Exploring and Exploiting Prokaryotic Immunity in *Salmonella*

THESIS

Submitted in partial fulfilment

of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

Simran Krishnakant Kushwaha

2018PHXF0406P

Under the Supervision of

Prof. Sandhya Amol Marathe

&

Co-supervision of

Prof. Franklin L. Nobrega

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

PILANI, RAJASTHAN, INDIA 333031

2024

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

CERTIFICATE

This is to certify that the thesis entitled **“Exploring and Exploiting Prokaryotic Immunity in *Salmonella*”,** submitted by **Ms. Simran Krishnakant Kushwaha**, ID No. **2018PHXF0406P,** for the award of the Ph.D. degree from the Institute embodies the original work done by her under our supervision.

Signature of the Supervisor:

Name: Prof. Sandhya Amol Marathe

Designation: Associate Professor

Date:

Place:

Signature of the Co-supervisor:

Name: Prof. Franklin L. Nobrega

Designation: Associate Professor

Date: 08 May 2024

Place: Southampton, UK

"To Twitter, my furry angel. Your love and presence fueled every step of my life".

ACKNOWLEDGEMENTS

"In the journey of academic pursuit, we stand on the shoulders of giants, propelled by the collective wisdom, encouragement, and support of those who have walked before us. This thesis is a testament to the power of mentorship, collaboration, and the intricate web of relationships that shape the growth of knowledge. To all those who have illuminated my path with their guidance and kindness, I extend my deepest gratitude."

I want to express my sincere gratitude to my supervisor, Prof. Sandhya Amol Marathe, for her invaluable guidance, support, and mentorship throughout this research. Her feedback, patience, and dedication have been instrumental in shaping the direction of my thesis. I distinctly recall the first year of my PhD journey, marked by a significant disparity in research expertise bridging the gap between my bachelor's and doctoral studies. During this formative phase, she played a pivotal role in elucidating the essence of a PhD and guiding me through its intricacies.

I am profoundly grateful to my co-supervisor, Prof. Franklin L. Nobrega, for his invaluable advice, continuous support, and patience during my PhD. He is the individual who guided me in embracing the true essence of scientific inquiry. I thank him for allowing me to collaborate within his laboratory and furnishing the utmost excellent resources. I gained not only scientific knowledge from him but also life skills, interpersonal communication, and the art of self-presentation.

I thank my father and mother for their unwavering support and enduring belief in my capabilities. Their sacrifices, encouragement, and guidance have shaped my academic journey. Their faith in me has empowered me to navigate challenges and strive for excellence.

With the most heartfelt gratitude, I thank my friends Arvind, Roshan and Hrishav. Arvind- My Essential Pillar. A profound gratitude to a person who quickly transformed into a vital cornerstone of my endeavours. Roshan- My Unwavering Anchor. His selfless friendship, support, and shared experiences have enriched every aspect of my life, from academia to personal growth. Hrishav- My Python Guru-The person I pestered the most with my coding challenges, and he graciously helped me navigate through them.

I extend my sincere gratitude to my lab mates Ankita, Shreya, and Shobham, for it was within the walls of our shared workspace that most of my PhD journey unfolded. Ankita- My Lab Rant Companion- The individual with whom I share the deepest bonds in BITS. I laugh at the memory of how we screwed up my cloning experiment and ended up having a good dinner in silence. Shreya- The Lab Giggle Luminary- The individual who entertained me the most in the lab. Shobham- The Captain Distraction- The person with whom I've indulged in countless junk food moments and who would unfailingly drop by my lab every morning to say Hi.

I wish to express my heartfelt gratitude to the members of the Director Advisory Committee, Prof. Prabhat Nath Jha and Prof. Syamantak Majumder, for their invaluable inputs and insightful suggestions. I genuinely appreciate the time and effort they have dedicated to my research. Prof. Jha has consistently served as an exceptional source of motivation throughout my PhD journey, and I cherish our enriching discussions about anything and everything. I would also like to thank Prof. A K Das for his comprehensive insights in molecular biology and the casual conversations we shared daily in the corridors.

I wish to express my sincere appreciation to the members of the EMBL, especially Dr. Nandita Sharma, Saumya Arora, Dr. Tripti Misra, Dr. Monika Sandhu and Dr. Shahid Khan, for their invaluable scientific inputs and engaging culinary sessions. I want to thank my collaborators, with whom I could engage in fascinating scientific endeavours and broaden the horizons of my research. I wholeheartedly extend my love to my students, whose support in my research has been an invaluable source of assistance and encouragement.

As I conclude this chapter of my academic journey, I am reminded of the profound impact that collective support and guidance have had on shaping my path. Each individual mentioned in these expressions of gratitude has played an integral role in moulding my experiences, fortifying my resilience, and expanding my horizons. I sincerely appreciate all who illuminated my path and made this journey possible.

Simran Krishnakant Kushwaha

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Symbol	Abbreviation
Abi	Abortive Infection
ABR	Antibiotic Resistance
Acr	Anti-CRISPR
BREX	Bacteriophage Exclusion
Cas	CRISPR-Associated
ChIP	Chromatin Immunoprecipitation
cDNA	Complementary Deoxyribonucleic Acid
CFU	Colony Forming Units
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CRP	Cyclic AMP Receptor Protein
crRNA	CRISPR RNA
CTAB	Cetyltrimethylammonium Bromide
CU	Chaperone/Usher
DISARM	Defence Island System Associated With Restriction–Modification
DNA	Deoxyribonucleic Acid
DND	DNA Degradation
DR	Direct Repeats
EDTA	Ethylenediamine Tetra Acetic Acid
EMSA	Electrophoretic Mobility Shift Assays
EPEC	Enteropathogenic *E. coli*
FGC	Fimbrial Gene Cluster
GFP	Green Fluorescent Protein
gyrB	DNA Gyrase B
H-NS	Histone-Like Nucleoid-Structuring
HGT	Horizontal Gene Transfer
IHF	Integration Host Factor
LB	Luria-Bertani Medium
LCA	Last Common Ancestor
LPS	Lipopolysaccharide
LRP	Leucine-Responsive Regulatory Protein
MGE	Mobile Genetic Element
ML	Maximum Likelihood
MLST	Multi-Locus Sequence Typing
mRNA	Messenger Ribonucleic Acid
NTS	Non-Typhoidal *Salmonella*
PAM	Protospacer Adjacent Motif
PARIS	Phage Anti-Restriction-Induced System
PI	Pathogenicity Island

PRS	Potential Regulatory Spacers
pSV	*Salmonella* Virulence Plasmids
QAC	Quaternary Ammonium Compounds
QS	Quorum Sensing
RGP	Regions of Genomic Plasticity
R-M	Restriction-Modification
RNA	Ribonucleic Acid
RT	Room Temperature
s-HSPs	Small Heat Shock Proteins
SD	Standard Deviation
SNP	Single Nucleotide Polymorphism
SPI	*Salmonella* Pathogenicity Island
Stn	*Salmonella* Enterotoxin
ssa	Secretion System Apparatus
ssc	Secretion System Chaperons
sse	Secretion System Effectors
ssr	Secretion System Regulators
STS	Self-Targeting Spacers
T3SS	Type III Secretion System
TCS	Two-Component Signal Transduction Systems
TMP-SMX	Trimethoprim/Sulfamethoxazole
TS	Typhoidal *Salmonella*
UPGMA	Unweighted Pair Group Method With Arithmetic Mean

ABSTRACT

Salmonella, a versatile bacterial pathogen, is a formidable threat due to its involvement in widespread outbreaks, impacting populations across developed and developing nations. Within the *Salmonella* genus, two distinct species, alongside a diverse array of subspecies and serovars, have a complex genomic landscape. This complexity is driven by myriad factors, with horizontal gene transfer (HGT) playing a pivotal role. This flexible genome (containing the accessory genes, present in <90% strains) is organised in regions of genomic plasticity (RGP) and serves as a potent facilitator of the dynamic evolution of bacterial genomes through gene acquisition and loss. Our study on the genomic plasticity across *Salmonella* lineages revealed a purposeful, non-random integration pattern of pathogenicity-related gene clusters into strategic locations (spots). Noteworthy examples include the correlation between the type I-E CRISPR-Cas system, gold tolerance, and specific spots. The scattered prevalence of RGP across *Salmonella* lineages profoundly shapes the pathogenicity makeup of *Salmonella* strains. The preferences of RGP seem guided by conserved flanking genes that likely share regulatory and functional coordination. For example, RGPs housing metal resistance genes are positioned near stress resistance genes, indicating a regulatory network to efficiently counter stressors. Additionally, we observed that different plasmid incompatibility types and prophage genera carry distinct pathogenicity genes. Similar to RGPs, their distribution across *Salmonella* lineages plays a critical role in defining pathogenicity.

Our analyses indicate the prevalence of the type I-E CRISPR-Cas system in *Salmonella*, with notable conservation in spot #22. We aimed to delve deeper into understanding the intricacies of the CRISPR-Cas system. To gain insights into the evolution of *Salmonella* in association with the CRISPR-Cas genes, we performed phylogenetic surveillance across strains belonging to *Salmonella* serovars. The strains differed in their CRISPR1-leader and *cas* operon features, assorting into two main clades, CRISPR1-STY/*cas*-STY and CRISPR1-STM/*cas*-STM, comprising mainly typhoidal and non-typhoidal *Salmonella* serovars, respectively. Serovars of these two clades displayed better relatedness concerning CRISPR1-leader and *cas* operon across genera than between themselves. This signifies that the CRISPR/Cas region acquisition could be through an HGT

event owing to the presence of mobile genetic elements flanking the CRISPR1 array. The observed discordance between the phylogenetic trees of various CRISPR-Cas components and the MLST phenogram suggests the differential evolution of the CRISPR-Cas system.

We extensively examined 7,624 unique CRISPR spacers in 52 *Salmonella* serovars to gain a profound understanding of the system's role in *Salmonella* physiology. The analysis revealed variability in spacer counts among serovars, with broader host-range (infecting multiple species) serovars displaying higher counts. Notably, only a small percentage of spacers (4.8%) show matches against plasmids, and 0.6% match phages, suggesting alternative functional roles. No distinct negative correlation between spacer count and prophage prevalence was observed. We found that the spacers show partial matches against their genomes, perhaps regulating the endogenous genes. Closer inspection in serovars Enteritidis, Typhimurium, and Typhi indicated potential regulation of genes associated with various biological functions by highly conserved spacers (by sequence) within the serovars. For instance, the genes linked to DNA repair processes (*recA*, *ruvB*), stress response (*mdtB*, *mrcB*), biofilm formation (*cadC, bcsC, ratB, pepB*), bacterial infections (lon protease, *sipD*), and directly interacting with the CRISPR-Cas system (*leuO, igaA*) have been targeted in a significant proportion of strains. The expression of some of these genes, like *bcsC, sipD,* etc., are reportedly affected by the CRISPR-Cas system. Furthermore, the flexibility in PAM recognition by Cas proteins is proposed to influence gene regulation.

We next explored the conditions activating *Salmonella*'s CRISPR-Cas system to exploit it for self-killing. The experimental verification of the system's activation under growth conditions like stress and biofilm showed a lack of detectable *cas* gene expression. The CRISPR-Cas system was robustly and functionally activated in various serovars by supplying LeuO, a transcriptional activator, in trans. Nevertheless, selecting the *Salmonella*-specific protospacers from its genome, we observed less than 35% self-killing.

In conclusion, our study unveils intricate connections among gene clusters, RGPs, mobile genetic elements, and pathogenic attributes while offering novel insights into the evolutionary trajectory of *Salmonella*. Further, the CRISPR-Cas system exhibits diverse evolutionary patterns and spacer functionalities. Despite attempts to leverage this system for species-specific eradication, challenges highlight the complexities of using endogenous CRISPR-Cas systems as an anti-microbial and warrant further strategic refinements.

Chapter 1

General Introduction

Publications from this Chapter-

1. Kushwaha SK, Narasimhan LP, Chithananthan C, Marathe SA. Clustered regularly interspaced short palindromic repeats-Cas system: diversity and regulation in *Enterobacteriaceae*. Future Microbiology. 2022 Oct;17:1249-1267. DOI: 10.2217/fmb-2022-0081. PMID: 36006039.

1.1 Evolution and classification of *Salmonella*

Enterobacteriaceae, a family of Gram-negative bacteria, includes both pathogenic and non-pathogenic bacteria. It is a member of the domain Bacteria, phylum Proteobacteria, class Gammaproteobacteria, and order Enterobacteriales (Donnenberg *et al.,* 2014), consisting of over 30 genera and 120 species. These bacteria are commonly found in the small and large gastrointestinal tracts and are often called enterics. They encompass beneficial commensal microbiota, opportunistic pathogens (that can cause significant harm to immunocompromised individuals) and primary pathogens (that can initiate illnesses even in healthy individuals). This range of pathogenicity is linked to the presence or absence of specific virulence factors that contribute to the disease process (Janda & Abbott, 2021). Around 95% of clinically essential strains are found within 10 genera and fewer than 25 species (Rock & Donnenberg, 2014). The most prominent Enterobacteriales include *Escherichia*, *Shigella,* and *Salmonella* (Dekker & Frank, 2015). Some *Escherichia* species are beneficial gut inhabitants, while some are potential pathogens causing foodborne illnesses and urinary tract infections. *Shigella* species are responsible for shigellosis, a disease characterised by severe diarrhoea and abdominal cramps. *Salmonella* is known for its role in salmonellosis, a foodborne disease associated with contaminated food products (Dekker & Frank, 2015).

Escherichia and *Salmonella* exhibit significant similarity due to their close evolutionary relationship within the *Enterobacteriaceae* family (Fukushima, Kakinuma, & Kawaguchi, 2002). The genomes of these two species are essentially superimposable, and genome sequencing has demonstrated an 80% median homology between non-pathogenic *E. coli* and *Salmonella enterica* subspecies *enterica* serovar Typhimurium genomes (McClelland *et al.*, 2000). Throughout their evolution, the integration of pathogenicity islands (PIs) and phage-associated genes into the genome influenced *Salmonella*'s virulence profile (**Fig. 1.1**) (Schmidt & Hensel, 2004). This drove its divergence from *E. coli* ~100–150 million years ago, as it developed strategies to invade varied hosts and develop resistance mechanisms (Lamas *et al.,* 2018).

Salmonella genus includes *S. bongori* and *S. enterica* (**Fig. 1.1**) (Tanner & Kingsley, 2018). The split between the two species is estimated to have occurred around 40 to 63.4 million years ago (McQuiston *et al.*, 2008). Evolutionarily, *S. bongori* is positioned between *E. coli* and *S. enterica* and have ancestrally retained basic virulence functions and lacks some specific *S. enterica* metabolic pathways involving biosynthesis of amino acids, carbohydrates, fatty acids and lipids. *S. enterica* possesses a full set of type III secretion systems (T3SS-1 and T3SS-2), unlike *S. bongori,* which lacks T3SS-2 (**Fig. 1.1**) that is vital for optimal replication in host macrophages (Fookes *et al.*, 2011). After divergence, *S. bongori* evolved, gaining twelve T3SS candidate effector proteins. Ten of these are absent in other *Salmonella* but relate to those found in enteropathogenic *E. coli* strains (Fookes *et al.*, 2011).

