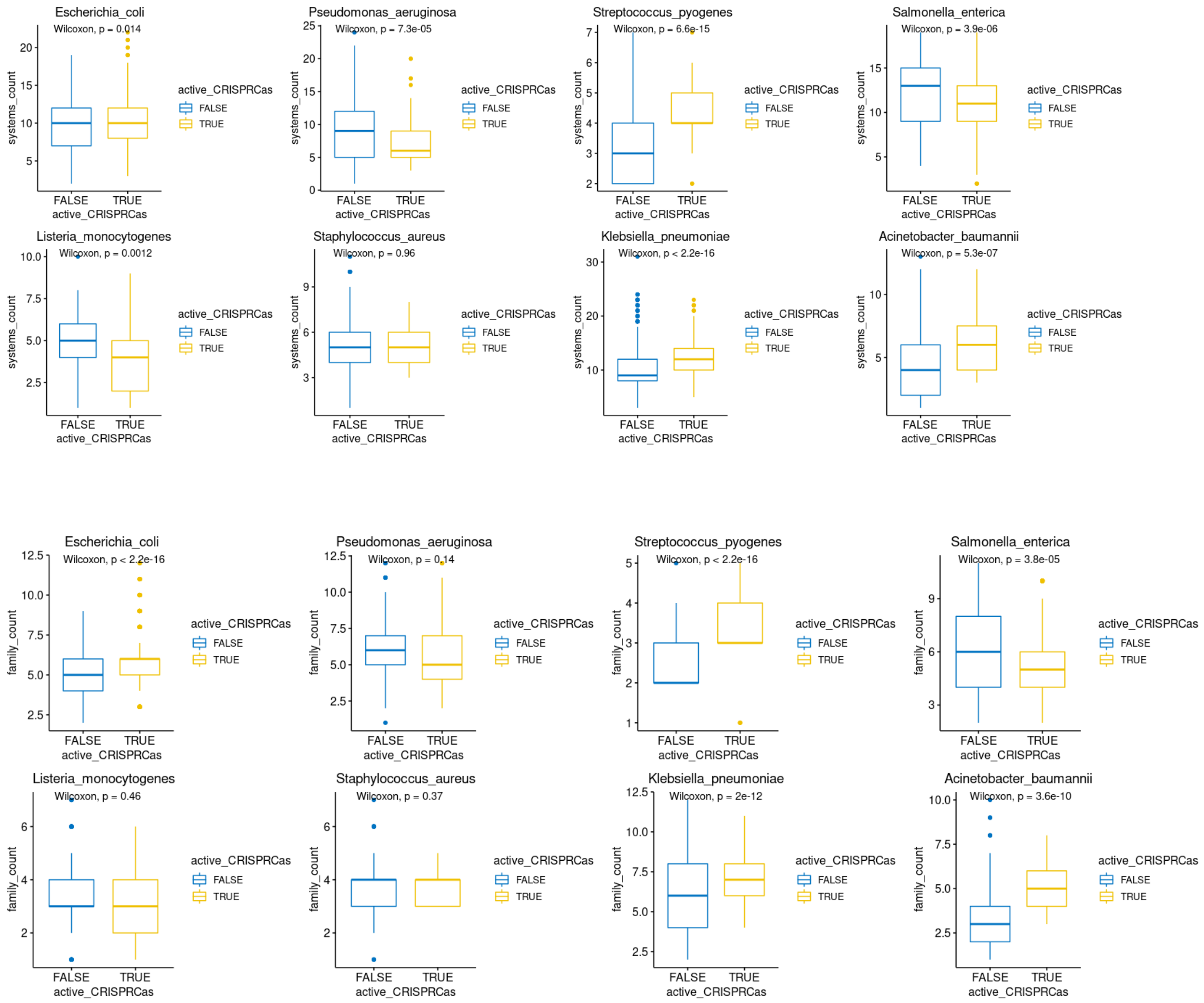


Figure A.8 The association active CRISPR-Cas and number of defence systems and defence families.