S. enterica is further categorised into six subspecies – I: *enterica,* II: *salamae,* IIIa: *arizonae,* IIIb: *diarizonae,* IV: *hountenae,* and VI: *indica*, with over 2600 serovars (**Fig. 1.1**) (Gal-Mor, Boyle, & Grassl, 2014)*. S. bongori* and all *S. enterica* subspecies except subspecies *enterica* infect poikilotherms and are generally found in the nonhost environment. Their presence in homeotherms is infrequent. In comparison, *S. enterica* subspecies *enterica* infects homeotherms (Tanner & Kingsley, 2018).

Salmonella serovars are classified as per the White-Kauffmann-Le Minor system based on the antigenic formulae for H (flagellar proteins) and O (oligosaccharides of lipopolysaccharide) antigens (Kaniuk *et al.,* 2002). Furthermore, based on their ability to adapt to different hosts, these serovars can be categorised into three distinct groups (Tanner & Kingsley, 2018).

(i) Adapted to humans and higher primates: *Salmonella* serovar Typhi, Paratyphi A and Sendai. These are categorised as typhoidal *Salmonella* serovars in humans owing to their ability to cause systemic infection and typhoid fever.

(ii) Adapted fully or predominantly to larger animals: *Salmonella* serovars Gallinarum and Pullorum targeting poultry, Dublin affecting cattle, Choleraesuis is associated with pigs, Abortusequi impacting horses, and Abortusovis infecting sheep.

(iii) Broad host range of animals: *Salmonella* serovar Typhimurium, Heidelberg, Enteritidis and Newport. These are categorised as non-typhoidal *Salmonella* (NTS) serovars in humans as they do not spread systemically and cause typhoid fever.

1.2 Versatility of *Salmonella* as a proficient pathogen

The ability of *Salmonella* to cause disease stems from a range of virulence factors that facilitate its colonisation and invasion of host tissues, as well as its ability to evade host immune responses (M. Wang, Qazi, Wang, Zhou, & Han, 2020). Furthermore, to live

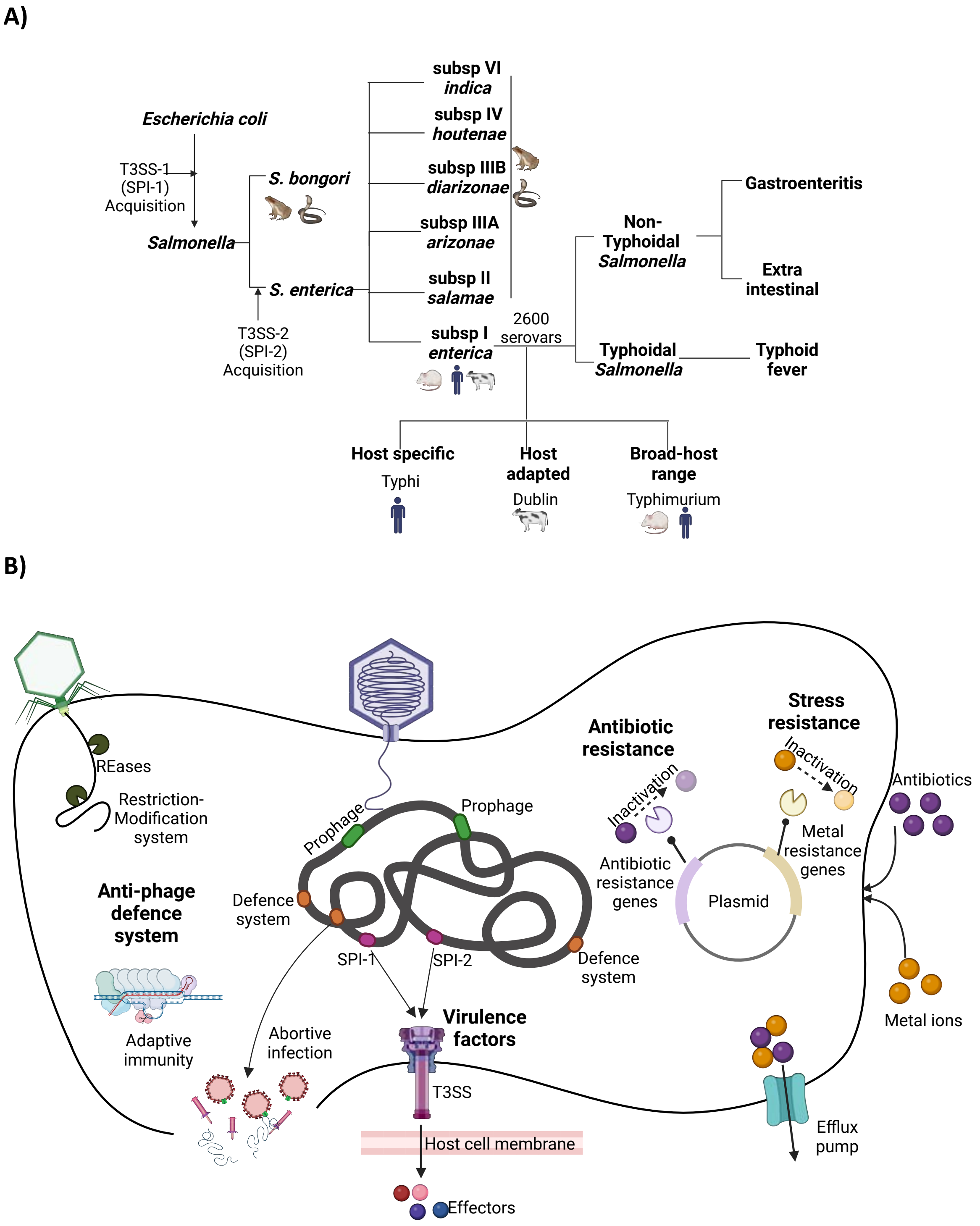


Figure 1.1 Classification and pathogenic determinants of *Salmonella*. A) Classification of *Salmonella. Salmonella* genus includes two species, *S. bongori* and *S. enterica.* Within S. *enterica*, six subspecies exist- I *enterica,* II *salamae,* IIIa *arizonae,* IIIb *diarizonae,* IV *hountenae,* and VI *indica.* Serovars of subspecies *enterica* are further categorised into typhoidal and non-typhoidal serovars. **B)** Pathogenetic determinants in *Salmonella*- virulence factors, antibiotic resistance genes, stress resistance genes and anti-phage defence systems.

in the nonhost environment, *Salmonella* displays remarkable resilience in adverse conditions such as low pH, high temperatures, and exposure to antimicrobials, thanks to various stress response genes (Andino & Hanning, 2015). It has also developed multiple defence strategies against foreign plasmids and bacteriophage infections (Bernheim & Sorek, 2020). Moreover, the therapeutic use of antibiotics against these bacteria has led to the rise of antimicrobial-resistant *Salmonella* strains (V T Nair, Venkitanarayanan, & Kollanoor Johny, 2018).

Previous data supported that genome plasticity, the bacterium's capacity to undergo rapid genetic changes, contributed to the divergence of *Salmonella* strains, allowing them to adapt swiftly to varying conditions (Ferreira, Buckner, & Finlay, 2012). Genome plasticity is facilitated by horizontal gene transfer (HGT), genetic recombinations, and mutations (Dobrindt, Zdziarski, Salvador, & Hacker, 2010). It enables *Salmonella* to acquire new genes, lose unnecessary ones, and modify existing genes. This dynamic genetic landscape facilitates the evolution of *Salmonella* to adjust its virulence factors, stress response mechanisms, and metabolic pathways in response to different environments and challenges (G. R. Liu *et al.*, 2006).

1.2.1 Mobile genetic elements in *Salmonella*

Salmonella's ability to survive in diverse environments and cause a range of infections has evolved as a sophisticated genetic toolkit. Prophages and plasmids stand out as crucial players, shaping its genetic landscape and pathogenic potential.

1.2.1.1 Prophages in *Salmonella*

As per the core genes analysis *Salmonella* phages can be classified into five main groups - P22-like, lambdoid, P27-like, T7-like and P2-like, with three outliers - ε15, KS7, and Felix O1 that are described below (Kropinski, Sulakvelidze, Konczy, & Poppe, 2007; Garcia-Russell, Elrod, & Dominguez, 2009; Wahl, Battesti, & Ansaldi, 2019).

P22-Like Phages: P22, formerly known as PLT 22, is a pioneering model demonstrating the transfer of genetic material between *Salmonella enterica* serovar Typhimurium mutants through generalised transduction. Despite morphological differences, P22 is identified as the archetype of the P22-like phage genus. Other phages in this group, such as ST104, ES18, and ST64T, exhibit unique characteristics. ST104 demonstrates induced broad host range capabilities, ES18 stands out for its distinct receptor and genome structure among transducing phages, and ST64T engages in generalised transduction with serotype-converting capabilities. P22 plays a role in modulating immune functions through dynamic alterations of bacterial lipopolysaccharide (LPS), notably the O-antigen. P22 carries a *gtrABC* operon for O-antigen glucosylation. This operon, encompassing *gtrA* and *gtrB* genes for membrane proteins and a variable *gtrC* gene for specificity, enables glucose attachment at distinct O-antigen sites. LPS undergoes transient surface changes, shaping *Salmonella*'s interaction with the host immune system.

Lambdoid Group: Three lambda-related prophages (Fels-1, Gifsy-1, and Gifsy-2) within the siphovirus family are identified in *Salmonella* genomes. Each prophage integrates into specific host genes and carries potential virulence genes impacting *Salmonella* pathogenesis. Gifsy1 prophage encodes three genes crucial for surviving within cells: *gogB, sarA*, and *pagK2*. GogB encodes an anti-inflammatory effector that mitigates tissue damage during prolonged infections, while short-term inflammation facilitates colonisation in the intestine. SarA is primarily secreted by the SPI-2-encoded T3SS, activating the eukaryotic transcription factor STAT3 inducing the host's anti-inflammatory pathway. PagK2, secreted in outer membrane vesicles, contributes to intracellular survival in macrophages through an unknown mechanism. Gifsy2 prophage encodes for GrvA, an anti-virulence factor responsible for decreasing *Salmonella*'s pathogenicity, probably by affecting resistance to toxic oxygen species.

P27 Group: The P27 group includes phage ST64B, morphologically similar to ST64T, with a 40 kb genome. Despite lacking a tail structure, ST64B shares genetic similarities with Shiga toxin-carrying siphovirus P27 and *Shigella flexneri* phage V. ST64B carries two genes, *sopE* and *sspH2* that play roles in SPI-1 and SPI-2 virulence-associated T3SS.

P2 Group: The P2-like phages, members of the myovirus family, encompass temperate phages like P2, 186, CTX, HP1, HP2, PSP3, and SopEφ. Fels-2, a prophage in *Salmonella* Typhimurium LT2, integrates into the host *ssrA* genes and carries a DAM methylase gene. SopEφ plays a role in *Salmonella*'s infection mechanism.

***T7 Group*:** T7's key role lies in its lytic life cycle, where it efficiently replicates by utilising host machinery. Their contributions to *Salmonella*'s virulence remain unknown.

1.2.1.2 Plasmids in *Salmonella*

Among the various types of plasmids found in *Salmonella*, IncA/C, IncF, IncHI, and IncI1 are prominent classes with distinct characteristics and functions (Rychlik, Gregorova, & Hradecka, 2006 McMillan, Jackson, & Frye, 2020; Robertson, Schonfeld, Bessonov, Bastedo, & Nash, 2023).

IncA/C Plasmids: IncA/C plasmids are notable for their large size, low copy number, broad host range, and frequent inclusion of antibiotic-resistance genes. *Salmonella* strains carrying IncA/C plasmids often exhibit resistance to multiple classes of antibiotics. These plasmids are found in multiple serovars preferably serovar Newport and cattle-specific serovars.

IncF Plasmids: IncF plasmids, characterised by their large size, low copy number, and host restriction to *Enterobacteriaceae*, play a crucial role in *Salmonella* virulence. These plasmids often carry virulence-associated genes, including *spv*, enhancing the bacteria's ability to cause infections. *Salmonella* virulence plasmids (pSV) are identified in *S. enterica* subsp. *arizonae* and *S. enterica* subp. *enterica* serovar Typhimurium, Sendai, Dublin, Enteritidis, Choleraesuis, Gallinarum and Pullorum (Libby *et al.*, 2002). It is heterogeneous in size (50-285 kb) but possesses a 7.8 kb region containing a *spvRABCD* operon essential for bacterial proliferation in endothelial cells and systemic infection. Other loci, the *pef* (fimbrial operon) and *rck* (resistance to complement killing) are sometimes found in the pSV of some strains (Silva, Puente, & Calva, 2017). IncF plasmids also carry antibiotic-resistance genes, those conferring resistance to fluoroquinolones.

IncHI Plasmids: First identified in *Salmonella* Typhi, IncHI plasmids are classified into three groups: HI1, HI2, and HI3. These plasmids are generally large, conjugative, and can contain up to 300 kb. IncHI plasmids often carry heavy metal resistance genes and are associated with antibiotic resistance, including genes for chloramphenicol, streptomycin, sulfonamides, and β-lactams.

IncI1 Plasmids: IncI1 plasmids, which are large, conjugative, and restricted to *Enterobacteriaceae*, exhibit a well-conserved genetic structure with variable accessory gene regions. IncI1 plasmids are classified into numerous sequence types using a pMLST system that relies on the genes *pilL* (pilus biosynthesis), *sogS* (primase), *ardA* (restriction-modification enzyme), *repI1* (RNAI), and a region situated between the *trbA* and *pndC* genes. They play a significant role in disseminating β-lactamase genes. IncI1 plasmids are frequently associated with antibiotic resistance in *Salmonella* strains, particularly those linked to poultry-related outbreaks.

1.2.2 Pathogenicity islands and virulence factors in *Salmonella*

1.2.2.1 Pathogenicity islands

Most of the genomic components responsible for *Salmonella*'s virulence are present as gene clusters within its chromosomal structure, forming designated regions termed *Salmonella* pathogenicity islands (SPI) (Groisman & Ochman, 1996; Marcus, Brumell, Pfeifer, & Finlay, 2000). *Salmonella* reportedly acquired these islands through intricate mechanisms of HGT (Vernikos & Parkhill, 2006), as evidenced by the deviation in the GC content of these regions from the average genomic composition (Groisman & Ochman, 1996). Additionally, the presence and association of mobilome genes and prophage segments within SPI plausibly suggest origins from extraneous sources, such as divergent bacterial species, bacteriophages, or plasmids (Groisman & Ochman, 1996; Ochman & Groisman, 1996; Sabbagh, Forest, Lepage, Leclerc, & Daigle, 2010).

The acquisition of SPI-1, a 40 kb fragment, marks a distinctive occurrence in the evolutionary progression of *Salmonella*, leading to its divergence from the shared ancestor with *E. coli* (Bäumler, 1997). *Salmonella enterica* diverged from *S. bongori*, by acquiring SPI-2 (**Fig. 1.1**). SPI-1 encompasses 39 genes that encode components of the T3SS-1, including exporter apparatus (encoded by *prg/org* and *inv/spa* operon), needle complex (composed of SipB, SipC, and SipD), secreted effectors (like Avr, Sips, and SptP), chaperons (SicA, InvB, and SicP), and regulators (HilA, HilC, HilD, and InvF). This system facilitates pathogen entry into host cells through membrane ruffling and cytoskeleton remodelling. In the intestinal environment, SPI-1 induces inflammation, aiding *Salmonella* to out-compete the gut microbiota by generating specific electron acceptors (Lou, Zhang, Piao, & Wang, 2019). During the proliferation phase inside the host, *Salmonella* switches to the SPI-2 secretion system while in the *Salmonella*-containing vacuole. SPI-2, divided into 15 kb and 25 kb segments, encodes genes for virulence and tetrathionate metabolism. The SPI-2 consists of categories like secretion system apparatus (*ssa*), secretion system effectors (*sse*), secretion system regulators (*ssr*) and secretion system chaperons (*ssc*). The *ssa* genes encode effector proteins that are responsible for encoding the structural components of the needle complex. The *sse* genes encode effector proteins and once inside the host cell, these effectors play roles in manipulating various cellular processes, such as preventing the activation of the host immune system. The *ssr* genes encode proteins that act as regulators, orchestrating the expression of both *ssa* and *sse* to finely control the production of their respective proteins. They regulate the timing of Sse release, ensuring that the T3SS activation aligns with the correct phase of the infection process. The *ssc* genes encode chaperone proteins, that facilitate the proper folding and stabilisation of effector proteins during their transport through the bacterial cytoplasm (Buckner, Croxen, Arena, & Finlay, 2011; Jennings, Thurston, & Holden, 2017). Hence, the intricate interplay of SPI-1 and SPI-2 genes holds a pivotal role in *Salmonella*'s replication and systemic dissemination by orchestrating the precise functioning of the T3SS and the activities of its associated effectors.

In addition to SPI-1 and SPI-2, *Salmonella* contains 22 more pathogenicity islands (SPI-3 to SPI-24), aiding its ability to cause infection (Fookes *et al.*, 2011; Hayward *et al.*, 2014; Urrutia *et al.*, 2014). However, the role in virulence has been verified only for some pathogenicity islands (**Table 1.1**) (Sabbagh *et al.*, 2010; Cheng, Eade, & Wiedmann, 2019).

1.2.2.2 Virulence Factors

Fimbriae or pili: Fimbriae are proteinaceous surface structures made of fimbrins arranged in a helical pattern (Collinson *et al.*, 1996). A particular fimbrial gene cluster (FGC) encodes proteins necessary for forming these fimbriae. FGCs usually consist of 4-15 genes, and *S. enterica* strains, on average, exhibit 12 FGCs (Nuccio & Bäumler, 2007). *Salmonella* utilises three distinct routes for fimbrial assembly: the chaperone/usher (CU) pathway, the nucleation/precipitation pathway to assemble curli fimbriae, and the type IV pathway (Fronzes, Remaut, & Waksman, 2008). Fimbriae play a significant role in pathogenesis, and different *Salmonella* serovars contain various combinations of fimbrial genes (Humphries *et al.*, 2003). Their functions include adherence to cells and inert surfaces, facilitating biofilm formation, colonisation, and evasion of the host immune system (Althouse, Patterson, Fedorka-Cray, & Isaacson, 2003; White, Gibson, Collinson, Banser, & Kay, 2003; Daigle, 2008).

Flagella: *Salmonella*'s flagella is a long filamentous structure consisting of basal body rings, an axial structure including a rod as a drive shaft, a hook acting as a universal joint, and a filament as a helical propeller (Horváth *et al.*, 2019). *Salmonella* have multiple

Table 1.1 Overview of *Salmonella* pathogenicity island

SPI	Approx. Size (kb)	Features	Centisome Location
SPI-1	40	Encodes a T3SS essential for bacterial-mediated enterocyte invasion and intestinal epithelial invasion	63
SPI-2	40	Encodes a T3SS crucial for surviving within macrophages and initiating systemic infection	31
SPI-3	36	Essential for *Salmonella*'s viability within the intracellular phagosomal environment during periods of nutritional deprivation	82
SPI-4	24	Required for adhesion to epithelial cells and gastrointestinal inflammation	92
SPI-5	8	Encodes effector proteins associated with SPI-1 and SPI-2 encoded T3SS	25
SPI-6	59	Encodes the type VI secretion system	7
SPI-7	134	Encodes for Vi antigen and constitutes *pil* gene cluster that encodes for putative virulence factors	-
SPI-8	8	Improves bacterial fitness during infection in humans	-
SPI-9	16	Encodes for virulence factors of type I secretion system	-
SPI-10	33	Responsible for attenuation of virulence	93
SPI-11	10	Includes the PhoP-activated genes *pagD* and *pagC* involved in intramacrophage survival	-
SPI-12	6.3	Required for systemic infection of mice	48
SPI-13	25	Involved in systemic infection of mice and replication inside murine macrophages	67
SPI-14	9	Associated with virulence by mediating invasion	19
SPI-15	6.5	Unknown	-
SPI-16	4.5	Required for intestinal persistence	-
SPI-17	5	Encodes genes responsible for LPS modification	-
SPI-18	2.3	Contains genes controlled by the virulence-related regulator PhoP	-
SPI-19, SPI-20, SPI-21 and SPI-22	45, 34, 55 and 20	Encodes the type VI secretion system	-
SPI-23	37	Plays a role in adherence and invasion of porcine tissues	-
SPI-24	25	Plays a role in fibronectin binding, murine intestinal colonisation, and intramacrophage survival	-

randomly positioned surface flagella comprised of numerous flagellin molecules (Dauga, Zabrovskaia, & Grimont, 1998). The flagella are responsible for the bacterium's motility, adhesion, biofilm formation and triggering immune responses in host cells (Elhadad, Desai, Rahav, McClelland, & Gal-Mor, 2015).

Siderophore: Iron plays a pivotal role in both bacteria and host cells. Although host cells possess ample iron, they are often sequestered in a form hardly accessible to the bacteria. *Salmonella* has evolved mechanisms to scavenge the sequestered iron by synthesising siderophores (enterobactin and salmochelin) that bind to iron ions within the surrounding environment, enabling their growth and survival (Mey, Gómez-Garzón, & Payne, 2021).

Toxins: *Salmonella* is known to produce both exotoxins and endotoxins. Exotoxins are further categorised into cytotoxins, which kill mammalian cells (Ashkenazi, Cleary, Murray, Wanger, & Pickering, 1988) and enterotoxins, which target the intestine. Examples of cytotoxin and enterotoxin include cytoxin styphnolysin, enterotoxin A (Stn) and enterotoxin B (SenB), respectively. The specific role of Stn in *Salmonella*'s pathogenesis remains unclear (Nakano et al., 2012). On the other hand, the endotoxin/LPS is composed of lipid A, core polysaccharide, and O-Antigen (Hitchcock *et al.*, 1986). LPS triggers the host's inflammatory immune responses (Buyse *et al.*, 2007).

Along with the abovementioned factors, *Salmonella* employs a range of mechanisms that contribute to adhesion, immune evasion, and infection establishment.

1.2.3 Environmental stress response factors in *Salmonella*

Salmonella demonstrates impressive adaptability to various environmental factors, ranging from pH fluctuations and temperature variations to antimicrobial peptides, nutrient scarcities, biocides, heavy metals, osmolarity changes, and redox shifts (**Fig. 1.1**). The responses to these stresses are regulated by alternative sigma factors, two-component signal transduction systems and transcriptional regulators (Michael, 2012).

Salmonella needs ions of metals like iron, zinc, copper, manganese, etc., for multiple physiological functions. However, excess or limited amounts of these can induce stress. *Salmonella* manages the stress due to the limitation of metal ions by expressing their respective transporters and scavenging molecules. For example, *Salmonella* produces siderophores to scavenge iron and express Mnt and Mgt transport systems to import Mn^{2+}, Fe^{2+}, and Mg^{2+} ions, respectively (Cunrath & Palmer, 2021). To overcome metal toxicity, it possesses CBA efflux systems, tripartite protein complexes that expel metal ions from cell compartments into the external environment, helping to effectively regulate their intracellular levels (Pontel, Audero, Espariz, Checa, & Soncini, 2007).

Salmonella employs DNA-binding proteins to repress transcription of multiple stress response genes until specific environmental conditions are met, thereby conserving energy by ensuring gene activation only when necessary (Lewis *et al.*, 1996). It also uses multiple promoters, sensor adaptability, counter-silencing mechanisms, and signalling cascades to navigate complex and seemingly unrelated environmental cues (Erickson & Gross, 1989; Bang, Frye, McClelland, Velayudhan, & Fang, 2005; Perez & Groisman, 2007). All these mechanisms fine-tune gene expression, reinforcing the systems adopted to thrive in unpredictable environments.

1.2.4 Anti-phage defence systems

In the natural environment, *Salmonella* is also attacked by bacteriophages (**Fig. 1.1**). Though not well characterised, *Salmonella* has evolved strategies/tools to tackle these attacks, including flagellar phase variation and O-antigen regulation (Kim & Ryu, 2012). A recent study on 1,564 *S.* Typhimurium identified at least eight anti-phage defence systems, with nucleic acid degradation and abortive infection systems being the most prevalent (Woudstra & Granier, 2023). These include Restriction-Modification (R-M), Bacteriophage Exclusion (BREX), phage anti-restriction-induced system (PARIS), Retron, and abortive infection (Abi). The R-M system in bacteria involves restriction enzymes recognising and cleaving foreign DNA while modification enzymes protect the bacterial DNA by adding methyl groups to its recognition sites (Oliveira, Touchon, & Rocha, 2014). BREX system involves a six-gene cassette and defends against a wide range of phages by allowing adsorption but hindering DNA replication (Barrangou & van der Oost, 2015; Goldfarb *et al.*, 2015). BREX type I, PARIS, Gabija, ietAS and AbiD systems were usually associated with integrases and were predominately found in MGEs (Woudstra & Granier, 2023).

In general, bacteria exhibit a panoply of defence mechanisms to counter phage assaults, including innate and adaptive systems, chemical defence, abortive infections, signalling systems, defence systems with homology to human innate immunity genes, toxin-anti-toxin systems and various other systems of unknown mechanisms (Doron *et al.*, 2018; Bernheim & Sorek, 2020; Gao *et al.*, 2020; Millman *et al.*, 2022). As a countermeasure, bacteriophages have developed diverse tactics, including rapid mutation, lytic enzymes, and lysogenic integration, to overcome bacterial defences and ensure their replication (Egido, Costa, Aparicio-Maldonado, Haas, & Brouns, 2022).

The bacterial adaptive anti-phage system is a recent and extraordinary revelation that challenges the conventional perception of bacteria as basic, single-celled entities with limited defence capabilities (Barrangou *et al.*, 2007). In contrast to innate immune systems, adaptive immunity in bacteria closely resembles the immune systems found in complex organisms such as animals (Netea, Schlitzer, Placek, Joosten, & Schultze, 2019). Within this mechanism, bacteria can "remember" prior encounters with pathogens, enabling them to formulate targeted counteractions when re-exposed. A prominent illustration of bacterial adaptive immunity is the clustered regularly interspaced short palindromic repeats (CRISPR) /CRISPR associated (Cas) system (Barrangou *et al.*, 2007).

1.2.4.1 Overview of the CRISPR-Cas system

The CRISPR-Cas system was initially identified in 1987 as an "unusual structure" containing repeats alternated with spacers of unknown function at the 3′ end of the *iap* gene locus of *E. coli* (Ishino, Shinagawa, Makino, Amemura, & Nakata, 1987) and named subsequently (**Fig. 1.2A**). Later (2005-2007), the CRISPR-Cas system was proposed to act as a guardian of the bacterial genome, regulating the tolerance of bacteria against environmental stresses and MGE attacks (**Fig. 1.2A**) (Barrangou *et al.*, 2007).

The CRISPR-Cas system prevails in ~90% archaea and 30-40% bacteria, consisting of three critical attributes - a set of *cas* genes, a leader sequence, and a succeeding CRISPR array (Barrangou *et al.*, 2007). The CRISPR array comprises partially palindromic direct repeat (DR) sequences interspaced by distinct spacer sequences (**Fig. 1.2B**) (Richter, Chang, & Fineran, 2012). The spacers are generally derived from MGEs like the bacteriophages and the plasmids when they first invade the bacteria (Hille *et al.*, 2018). Then onwards, they act as a memory, providing immunity against subsequent attacks by the invading MGE (Hille *et al.*, 2018). According to the 2019 classification of the CRISPR-Cas by Makarova *et al.*, the system is highly diverse and categorised into two classes, six types and 33 subtype (Makarova *et al.*, 2020). Most (~90%) CRISPR-Cas systems belong to

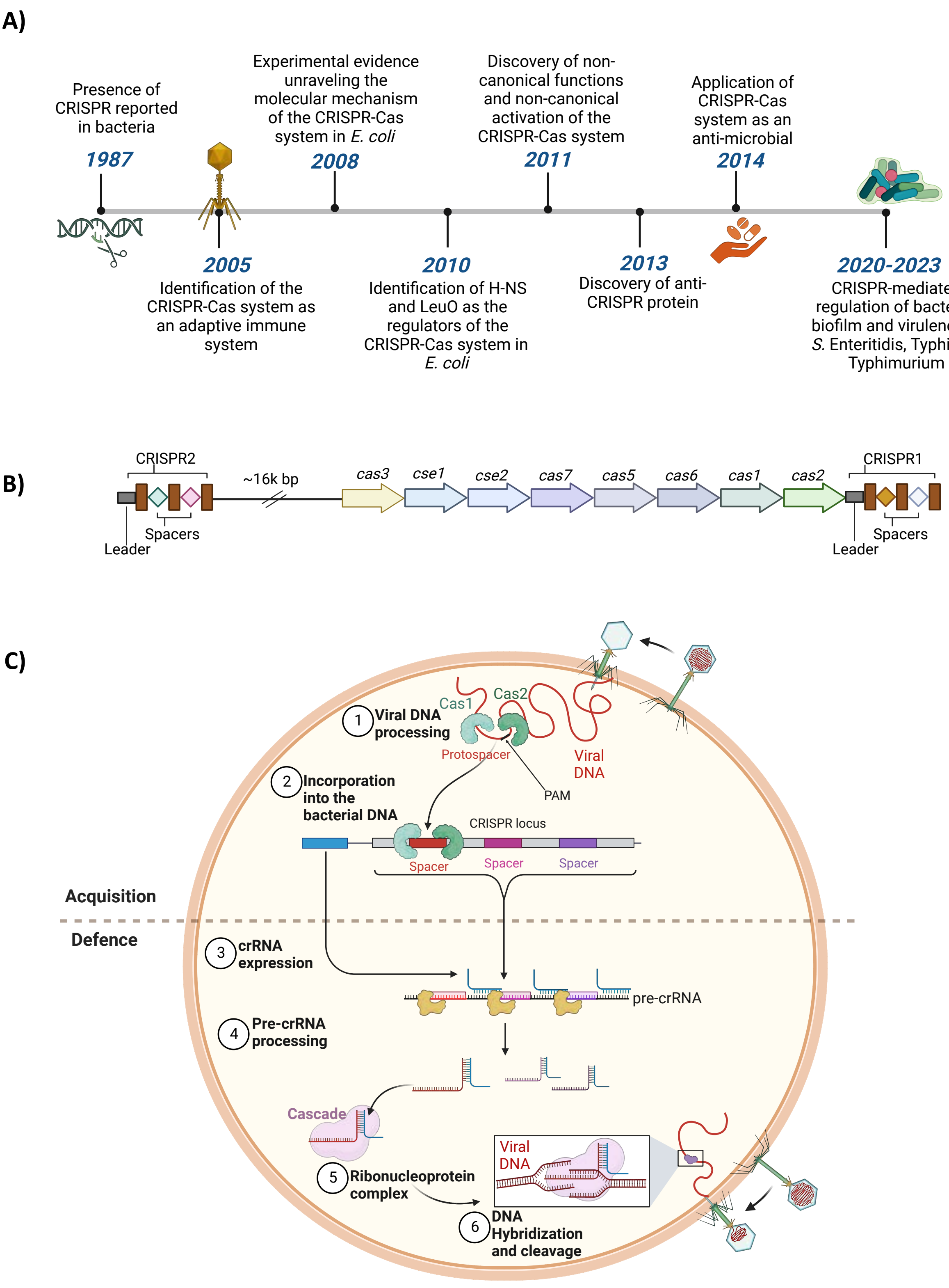

A)
Presence of CRISPR reported in bacteria
1987
2005
Identification of the CRISPR-Cas system as an adaptive immune system
Experimental evidence unraveling the molecular mechanism of the CRISPR-Cas system in *E. coli*
2008
2010
Identification of H-NS and LeuO as the regulators of the CRISPR-Cas system in *E. coli*
Discovery of non-canonical functions and non-canonical activation of the CRISPR-Cas system
2011
2013
Discovery of anti-CRISPR protein
Application of CRISPR-Cas system as an anti-microbial
2014
2020-2023
CRISPR-mediated regulation of bacterial biofilm and virulence in *S.* Enteritidis, Typhi and Typhimurium
B)
CRISPR2
~16k bp
cas3
cse1
cse2
cas7
cas5
cas6
cas1
cas2
CRISPR1
Spacers
Leader
Spacers
Leader
C)
Cas1
Cas2
1 Viral DNA processing
Protospacer
Viral DNA
PAM
2 Incorporation into the bacterial DNA
CRISPR locus
Spacer
Spacer
Spacer
Acquisition
Defence
3 crRNA expression
pre-crRNA
4 Pre-crRNA processing
Cascade
5 Ribonucleoprotein complex
Viral DNA
6 DNA Hybridization and cleavage

Figure 1.2 CRISPR-Cas system. A) Chronological representation of significant milestones in the field of CRISPR-Cas biology. **B)** Arrangement of the CRISPR-Cas system in *Salmonella. Salmonella* comprises two CRISPR loci (CRISPR1 and CRISPR2) and eight *cas* genes. The *cas* locus is in the neighbourhood of the CRISPR1 loci, while the CRISPR2 locus is an orphan. The diamonds represents the spacer sequences, while the rectangles represent the direct repeats (DR). **C)** Mechanism of action of the CRISPR-Cas system in *Salmonella.* The mechanism of action is divided into three stages: adaptation, crRNA biogenesis, and interference.

the class 1 category, exhibiting their effect through multiple subunit effector complexes containing four to seven Cas proteins. Conversely, the less prevalent class 2 system relies on a single multi-domain effector protein (Makarova *et al.*, 2020).

The mechanism of the CRISPR-Cas system can be divided into three stages: adaptation, crRNA biogenesis, and interference (**Fig. 1.2C**) (Xue & Sashital, 2019). During the adaptation step, protospacers (pieces of invading genetic elements) are incorporated into the CRISPR array with the help of Cas proteins. The Cas proteins recognise a distinct small motif, protospacer adjacent motif (PAM), in the invading DNA, thereby cleaving it and incorporating the protospacer in the array (J. Wang *et al.*, 2015). The crRNA biogenesis yields crRNAs guiding the Cas proteins to sequence-specifically target the invading MGEs. The CRISPR array is transcribed into long precursor crRNAs (pre-crRNA) that are further processed into mature crRNAs (Brouns *et al.*, 2008). A single crRNA, comprising a DR and a spacer, acts as a docking centre for a Cascade complex (made of multiple Cas proteins) to bind and form a surveillance complex (Koonin, Makarova, & Zhang, 2017; Xue & Sashital, 2019). Unlike other types, the surveillance complex of the type I system does not perform the interference step by itself (Westra *et al.*, 2012; Hochstrasser *et al.*, 2014; Redding *et al.*, 2015). The Cas3 nuclease is recruited after accurate target recognition, thereby targeting and cleaving the invader MGE (**Fig. 1.2C**) (Xue & Sashital, 2019).

1.2.4.2 CRISPR-Cas system in *Salmonella*

Salmonella contains the type I-E CRISPR system comprising eight *cas* genes and two CRISPR loci (CRISPR1 and CRISPR2) (**Fig. 1.2B**) (Shariat, Timme, Pettengill, Barrangou, & Dudley, 2015). Typically, the *cas* locus is in the neighbourhood of the CRISPR1 loci, while the CRISPR2 locus is an orphan (Shariat *et al.*, 2015; Tanmoy *et al.*, 2020). Over 7,500 spacers have been detected in *Salmonella* (Zhang *et al.*, 2021). A study by Pettengill *et al.*, on 431 *Salmonella* strains revealed two *cas* profiles, 878 CRISPR1 and 1,241 CRISPR2 unique spacers. However, only ~75% had complete *cas* genes, while ~2.3% had no *cas* genes (Pettengill *et al.*, 2014). Later, in 2015, an analysis of over 600 *Salmonella* strains belonging to four serovars, Typhimurium, Heidelberg, Enteritidis, and Newport, by Shariat *et al.*, identified 179 unique spacers and a distinct CRISPR1 leader for serovar Newport II. Further, the authors speculated that the CRISPR system is not immunogenic, probably having auxiliary functions (Shariat *et al.*, 2015).

Tanmoy *et al.*, analysed 1,059 serovar Typhi isolates identifying 1,919 CRISPR array while grouping them into two types, group-A (evidence score 3/4) and group-B (evidence score 1/2) based on the evidence score for CRISPR detection (Tanmoy *et al.*, 2020). However, Fabre *et al.*, indicate contamination of ~47% isolate genomes with serovars Enteritidis, Paratyphi A and Worthington, as well as the differences in the CRISPR/DR sequences and the CRISPR loci presented by Tanmoy *et al.*, thus, explaining the discrepancies in the CRISPR profiles and loci reported for serovar Typhi (Fabre, Njamkepo, & Weill, 2021). Nonetheless, some interesting observations were reported by Tanmoy *et al.,* The protospacers for group-A loci were in phage sequences, whereas for group-B loci, they were in plasmid sequences. The predicted PAM sequence (TTTCA/T) identified for Typhi serovars was distinct from that (AWG) of serovar Typhimurium and *E. coli* (also contains the type I-E CRISPR-Cas system). Of the identified spacers among 1,919 CRISPR loci, only 47 spacers were unique, and a few had 100% identity to the phage, plasmid, viral and antimicrobial resistance-related gene sequences.

The CRISPR-Cas system of serovar Typhimurium is predicted to encode three transcriptional units defined by three promoters: P_{cas3}, P_{casA}, and P_{CRISPR} (Dillon *et al.*, 2012). In contrast, serovar Typhi has five transcriptional units encoding *cas3,* sense *cse2* (*scse2*), anti-sense *cas2-cas1* (*ascas2-1*), anti-sense *cse2-cse1* (*ascse2-1*), *and cse1–cse2–cas7–cas5–cas6e–cas1–cas2*-CRISPR (*cas*-CRISPR operon) (Medina-Aparicio *et al.*, 2017). Intriguingly, *cas* genes of other CRISPR-Cas types like DEDDh, DinG (type IV-A), and WYL (type-I system) were reported in the Typhi isolates (Tanmoy *et al.*, 2020). Reportedly, the WYL domain transcriptionally regulates the CRISPR-Cas system. It is predicted that the DEDDh exonuclease domains (that can fuse with *cas1* and *cas2*) could compensate for the shorter *cas3* (an exonuclease) gene in this serovar (Makarova, Anantharaman, Grishin, Koonin, & Aravind, 2014).

1.2.4.3 Mechanism of CRISPR-Cas regulation in *Salmonella*

In *Salmonella*, LeuO, histone-like nucleoid structuring protein (H-NS) and leucine responsive regulatory protein (LRP) regulate the CRISPR-Cas expression (**Fig. 1.3**) (Medina-Aparicio *et al.*, 2011). Both H-NS and LRP simultaneously bind upstream and downstream of the transcription initiation site of the *cas* gene, possibly forming a nucleosome structure. This could promote the repression (like that of 16S rRNA) of the CRISPR-Cas system (Medina-Aparicio *et al.*, 2011). H-NS binding reduces the access of RNA polymerase to the promoter, thereby inhibiting the transcription of genes like *casA* (*cse1*) and *crispr1* (Y. Liu, Chen, Kenney, & Yan, 2010). It is hypothesised that H-NS on invasion binds to MGEs with high AT content (Navarre *et al.*, 2006; Richter *et al.*, 2012). According to the model, the binding of LeuO triggers fine-restructuring of the nucleoprotein complex, thereby surmounting the H-NS mediated repression without ripping it off from the DNA. Here, the binding of LeuO to the two binding sites loops out the DNA containing the H-NS behind the LeuO barrier. This interferes with H-NS activity, thus preventing obstructions of a nearby promoter(s) and inducing gene expression (Dillon *et al.*, 2012). Nevertheless, the natural growth conditions activating the CRISPR-Cas system in *Salmonella enterica* and *E. coli* are unknown, and *leuO* expression is also low under standard laboratory conditions (Guadarrama, Medrano-López, Oropeza, Hernández-Lucas, & Calva, 2014).

In *S. enterica* subsp. *enterica* serovar Typhi, LeuO binds to *cse1* and *cas3* promoters, while in serovar Typhimurium, it binds to the CRISPR promoter with negligible binding to *cse1* and *cas3* promoters (Dillon *et al.*, 2012). However, when present in higher concentrations, LeuO regulates both *cse1* and *cas3* expression of *S.* Typhimurium. Under conditions mimicking the intra-macrophage environment, the system is activated in a LeuO-independent manner, at least in *S.* Typhi (Medina-Aparicio *et al.*, 2011). Introduction of LacI repressor (absent from *Salmonella* genome) in *S.* Typhimurium induced the expression of *cas* genes indicating direct/indirect regulation by LacI repressor (Eswarappa, Karnam, Nagarajan, Chakraborty, & Chakravortty, 2009; Louwen, Staals, Endtz, van Baarlen, & van der Oost, 2014).

1.2.4.4 Association of the CRISPR-Cas system in endogenous gene regulation

Recent studies hint at the involvement of the CRISPR-Cas system in regulating bacterial physiology, virulence, and biofilm (Cui *et al.*, 2020; Stringer, Baniulyte, Lasek-Nesselquist, Seed, & Wade, 2020; Medina-Aparicio *et al.*, 2021; Sharma, Das, Raja, & Marathe, 2022).

The Cas3 nuclease of *S. enterica* subsp. *enterica* serovar Enteritidis is observed to influence its virulence by regulating key T3SS genes, its effectors, and chaperones (**Fig. 1.3**) (Cui *et al.*, 2020). In the *cas3* knockout strain, the fimbrial subunit genes are downregulated, while the biofilm-dependent modulation protein is upregulated, thereby

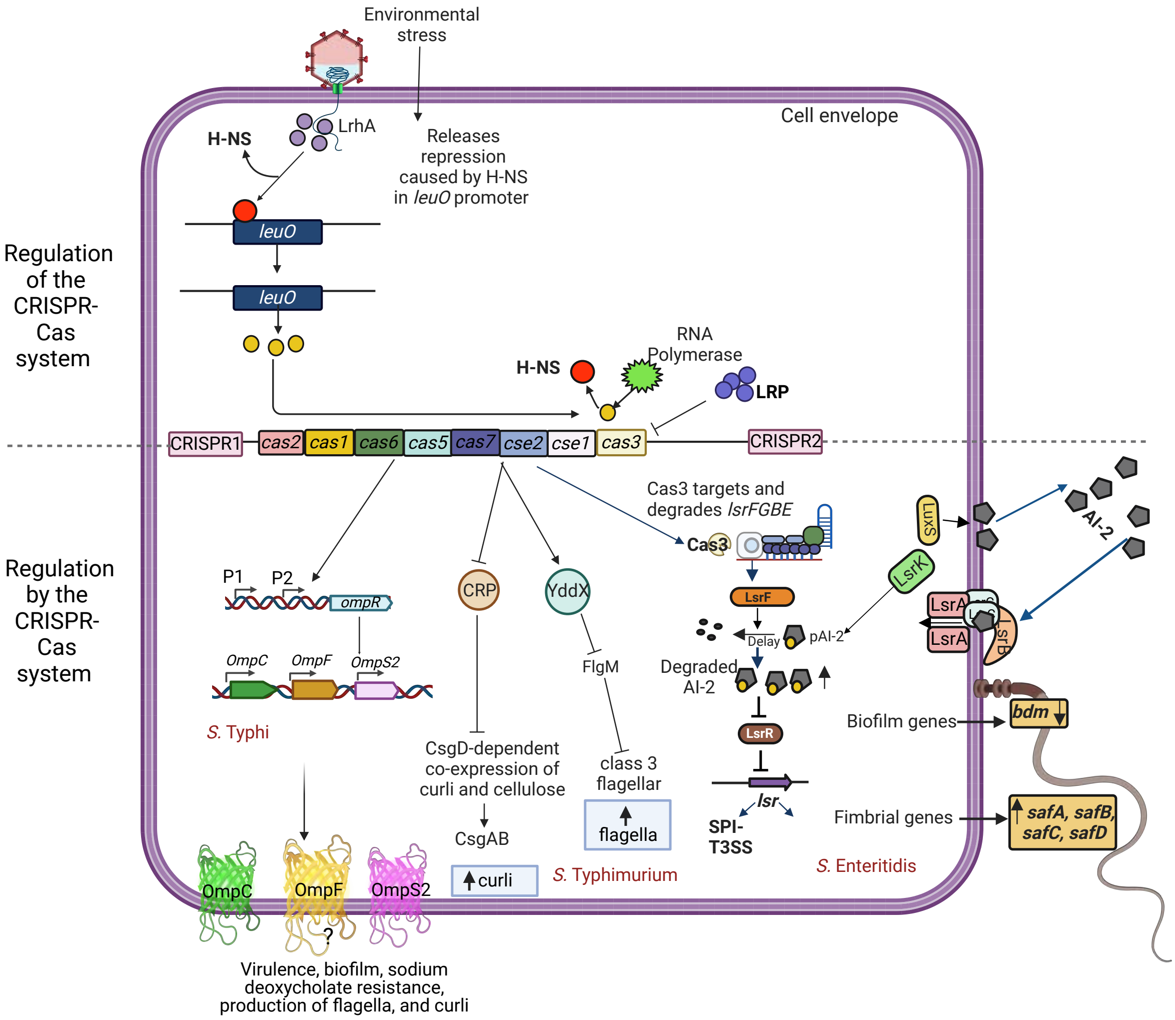


Figure 1.3 Mechanism of CRISPR-Cas regulation in *Salmonella*. H-NS, LeuO and LRP regulate the expression of the CRISPR-Cas system. Endogenous gene regulation by the CRISPR-Cas system in *Salmonella enterica* subsp. *enterica* serovar Enteritidis, Typhi and Typhimurium.

reducing biofilm formation. The CRISPR-Cas system (especially *cas3*) is believed to regulate the LuxS/AI-2 type quorum sensing (QS) system by silencing lsrF-mRNA that degrades auto-inducer-2. The active QS system enhances the expression of T3SS and biofilm-related genes. This supports virulence and biofilm formation, thereby explaining the observed effects of *cas3* mutant (**Fig. 1.3**) (Cui *et al.*, 2020).

Different studies *in S. enterica* subsp. *enterica* serovar Typhi predicts the role of CRISPR-Cas in endogenous gene regulations. The *cas* expression was observed in bacteria within human macrophages (Faucher, Curtiss, & Daigle, 2005) and conditions of pH (7.5) identical to the distal ileum, a colonising site of this bacteria (Medina-Aparicio *et al.*, 2017). CRISPR-Cas reportedly regulated outer membrane proteins, OmpC, OmpF, and OmpS2 *via* OmpR (**Fig. 1.3**) (Medina-Aparicio *et al.*, 2021). The authors suggested that Cas proteins associate in different combinations to form diverse protein complexes. These complexes bind and influence the *ompR* mRNA stability, thereby modulating OmpF, OmpC, or OmpS2 differently. In addition, the authors report the sensitivity of *crispr* and *cas* null mutants to human bile salt while showing enhanced biofilm formation. This suggests that the CRISPR-Cas system negatively regulates biofilm genes. The authors concluded the moonlighting of Cas proteins acting in diverse combinations by controlling the *omp* RNA or binding to and tweaking the *ompR* promoter (**Fig. 1.3**) (Medina-Aparicio *et al.*, 2021).

Sharma *et al.*, explored the roles of the CRISPR-Cas system of *S.* Typhimurium in biofilm formation by knocking out various components of the system, Δ*crisprI,* Δ*crisprII,* ΔΔ*crisprI crisprII,* and Δ*cas op* (Sharma *et al.*, 2022). The study concluded that the CRISPR-Cas system positively modulates the surface-attached biofilm formation while negatively modulating the pellicle-biofilm (**Fig. 1.3**). The results contradict previously reported studies by Cui *et al.*, 2020 and Medina-Aparicio *et al.*, 2021 on *Salmonella enterica* serovars Enteritidis and Typhi, respectively. Sharma *et al.*, attributed the discrepancy to the difference in the CRISPR-Cas arrangement and the knockout strains used, leading to variation in *cas* gene expression. In serovar Typhimurium, a complete *cas* operon was deleted (Sharma *et al.*, 2022), and in serovar Enteritidis, only *cas3* was deleted with simultaneous upregulation of other *cas* genes (Cui *et al.*, 2020). Thus, both studies ultimately show that Cas inhibits pellicle-biofilm formation. Serovars Typhimurium and Typhi differ in the *cas* gene sequences and arrangements, probably explaining the difference in the surface-attached biofilm regulation by the Cas system (Medina-Aparicio *et al.*, 2021; Sharma *et al.*, 2022).

1.2.5 Treatment strategies for *Salmonella* infections

Generally, antibiotic treatment is unnecessary for NTS infections due to the self-limiting nature of these infections (Antony *et al.*, 2018). However, antibiotic intervention becomes necessary if an NTS infection progresses to conditions like meningitis and septicaemia. Typhoidal *Salmonella* infections are typically dealt with by using cephalosporins such as cefixime, cefotaxime, or ceftriaxone, as well as chloramphenicol, amoxicillin, trimethoprim/sulfamethoxazole (TMP-SMX), azithromycin, or aztreonam. However, the emergence and dissemination of antibiotic-resistant *Salmonella* strains have introduced a complex dimension to the treatment landscape (Gut, Vasiljevic, Yeager, & Donkor, 2018).

The escalating global concern stems from the increasing resistance rate of *Salmonella* to antibiotics, resulting in heightened health risks (X. Wang *et al.*, 2019). Recent data published in 2018 indicated a 65% increase in antibiotic consumption from 2000 to 2015, with China, India, and Pakistan largely contributing to this surge (Klein *et al.*, 2018). Notably, in the Indian context, the employment of cephalosporin antibiotics against *Salmonella* has exhibited a three to fourfold increase between 2000 and 2014 (Britto, Wong, Dougan, & Pollard, 2018). Alarmingly, research has highlighted patterns of tetra- and penta-drug resistance against commonly available antibiotics (Xiang *et al.*, 2020). Moreover, individual pathogenic strains of *Salmonella* respond diversely to the array of antibiotics. For example, specific serovars such as Typhimurium, Newport, and Heidelberg account for a substantial proportion (about 75%) of antibiotic-resistant infections (Gut *et al.*, 2018). These concerns are compounded by dysbiosis (the perturbation of the gut microbiome) that results from using antibiotics during infancy (Vangay, Ward, Gerber, & Knights, 2015). Dysbiosis has the potential to impede the development of crucial immune system components like Peyer's patches and mesenteric lymph nodes, which play pivotal roles in preventing *Salmonella* infection. Consequently, employing antibiotics for uncomplicated cases of *Salmonella* gastroenteritis is not recommended (Vangay *et al.*, 2015; Bruzzese, Giannattasio, & Guarino, 2018). In light of dysbiosis associated with the use of antibiotics, probiotics offer a promising solution for the prophylactics and therapeutics for salmonellosis (Sanders *et al.*, 2010; Shi, Li, Shen, & Sun, 2020). However, selecting appropriate probiotic strains is important, given their specific biogeography and strain-specific activity. Using the wrong strain may not yield benefits. The safety of probiotics, especially in immunocompromised individuals, is a concern, as it may lead to cases of septicemia. Phage therapy involving the use of bacteriophages or phage cocktails to target and kill *Salmonella* is another alternative to antibiotics (Khan & Rahman, 2022). However, it has limited applications considering the potential for bacterial resistance, regulatory hurdles, limited clinical data, sensitivity to environmental conditions, dosing complexities, potential side effects, and ethical and legal considerations (Lin, Du, Long, & Li, 2022). These collectively pose challenges to its widespread adoption and effectiveness in treating *Salmonella* infections.

Considering the factors mentioned above, it becomes imperative to formulate innovative antimicrobial strategies capable of effectively addressing both antibiotic-sensitive and antibiotic-resistant strains of *Salmonella*. This endeavour may encompass exploring alternative treatment methodologies, such as harnessing the CRISPR-Cas system to eliminate *Salmonella* selectively (Gomaa *et al.*, 2014). The system, renowned for its precision and adaptability (Xue & Sashital, 2019), can potentially emerge as a focused and promising approach against antibiotic-sensitive and resistant strains of *Salmonella*. Other treatment methods discussed above generally work at the level of entire organisms or broad bacterial populations. But the CRISPR-Cas system enables targeted and specific alterations to an organism's DNA with unparalleled accuracy (Gomaa *et al.*, 2014). Thus, making it potentially effective to target *Salmonella* precisely while preventing disruptions to the intricate equilibrium of the gut microbiota. This fidelity may also allow us to potentially combat antibiotic resistance by selectively targeting resistance genes (Tao, Chen, Li, & Liang, 2022).

1.3 Gaps in existing research and objectives of the thesis

The scientific literature reveals the presence of diverse genes in *Salmonella*, enhancing its proficiency as a pathogen. Reportedly, the *Salmonella* subspecies have a well-conserved genome structure. Nevertheless, a few serovars, like the host-adapted serovar Typhi, have incredible variations in genome structures with different arrangements of DNA segments (G.-R. Liu *et al.*, 2005). Phylogenetic and genomic analyses of diverse *Salmonella* serovars reveal substantial inter- and intra-serovar genomic variability, which contributes to genomic plasticity (Chan *et al.*, 2003; W. Q. Liu *et al.*, 2007; Mastrorilli *et al.*, 2020). This variability within *Salmonella* strains can result in the presence of novel (accessory) genes in certain individuals while absent in others. Pangenome analysis facilitates the exploration of bacterial evolution within a species by assessing the core and accessory genes across multiple genomes. Researchers have analysed *Salmonella*'s pangenome with limited strains and serovars (Laing, Whiteside, & Gannon, 2017; Vila Nova *et al.*, 2019; Vaid, Thakur, Anand, Kumar, & Tripathi, 2021; Turcotte *et al.*, 2022). The limitations of these studies include (i) selection bias in strain representation that could inadvertently skew our perception of the species genetic diversity, (ii) constrained grasp on the understanding of the *Salmonella*'s evolutionary history and (iii) limited understanding of the prevalence of gene clusters associated with traits like virulence, antibiotic resistance, stress resilience and anti-phage defence systems. Against this backdrop, we aim to ***study the pangenome of Salmonella to unveil key genomic regions subjected to plasticity while shedding light on the*** **presence** ***of gene clusters associated with pathogenic determinants***.

The adaptability of the bacterial genome is driven by genome plasticity, with the CRISPR-Cas system standing out as a key influencer. Recent research on *E. coli* revealed significant conservation of this system (conserved in ~70% of strains) at a specific hotspot within the core genome (Hochhauser, Millman, & Sorek, 2023). This may play a role in shaping the genome plasticity and underpins bacterial adaptive responses. In *Pectobacterium atrosepticum,* the CRISPR-Cas systems exhibit self-genome targeting, exerting strong selective pressure on the bacterium (Dy, Pitman, & Fineran, 2013). This leads to mutations, genome rearrangements, and deletions of genomic fragments that can even be large-scale DNA deletions, like the pathogenicity islands. Such genome remodelling contributes to adaptation to diverse niches, leading to bacterial evolution and genetic diversity. In *Salmonella*, the structure and evolution of the CRISPR-Cas system have been studied and efficiently used for serotyping different isolates (Touchon & Rocha, 2010; Shariat *et al.*, 2015; Karimi, Ahmadi, Najafi, & Ranjbar, 2018). These studies discuss two patterns of the CRISPR-Cas arrangement, various system attributes (e.g., length and conservation of the leader sequence, spacers, and DR), protospacers, and the correlation between the CRISPR arrays and phylogeny of *S. enterica* isolates. Yet, the potential connection between the attributes of the CRISPR array and species/serovar host range and habitat diversity remains uncharted, as does the correlation of the spacer content with bacterial habitat and its host diversity. Particularly, whether similar environments yield serovars with matching spacers or protospacer sources is still unknown. Therefore, we plan to ***assess the diversity of the CRISPR-Cas system in Salmonella, perform phylogenomics to study the CRISPR diversity and derive any correlations with the serovar diversity. We also aim to inspect the evolutionary trajectory of the CRISPR-Cas system within the Enterobacteriaceae family.***

The study by Touchan and Rocha found that 53% of the CRISPR protospacers in *Salmonella* and *Escherichia* were within the genome (Touchon & Rocha, 2010), implying a potential role of the CRISPR-Cas system in endogenous gene regulation. Furthermore, a computational analysis predicting CRISPR targets in *E. coli* suggests that the type I-E system predominantly targets endogenous genes by attacking the reverse strand of the target DNA (Bozic, Repac, & Djordjevic, 2019). Subsequent evidence on endogenous gene regulation by the type I-E CRISPR-Cas system is provided through the wet lab studies on the system in the *Enterobacteriaceae* family. A *cheY* mutant in *S.* Typhi shows altered *cas* gene expression (Louwen *et al.*, 2014). There are indications that the CRISPR-Cas system in *Salmonella* regulates its physiology, like the biofilm formation and invasion of the host, by regulating the QS and invasion genes (Cui *et al.*, 2020; Medina-Aparicio *et al.*, 2021; Sharma *et al.*, 2022). Moreover, ChIP seq analysis indicated the binding of the Cascade complex to different genome locations in *S.* Typhimurium str. 14028 (Stringer *et al.*, 2020), implying its role in endogenous gene regulation. ***To gain insights into endogenous genes potentially regulated by the CRISPR-Cas system in different Salmonella serovars*, *we aim to do computational analyses to identify (i) if some spacers are self-targeting and (ii) the potential target genes/pathways regulated by the CRISPR-Cas system.*** Analysing self-targeting spacers may shed light on the co-evolution of *Salmonella* and the CRISPR system.

Apart from understanding the role of the CRISPR-Cas system in endogenous gene regulation, researchers are exploring the utilisation of this system as an antimicrobial. Endogenous and heterologous (exogenous system supplied on a plasmid/bacteriophage) CRISPR-Cas systems have been explored to kill the pathogen or eliminate the plasmid containing antibiotic resistance genes (Wu *et al.*, 2021; Tao *et al.*, 2022). The endogenous system was harnessed as an antimicrobial in *E. coli* but in the H-NS (a repressor of the CRISPR-Cas system) null mutant (Gomaa *et al.*, 2014). The heterologously expressed

CRISPR-Cas9 system has been exploited to specifically kill *Salmonella* (Hamilton *et al.*, 2019). They used a plasmid encoding the conjugative machinery and CRISPR component, TevCas9 nuclease, and the guide RNA to target the genome. The authors observed ~100% conjugation frequencies depending on the genes targeted, with killing efficiencies ranging from 1%-100%. However, there are problems associated with the heterologous system: (i) A constitutive or leaky expression of the heterologous Cas9 is toxic to the conjugation donor, thereby reducing the conjugation efficiencies and leading to the selection of inactive CRISPR-Cas9 plasmids (Pursey, Sünderhauf, Gaze, Westra, & van Houte, 2018; Hamilton *et al.*, 2019) (ii) with a heterologous system, a huge-sized DNA (owing to the size of Cas9) is transferred as an antimicrobial tool. Utilisation of endogenous CRISPR-Cas3 system would significantly reduce the size of DNA to be transferred, requiring only the CRISPR array (containing spacers to target *Salmonella*-specific genes) to be supplied in trans. Furthermore, the crRNA vital for Cas9 activity is 20 bp long (Jiang & Doudna, 2017), whereas the Cas3 of *Salmonella* is ~32 bp (Kushwaha, Bhavesh, Abdella, Lahiri, & Marathe, 2020). This would provide leverage to increase the target specificity. Reportedly, the resistance against exogenous Cas9 could be through (i) the occurrence of the anti-CRISPR genes in the target bacteria, (ii) protospacer mutation, and (iii) mutations in the Cas9 nuclease (Uribe *et al.*, 2021). These problems are expected to be rare with the endogenous CRISPR-Cas3 system as it regulates essential physiological functions of the bacteria, including biofilm formation and virulence in *Salmonella*. Moreover, the double-stranded DNA breaks generated by Cas9 can be repaired and thus may result in an inefficient killing (Wimmer & Beisel, 2019). The probable solution is to express a protein that inhibits the repair of cleaved DNA, but this would further increase the DNA size to be transferred. Moreover, Cas3 degrades the DNA away from the targeted region (Caliando & Voigt, 2015), and utilising the endogenous system is expected to have better killing efficiency. On this account, we ***aim to generate a foolproof method to harness endogenous CRISPR-Cas3 to selectively kill Salmonella using a customised CRISPR array targeting its highly conserved essential genes.*** This strategy might be an effective alternative therapeutics to exclusively eliminate antibiotic-sensitive and resistant *Salmonella* while differentiating between beneficial and pathogenic microorganisms.

Given the identified gaps in the literature, we have strategically designed the following research objectives:

1. **Decoding the genome plasticity of *Salmonella* to unravel adaptive mechanisms and functional specialisation.**

   The research outcomes of this objective are discussed in Chapter 2: Within the *Salmonella* genome, the gene clusters are associated with virulence, stress resistance, antibiotic resistance, and anti-phage defence while exhibiting distinct preferences for regions of genome plasticity integrated into specific genomic spots.

2. **Studying the phylogenomics to understand the CRISPR-Cas diversity in *Salmonella.***

   The research outcomes of this objective are discussed in Chapter 3: The CRISPR-Cas system shows differences in its spacers and *cas* genes within subspecies and serovars, possibly rendering a competitive advantage to the bacteria under stressful situations like the presence of antibiotics, different environmental factors and hostile conditions within the host.

3. **Analysing self-targeting CRISPR spacers in *Salmonella* to understand their role in endogenous gene regulation.**

   The research outcomes of this objective are discussed in Chapter 4: The CRISPR spacers conserved in most *Salmonella* strains of serovars Enteritidis, Typhimurium, and Typhi appear to extend their functional impact in diverse cellular processes, including DNA repair, stress response modulation, biofilm formation, and the regulation of pathogenic behaviour.

4. **Investigating the functional activation of the CRISPR-Cas system and repurposing it for *Salmonella*-specific killing.**

   The research outcomes of this objective are discussed in Chapter 5: The CRISPR-Cas system is not induced to detectable levels under lab conditions, conditions mimicking hostile intracellular conditions, but is functionally activated by LeuO, a transcriptional activator. However, we could observe only 25-35% self-killing for different *Salmonella* serovars.

Chapter 2

Decoding the genome plasticity of *Salmonella* to unravel adaptive mechanisms and functional specialisation

Publications from this objective-

1. Kushwaha SK, Anand A, Wu Y, Avila HL, Sicheritz-Ponten T, Millard A, Marathe SA, Nobrega FL. Genomic plasticity is a blueprint of diversity in *Salmonella* lineages. bioRxiv. 2023 Dec 3; DOI: 10.1101/2023.12.02.569618 (Under revision in PLOS Biology).

2.1 Introduction

The interplay between conserved and variable features in bacterial genomes plays a crucial role in shaping the diversity and adaptability of different species (Francino, 2012). Within a species, the core genome, comprising genes universally present, handles essential cellular functions. In contrast, the flexible genome consists of genes that vary between individual strains, allowing bacteria to adapt to specific environments and acquire pathogenic traits (Hacker & Carniel, 2001; Ulrich Dobrindt, Hochhut, Hentschel, & Hacker, 2004; Abby & Daubin, 2007). These variable genes are often organised into regions of genomic plasticity (RGP) (Mathee *et al*., 2008), typically associated with mobile genetic elements (MGEs). These elements serve as potent facilitators for acquiring genes related to virulence, antibiotic and stress resistance, and anti-phage immunity, contributing to the dynamic evolution of the bacterial genome (U. Dobrindt *et al*., 2003; Lin *et al*., 2011; Das *et al*., 2022; Johnson *et al*., 2023). Exploring this genomic plasticity is crucial for understanding bacterial evolution, phylogeny, and pathogenic potential.

Salmonella offers an excellent model for studying these variable genomic features. Its diverse spectrum of species, subspecies, and serovars showcases the inherent flexibility in its genome, a pivotal factor in shaping both the phylogeny and pathogenic potential of *Salmonella* (Fierer & Guiney, 2001; Tanner & Kingsley, 2018; Li *et al*., 2019). Consequently, exploring the genomic plasticity of *Salmonella* becomes a key avenue for gaining insights into its evolution as a pathogen.

To gain a further understanding of the structural and functional features of RGP in *Salmonella*, we carried out a comprehensive analysis of 12,244 *Salmonella* spp. genomes. Our findings revealed that gene clusters associated with virulence, stress resistance, antibiotic resistance, and anti-phage defence exhibit specific preferences for RGP integrated into distinct genomic spots. These preferences seem to be influenced by neighbouring genes that likely share regulatory and functional coordination. The irregular distribution of these genomic spots across diverse *Salmonella* lineages establishes a blueprint for pathogenicity and survival strategies. Deciphering the complex interplay between pathogenicity-related gene clusters and RGP not only improves our understanding of *Salmonella* evolution but also enables us to uncover novel pathogenicity genes, anticipate future adaptations, and identify targets for disease prevention, management, and therapeutic interventions.

2.2 Materials and Methods

2.2.1 Data collection

A total of 16,506 *Salmonella* genomes were downloaded from the PathoSystems Resource Integration Center (PATRIC) (Wattam *et al*., 2014) and NCBI genome databases in May 2021. Duplicate entries from both databases were removed. The completeness of the genome assemblies was assessed using BUSCO (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015), and strains with a recommended quality score of 95 or higher were selected. ANI scores were calculated using FastANI (Jain, Rodriguez-R, Phillippy, Konstantinidis, & Aluru, 2018), comparing each strain against the reference genome of *Salmonella* (*S. enterica* subsp. *enterica* serovar Typhimurium str. LT2, accession nr. CP060507.1). Strains with an ANI score of 90% or higher were retained (Pearce *et al*., 2021). After applying these filters, the final dataset consisted of 12,244 genomes. The MASH tool (Ondov *et al*., 2016; Ondov *et al*., 2019) was used to calculate the mash distance between these strains, with a threshold of 0.1. The serovar identification and country of isolation for each strain were obtained from the information available in the PATRIC and NCBI databases. The host-specificity of the species, sub-species, or strains was determined through an extensive literature review (Uzzau *et al*., 2001; Eswarappa, Karnam, Nagarajan, Chakraborty, & Chakravortty, 2009; Andino & Hanning, 2015; R. A. Cheng, Eade, & Wiedmann, 2019), categorising them as host-specific (human or poultry specific), host-adapted, or broad host range. Serovars with insufficient details were categorised as having an unknown host range.

2.2.2 Phylogeny building and pangenome analysis

The genomes were clustered using the K-mer-based tool PopPUNK v2.5.0 (Lees *et al*., 2019). The model for *Salmonella* was fitted using dbscan, and the phylogeny was visualised using Microreact (Argimón *et al*., 2016) and iTOL (Letunic & Bork, 2019). The pangenome analysis was performed using PPanGGOLin v1.2.74 (Gautreau *et al*., 2020), and the pangenome graph was visualised using Gephi software (https://gephi.org) with the ForceAtlas2 algorithm. The RGP and the spots of insertion were extracted using the panRGP (Bazin *et al*., 2020) subcommand of PPanGGOLin. RGP without a corresponding spot were excluded from further analysis. These included RGP on a contig border (i.e., likely incomplete) and instances in which the RGP is an entire contig (e.g., a plasmid, a region flanked with repeat sequences, or a contaminant). The frequency of the spot border gene and genes belonging to RGP were calculated using custom Python scripts.

2.2.3 Genome annotation and detection of genes of interest

The genomes were annotated using Prokka v1.14.6 (tool for annotating proteins in the bacterial genome) (Seemann, 2014). The virulence factors, antibiotic resistance, and stress resistance genes were identified using Abricate v1.0.1 (https://github.com/tseemann/abricate) against the Comprehensive Antibiotic Resistance Database (CARD) (Jia *et al*., 2017), NCBI AMRFinderPlus (Feldgarden *et al*., 2019) and Virulence Factor Database (VFDB) (Chen, Zheng, Liu, Yang, & Jin, 2016). The defence systems in the genomes were identified using PADLOC v1.1.0 (Payne *et al*., 2021) and DefenseFinder v1.0.9 (Tesson *et al*., 2022). The genes classified as adaptation or other categories were removed. Duplicate hits with the same gene name and location were removed using custom Python scripts. Quorum-sensing genes were detected using the automatic annotation process of QSP v1.0 (https://github.com/chunxiao-dcx/QSAP) from the QS-related protein (QSP) database (Dai *et al*., 2023). The virulence factors, antibiotic resistance genes, stress resistance genes, and defence systems were considered to be part of a particular RGP if the entire system was within the RGP.

2.2.4 Detection of plasmid, prophage and mobilome

Platon v1.6 (Schwengers *et al*., 2020), with default settings, was used to detect and annotate plasmids in the assemblies. Plasmid PubMLST (Jolley, Bray, & Maiden, 2018) was used for plasmid typing to determine the incompatibility groups. The *Salmonella* plasmid virulence (spv) region was identified by referencing the VFDB database and mapped onto the plasmid contigs to identify pSV. A representative pSV was visualised using Geneious Prime v2022.2.2. Prophage regions were identified using Phigaro v2.2.6 (Starikova *et al*., 2020) on default mode and PhageBoost v0.1.3 (Sirén *et al*., 2021) with a score >0.7 and a subsequent filtering with Phager (https://phager.ku.dk). From phages identified, duplicates were removed using Dedupe (https://github.com/dedupeio/dedupe) with a minimum identity of 100% and clustered at 95% identity across the region. taxmyPHAGE (https://github.com/amillard/tax_myPHAGE) was run on these regions to identify the phage kingdom, phylum, class, genus, species and name. For virulence factors, antibiotic or stress resistance genes, or defence systems to be considered within the prophage, the entire gene cluster had to be located within the prophage region. Heat maps were generated using GraphPad Prism v9.2.0.

2.2.5 Statistical analyses

Statistical analyses were performed using GraphPad Prism v9.2.0, employing simple linear regression and Pearson correlation analysis with a significance level set at a two-tailed P-value with a confidence interval of 95% for the correlation between the count of plasmid and antibiotic resistance genes.

2.2.6 Data availability

Refer to **Appendix II** for supplementary tables, interactive visualisation of the gene content of the spots and the interactive metadata of the isolates.

2.3 Results

2.3.1 The mobilome of *Salmonella* is highly variable across lineages

MGEs play a pivotal role in driving genetic diversity and shaping the evolutionary trajectories of bacteria, enabling them to adapt to various environmental challenges (Rocha & Bikard, 2022). One significant way through which MGEs exert their influence is by facilitating the horizontal transfer of genes associated with pathogenicity traits, directly impacting the potential of bacterial pathogens. To determine the broad relevance of specific MGEs in defining specific pathogenicity attributes of *Salmonella*, we explored the variation in plasmids and prophages across 12,244 *Salmonella* genomes (**Supplementary Table 2.1**). Our dataset included representative strains from the two species and six subspecies of *Salmonella*, as well as 46 serovars of *Salmonella enterica* subsp. *enterica* (**Fig. 2.1A**). These serovars were grouped into host-specific, host-adapted and broad-host range, with those lacking sufficient information categorised within the host range of unknown origin. As expected, the most prevalent serovars were Typhi (2,440 strains) and Typhimurium (2,170 strains), the main causative agents of typhoid fever (Stanaway *et al.*, 2019) and gastroenteritis (Eng *et al.*, 2015) in humans, respectively. The genome sequence of these strains was used to infer a phylogenetic topology representing the genomic diversity within the genus *Salmonella* (**Fig. 2.1B**). The overall topology of the phylogeny is

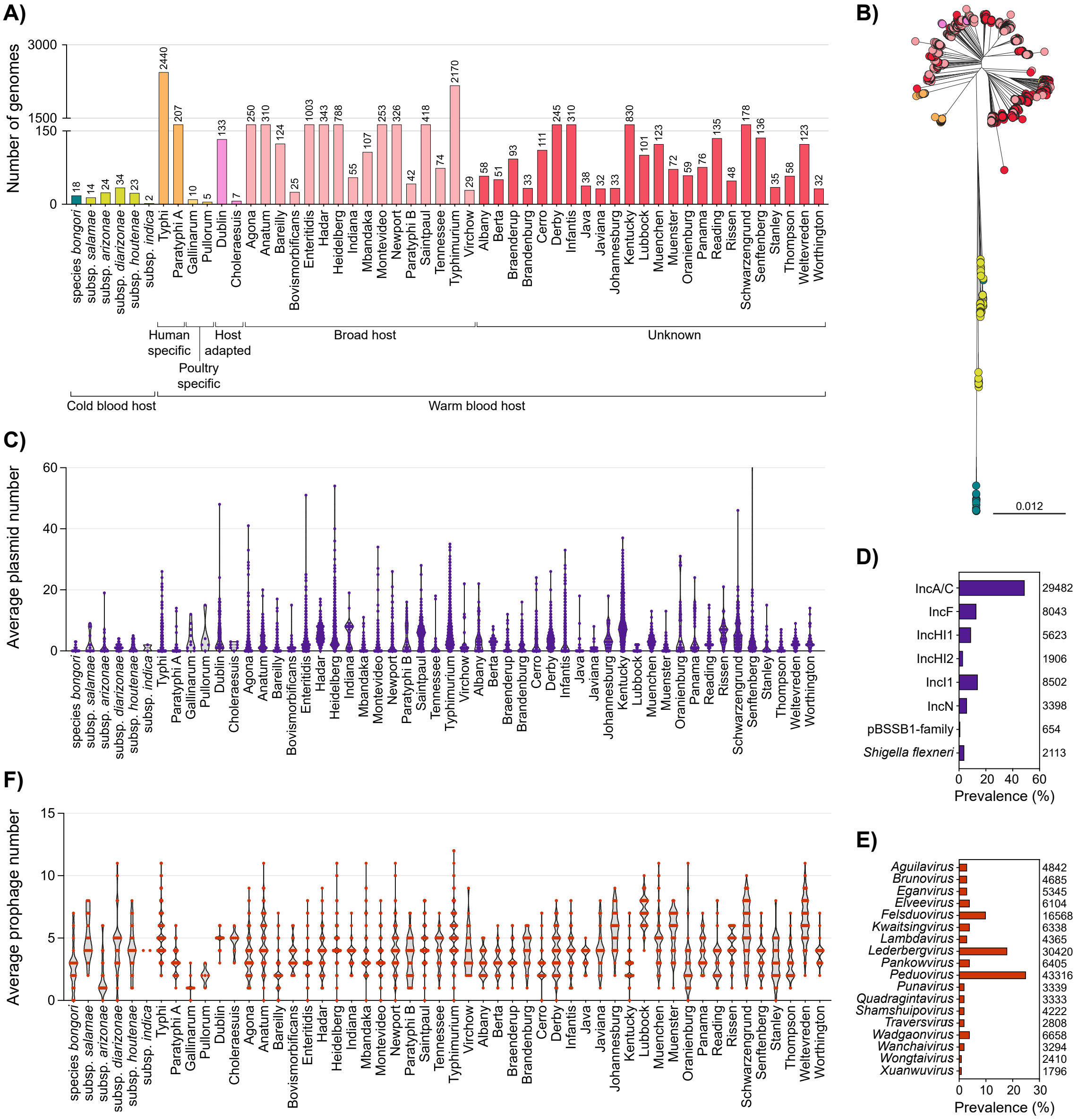


Figure 2.1 Classification of the 12.2K *Salmonella* strains. A) Distribution of strains from various *Salmonella* subspecies and serovars, categorised by their host specificity. **B)** Phylogenetic distribution of the *Salmonella* strains with a colour scheme analogous to (A). **C)** Average plasmid content in *Salmonella* subspecies and serovars. **D)** Prevalence of plasmid incompatibility groups in *Salmonella*. The *Shigella flexneri* plasmid incompatibility group refers to virulence plasmid pINV. **E)** Average prophage content in *Salmonella* subspecies and serovars. **F)** Prevalence of prophage genera in *Salmonella*.

in accordance with the phenogram created previously from concatenated MLST genes of a smaller number of genomes (Fookes *et al*., 2011).

Analysis of plasmid prevalence across different *Salmonella* subspecies and serovars revealed a variable abundance of these plasmid contigs (**Fig. 2.1C, Supplementary Table 2.2**). For instance, serovar Kentucky exhibited an average of 10 plasmid contigs, whereas most strains of serovar Paratyphi A lacked any plasmid contigs. Among the identified plasmids, the most prevalent were those belonging to the IncA/C group (39%, 29,482), IncI group (11%, 8,502), and IncF group (11%, 8,043) (**Fig. 2.1D**, **Supplementary Table 2.3**). Notably, IncA/C plasmids were predominantly found (58%, 17,091) in serovar Typhimurium, while IncHI1 (40%, 2,244) and IncN (30%, 1,025) were more commonly observed in serovar Typhi (**Supplementary Table 2.3**). Other plasmid types exhibited a more even distribution across different species.

Analysis of prophage prevalence in *Salmonella* shows that the vast majority of strains (96.6%, 11,829) carry at least one prophage, accounting for a total of 52,555 prophage regions (**Supplementary Table 2.4**). From this total, the taxonomy of 18,785 complete dsDNA prophage regions was determined using taxmyPHAGE (https://github.com/amillard/tax_myPHAGE). In most cases, we identified prophage regions associated with phages from multiple families, genera and species, resulting in a total of 172,862 entries. All these phages belong to the kingdom *Heunggongvirae*, phylum *Uroviricota*, and class *Caudoviricetes*. Within *Caudoviricetes*, 25% (43,316) of the phage regions belong to the genus *Peduovirus*, 18% (30,402) to the genus *Lederbergvirus* and 10% (16,568) to the genus *Felsduovirus* (**Fig. 2.1E**). The remaining phages are distributed across 68 other identified genera, though in smaller quantities. The most commonly identified phages include *Lederbergvirus Salmonella* phage BTP1, SE1Spa, P22, ST64T, *Enterobacteria* phage HK620 and *Shigella* phage Sf6. Similar to plasmids, the average number of prophages per strain varies among serovars, with serovar Lubbock averaging seven prophages, whereas serovar Gallinarum has only one (**Fig. 2.1F**).

In summary, our findings highlight the remarkable diversity observed in the mobilome of *Salmonella*. This diversity is reflected in the abundance and types of MGEs present per species, subspecies, and serovars. The variable nature of the mobilome and the resulting diversity in gene composition are expected to play a critical role in shaping the pathogenicity, adaptation, and distribution of *Salmonella*.

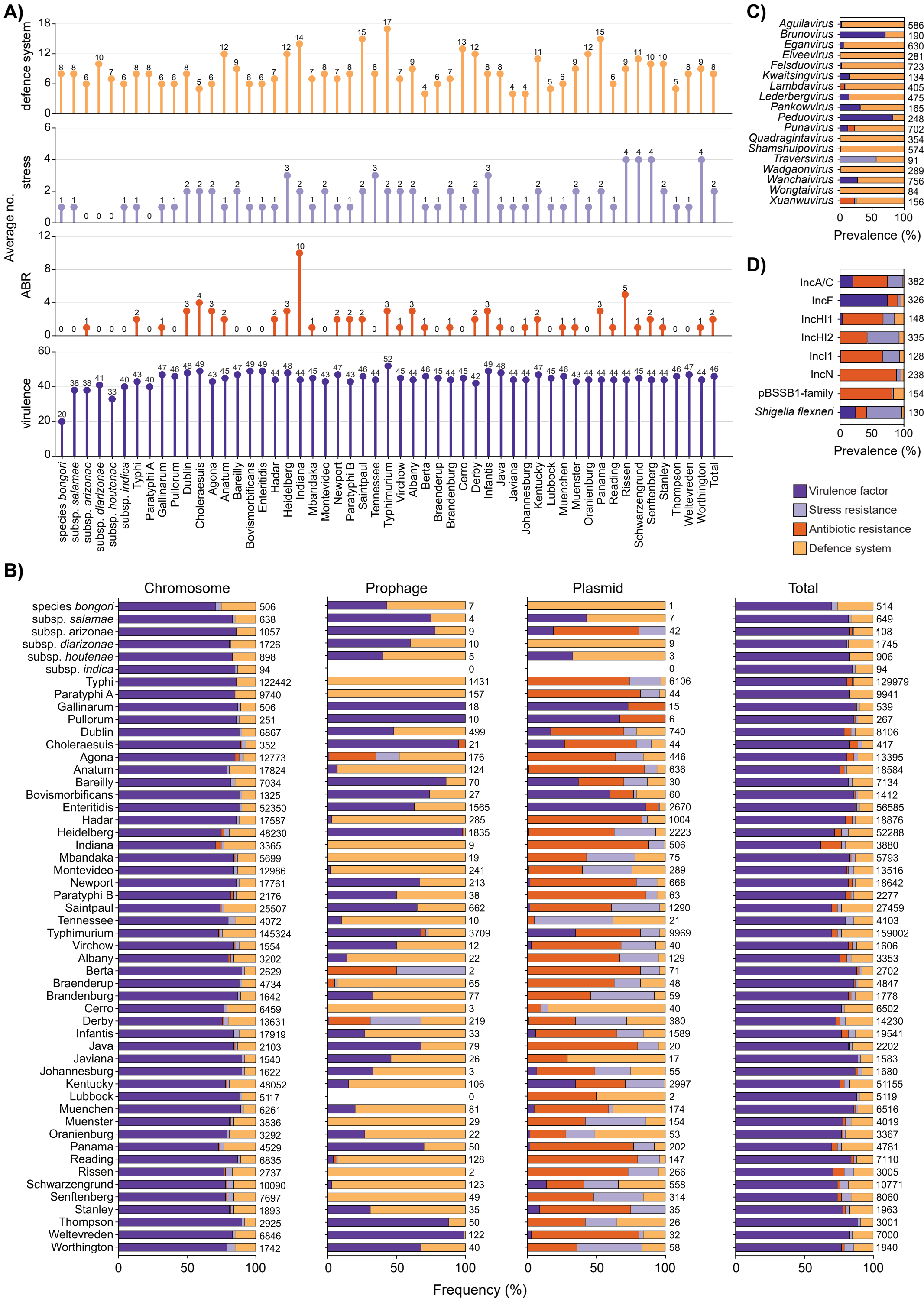


Figure 2.2 Prevalence and distribution of pathogenicity determinants in *Salmonella*. **A)** Prevalence of virulence factors, stress resistance genes, antibiotic resistance genes, and anti-phage defence systems across *Salmonella* subspecies and serovars. **B)** Distribution of the pathogenicity determinants across chromosomes, prophages, and plasmids. **C)** Distribution of the pathogenicity determinants across prophage genera present in *Salmonella* at >1% abundance. **D)** Distribution of the pathogenicity determinants across plasmid incompatibility groups present in *Salmonella*. In all panels, virulence factors, stress resistance genes, antibiotic resistance genes, and defence systems are coloured according to the key.

2.3.2 Virulence determinants are more prevalent in chromosomal regions

We next analysed the presence of factors contributing to the survival and adaptation of *Salmonella* to the environment. These included a set of virulence factors, antibiotic resistance genes, stress response genes, and phage-resistance genes (i.e., anti-phage defence systems) (the complete list of genes can be found in **Supplementary Table 2.5**). This analysis revealed the presence of virulence factors predominantly in *S. enterica* subsp. *enterica*, with an average of 46 virulence factors per strain (**Fig. 2.2A**, **Supplementary Table 2.6 & 2.7**). Interestingly, *S. bongori* has the lowest number of virulence genes, 20. In comparison to most *S. enterica* subsp. *enterica* strains, *S. bongori,* lacks the *Salmonella* pathogenicity island 2 (SPI-2), which encodes a type III secretion system (T3SS) that plays a central role in systemic infections and the intracellular phenotype of *S. enterica*, except for one strain (accession no. 1173775.3) that also groups with cold-blooded subspecies of *S. enterica* in the phylogenetic tree (**Fig. 2.1B**). The ability of SPI-2 to transfer to, and be functional in, *S. bongori* has been previously demonstrated experimentally (Hansen-Wester, Chakravortty, & Hensel, 2004) but, to our knowledge, not yet found in nature (Fookes *et al*., 2011). *S. bongori* do contain the SPI-1 with a T3SS that promotes invasion of epithelial cells through the secretion of different effector proteins (Lou, Zhang, Piao, & Wang, 2019). Curiously, SPI-1 is prevalent across all *Salmonella* species, subspecies and serovars but with variations in the presence of secreted effectors encoded by *spt* and *slr*, as well as *avr* and *ssp* genes, especially the latter (**Fig. 2.3**). Poultry-host-specific serovars Gallinarum and Pullorum share a more recent common ancestor with the broad host range serovar Enteritidis, but the gene *rck*, contributed by *Salmonella* virulence plasmid and responsible for evading the host immune response and surviving inside the host (Mambu *et al*., 2017), is absent in Gallinarum and Pullorum (**Fig. 2.3**). This gene is likely involved in the broader host range of Enteritidis strains.

Our analysis revealed that the majority of virulence factors are in chromosomal regions (**Fig. 2.2B & 2.3**, **Supplementary Table 2.6**). However, certain virulence genes are more commonly associated with prophages or plasmids. For example, genes *sod* and *grv,* critical to the bacterial response to oxidative stress and their ability to survive within immune cells (De Groote *et al*., 1997; Ho & Slauch, 2001), are frequently found in prophage regions, mostly of *Peduovirus* (80.1% *sod*, 86.1% *grv*) and *Brunovirus* (3.7% *sod*,

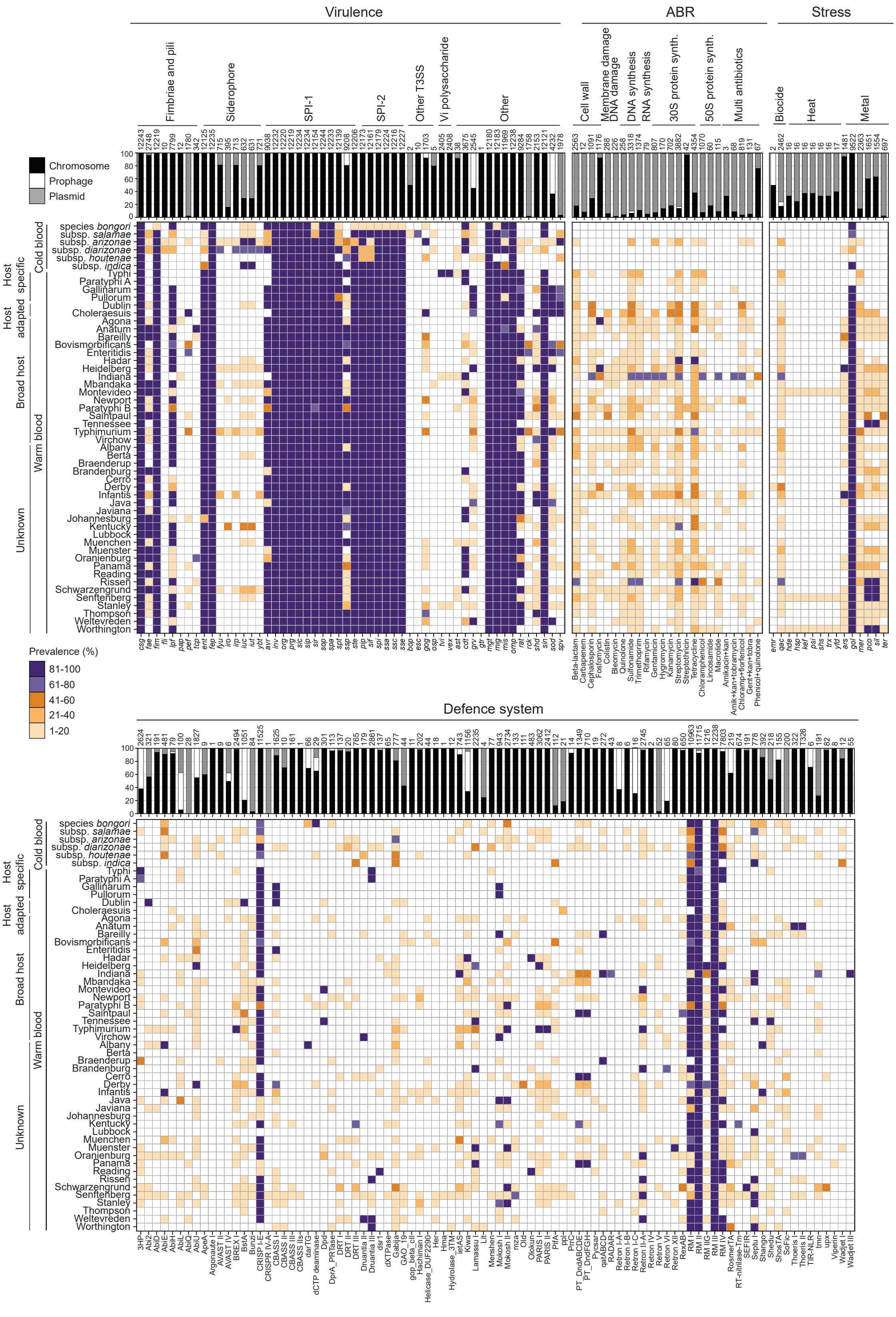


Figure 2.3 Distribution of virulence factors, antibiotic resistance (ABR) genes, stress resistance genes, and defence system across *Salmonella.* The prevalence of the specific gene (cluster) in chromosome, prophage, or plasmid is shown as a bar graph.

3.7% *grv*) genera (**Fig. 2.3**). These prophages seem to preferentially carry virulence factors (**Fig. 2.2C**). Surprisingly, in contrast to existing literature (Coombes *et al*., 2005), we found that the majority of *gog* genes, which are associated with an anti-inflammatory function (Pilar, Reid-Yu, Cooper, Mulder, & Coombes, 2012), are located on the chromosome (1,574 out of 1,703) rather than a prophage region (**Fig. 2.3, Supplementary Table 2.6**). The well-known virulence genes *spv* (involved in intracellular survival and evasion of the host immune response (D. Guiney & Fierer, 2011), *pef* (plasmid encoded fimbriae, important for colonisation of the host and establishment of infection (Bäumler *et al*., 1996; Ledeboer, Frye, McClelland, & Jones, 2006), and *rck* (contributing to evasion of the host immune response (Wiedemann *et al*., 2016) are predominantly found on plasmids (**Fig. 2.3**), particularly those belonging to the IncF group (**Fig. 2.2D, Supplementary Table 2.8**). These *Salmonella* virulence plasmids (pSV) containing the *spv* genes were identified in *S. enterica* subsp. *arizonae* and *S. enterica* subp. *enterica* serovar Typhimurium, Dublin, Enteritidis, Choleraesuis, Gallinarum, and Pullorum, consistent with the existing literature (Gulig *et al*., 1993; D. G. Guiney, Fang, Krause, & Libby, 1994; D. G. Guiney *et al*., 1995; Libby *et al*., 2002). Genes *pef* and *rck* have also been reported previously in pSV (Feng *et al*., 2012). The *fyu* and *ybt* genes involved in iron acquisition (Oelschlaeger *et al*., 2003) were predominantly associated with IncA/C plasmids. Notably, we did not find any virulence factors on IncHI2, IncN, and pBSSB1-family plasmids (**Fig. 2.2D**, **Supplementary Table 2.8**).

2.3.3 Antibiotic resistance genes are primarily located within plasmids

Plasmids serve as the primary reservoir for antibiotic resistance (ABR) genes in *Salmonella* (**Fig. 2.2B**). Specifically, 78% of the ABR genes identified in *Salmonella* were located on plasmids, with 21% found on chromosomal regions (predominantly streptothricin and fosfomycin), and 1% associated with prophages (mostly of the *Xuanvirus* genus) (**Fig. 2.2C**). While ABR levels in *Salmonella* prophages are lower compared to those in plasmids or the chromosome, they surpass previously reported general prophage analyses (Enault *et al*., 2017; Cook, 2021). Most ABR genes were found

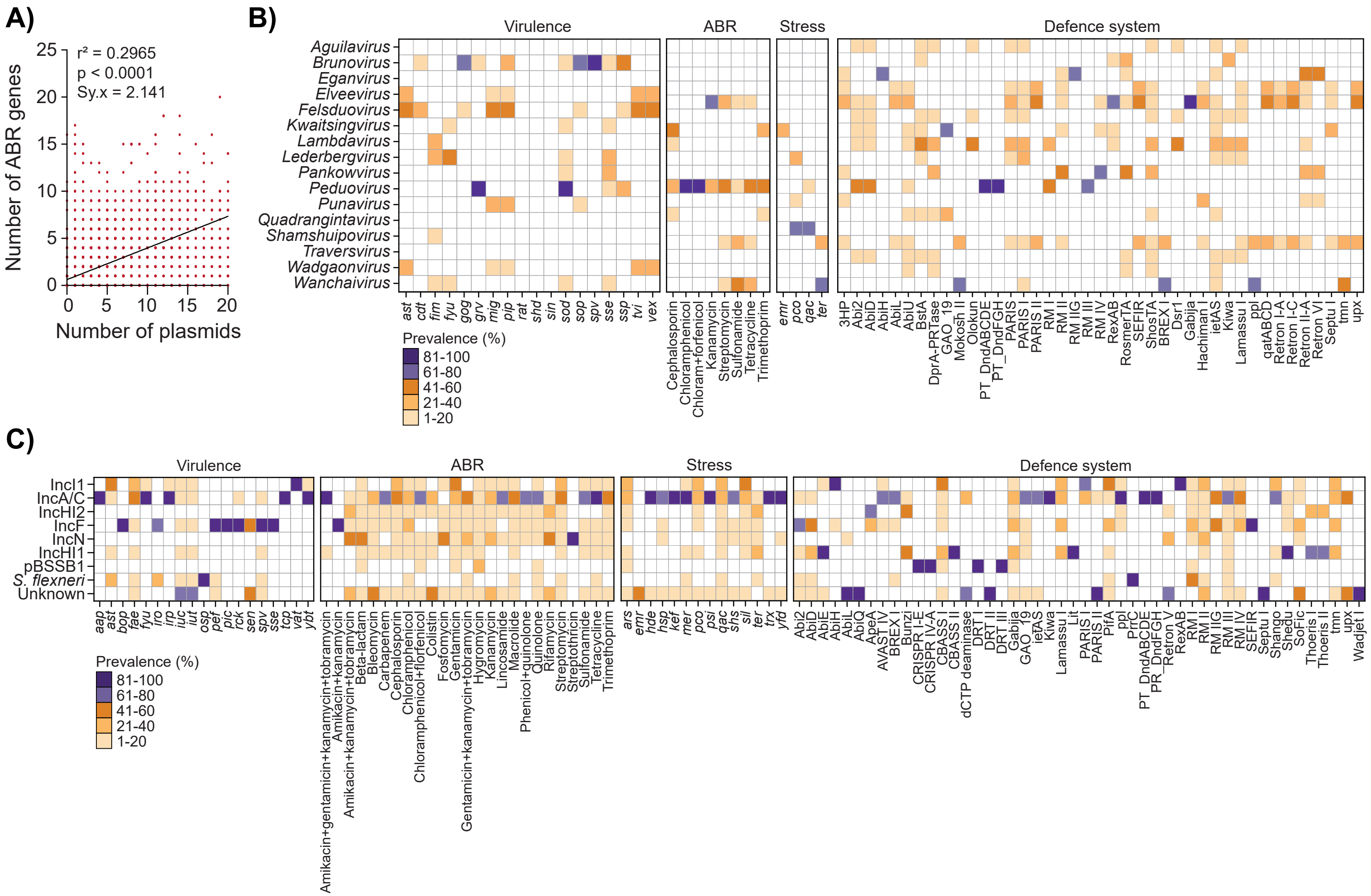


Figure 2.4 Distribution of pathogenicity determinants on plasmid and prophage classes. A) Correlation between the number of plasmids and the number of ABR genes in *Salmonella* strains. **B)** Frequency distribution of virulence factors, antibiotic resistance (ABR) genes, stress resistance genes, and defence systems on different prophage genera. **C)** Frequency distribution of virulence factors, ABR genes, stress resistance genes, and defence systems on different plasmid incompatibility groups.

across multiple Inc plasmid schemes, but majorly in IncN and IncA/C (**Fig. 2.2D, Supplementary Table 2.8**).

Consistently, the strains exhibited the highest resistance to tetracycline, streptomycin, sulphonamide, and beta-lactam antibiotics (**Fig. 2.3**), which aligns with previous reports (V. T. Nair, Venkitanarayanan, & Kollanoor Johny, 2018). Importantly, ABR genes were prevalent in strains of *S. enterica* subsp. *enterica* but negligible (≤ 1 ABR gene) in other *S. enterica* subspecies (except *S. enterica* subsp. *arizonae*) and *S. bongori*. This pattern does not seem to be strongly driven by a lower abundance of plasmids (r^2 = 0.2965, p = <0.0001) (**Fig. 2.4A**). Notably, serovar Paratyphi A showed minimal presence of ABR genes, as plasmids are also almost absent in this serovar. On the other hand, serovars Indiana and Rissen carried an average of 10 and 5 ABR genes, respectively, indicating multidrug resistance, consistent with previous reports (Gong *et al*., 2019; Xu *et al*., 2020) (**Fig. 2.3**, **Supplementary Table 2.5, 2.6 & 2.7)**. The serovars Heidelberg, Typhimurium, Newport, and Enteritidis are known to cause the majority of outbreaks, and 89%, 75%, 32% and 10% of strains from these serovars contain ABR genes (**Fig. 2.3, Supplementary Table 2.6**). Importantly, the presence of resistance to colistin, an antibiotic of last resort, was detected in 2.4% (288) of strains belonging to *S. enterica* subsp. *enterica*, with a predominant occurrence in serovars Saintpaul, Cholerasuis, and Paratyphi B (**Fig. 2.3**).

In summary, our results reinforce the role of plasmids in influencing antibiotic resistance patterns in *Salmonella* and highlight that plasmids of all schemes are drivers of ABR dissemination. Moreover, the higher prevalence of antibiotic resistance in *S. enterica* subsp. *enterica*, compared to other *Salmonella* species and subspecies, underscores the influence of human antibiotic usage in promoting the spread of antibiotic resistance.

2.3.4 Stress-resistance genes are primarily located on plasmids and chromosomal regions

The presence of acid, biocide, and heavy metal resistance genes is closely linked to the maintenance and spread of antimicrobial resistance (Hasman & Aarestrup, 2002; Campos, Cristino, Peixe, & Antunes, 2016; Yang, Agouri, Tyrrell, & Walsh, 2018). Interestingly, we observed that the two multidrug-resistant serovars, Indiana and Rissen, exhibit the highest prevalence of *qac* genes, which are small multidrug resistance efflux proteins associated with increased tolerance to quaternary ammonium compounds (QAC) and other cationic biocides (Jaglic & Cervinkova, 2012) (**Fig. 2.3**). *qac* genes are generally found in MGEs; here, they were found on IncI1 and IncA/C plasmids in over 75% of the cases (**Supplementary Table 2.8**).

Curiously, most strains in our dataset do not carry any heat-resistant genes, except for a small percentage (<20%) of strains from serovars Montevideo, Senftenberg, and Worthington, and the majority of these genes are located on IncA/C plasmids. On the other hand, *Salmonella* strains commonly exhibited resistance to heavy metals, with approximately 80% of the strains carrying genes conferring resistance to gold (**Fig. 2.3**). The only exceptions are *S. enterica* subsp. *houtenae* and *S. enterica* subsp. *enterica* serovars Typhi and Paratyphi A, which do not carry the *gol* cluster responsible for gold resistance. Serovars Heidelberg and Infantis show a high incidence (>95%) of arsenic resistance genes (*ars*), while serovars Tennessee, Rissen, Schwarzengrund, Worthington, and Senftenberg exhibit frequent (>80%) copper (*pco*) and silver (*sil*) resistance (**Fig. 2.3**).

Curiously, genes that confer resistance to heavy metals such as gold, arsenic, copper, and silver are predominantly located in chromosomal regions, while those associated with mercury and tellurite resistance are commonly found on plasmids (IncA/C, and IncHI1 and IncHI2, respectively) (**Fig. 2.3**). Among the different plasmid schemes, *Shigella flexneri* (virulence plasmid pINV) (Pilla, Arcari, Tang, & Carattoli, 2022) and IncHI2 plasmids are those most frequently associated with stress resistance genes, but IncA/C plasmids, due to their abundance, are responsible for the movement of most stress resistance genes (**Fig. 2.2D**, **Supplementary Table 2.8**).

In summary, our findings reveal that genes associated with resistance to biocides, heat, and heavy metals such as mercury and tellurite are primarily found on plasmids, while resistance to gold, arsenic, copper, and silver is commonly found within chromosomal regions. The elevated levels of heavy metal resistance observed in specific serovars raise concerns about the use of heavy metal-based compounds in animal-production settings.

2.3.5 Anti-phage defence systems are more prevalent in chromosomal regions

Anti-phage defence systems were found to be prevalent among *Salmonella* strains, with an average of eight defence systems per strain. This is higher than the average found in *Escherichia coli* (six) (Wu *et al*., 2023) or *Pseudomonas aeruginosa* (seven) (Costa *et al*., 2023) in previous studies. However, there is considerable variation in the number of defence systems carried by different subspecies and serovars. For example, serovars Typhimurium (17), Saintpaul (15), Panama (15), and Indiana (14) exhibit the highest prevalence of defence systems, while serovars Berta, Javiana, and Johannesburg have the lowest (4) (**Fig. 2.2A**, **Supplementary Table 2.6 & 2.7**).

Among the 90 defence system subtypes identified in *Salmonella* strains, the most prevalent were the restriction-modification (RM) and type I-E CRISPR-Cas systems, which are present in almost all subspecies and serovars (**Fig. 2.3**). However, the CRISPR-Cas system is absent in serovars Brandenburg, Lubbock, and Worthington. We noted significant variation in the prevalence of other defence systems across the *Salmonella* genus (**Fig. 2.3**). Notably, each serovar appears to have a distinct profile of defence systems, suggesting the selection of the most beneficial systems in specific environments or host interactions, as previously observed for distinct *E. coli* phylogroups (Wu *et al*., 2023). For example, in serovars Typhi and Paratyphi A, the 3HP and Druantia type III systems are highly abundant. On the other hand, in Typhimurium, we observed a predominance of the BREX type I, Mokosh type II, PARIS types I and II, and Retron II-A defence systems. Strains of Enteritidis exhibit enrichment in CBASS type I, while Gallinarium and Pullorum frequently harbour Mokosh type I in addition to CBASS type I. Additionally, we found specific defence systems enriched in particular species and subspecies. For instance, dCTP deaminase is more prevalent in *S. bongori*, Septu type I in *S. enterica* subsp. *indica* and *salamae*, and Gabija in *S. enterica* subsp. *arizonae* (**Fig. 2.3**).

In general, defence systems, including the abundant RM and CRISPR-Cas systems, are more frequently found within chromosomal regions (94%) (**Fig. 2.2B**) compared to prophages or plasmids. However, prophages of all genera except *Brunovirus*, *Peduovirus* and *Traversvirus* show a clear preference for carrying defence systems over other types of pathogenicity-related genes (**Fig. 2.2C**), and the defence systems 3HP, AbiL, BstA, Kiwa, Retron types I-A, I-C, and VI are predominantly found on prophages (**Fig. 2.3 & 2.4B**). Other defence systems, such as AbiQ, Bunzi, Gao_19, Lit, PifA, ppl, retron type V, SoFic, and tmn are frequently associated with plasmids. Among these, the systems ppl (89%) and GAO_19 (64%) are primarily linked to IncA/C plasmids, while pifA is mostly found (58%) on IncI1 plasmids. Lit (100%) and Bunzi (49%) are often identified on IncHI plasmids

(**Fig. 2.4C**, **Supplementary Table 2.8**). Interestingly, although plasmids of all types often accommodate a greater abundance of other pathogenicity-related elements (**Fig. 2.2D**), it is noteworthy that the IncHI1 and pBSSB1-family plasmids demonstrate a higher inclination toward carrying defence systems compared to other plasmid schemes.

In summary, anti-phage defence systems are widespread in *Salmonella*, with a notable prevalence of the R-M and the CRISPR-Cas systems. The significant variation in defence system repertoire across *Salmonella* species and serovars observed here highlights the significance of these systems in the evolution and adaptation of this pathogenic bacterium. This is likely influenced by their differential prevalence in distinct MGEs.

2.3.6 Gene clusters integrate into preferential spots in the *Salmonella* genome

Our analysis uncovered substantial variability in the presence and arrangement of genes associated with virulence, antibiotic resistance, stress response, and anti-phage defence genes among different *Salmonella* strains. This variability strongly suggests the occurrence of genomic rearrangements involving the insertion and deletion of genes. To gain a deeper understanding of genome plasticity within *Salmonella*, we performed a comprehensive mapping analysis using PPanGGoLiN (Gautreau *et al*., 2020) and panRGP (Bazin, Gautreau, Médigue, Vallenet, & Calteau, 2020) to identify RGP.

Our findings show that only 4.6% of the genes are conserved in nearly all *Salmonella* genomes (3,575 persistent genes, among which 65 are core genes), while 5.5% (4,256) were present at intermediate frequencies (shell genes), and ~90% (69,678) at low frequency (cloud genes) (**Fig. 2.5A**). Analysis of the average gene length showed that persistent and core genes (873 bp) are significantly longer than shell genes (558 bp) and cloud genes (420 bp) (**Fig. 2.5B, Supplementary Table 2.9**). In eukaryotes, longer genes are suggested to be more evolutionarily conserved and associated with important biological processes (Wolf, Novichkov, Karev, Koonin, & Lipman, 2009; Vishnoi, Kryazhimskiy, Bazykin, Hannenhalli, & Plotkin, 2010; Gorlova, Fedorov, Logothetis, Amos, & Gorlov, 2014; Grishkevich & Yanai, 2014). This observation aligns with our findings in *Salmonella*, as functional analysis of the persistent genes revealed their essential roles in survival and fitness (**Supplementary Table 2.9**). In contrast, the shorter gene length is associated with high gene expressions (Urrutia & Hurst, 2003), providing an advantage in

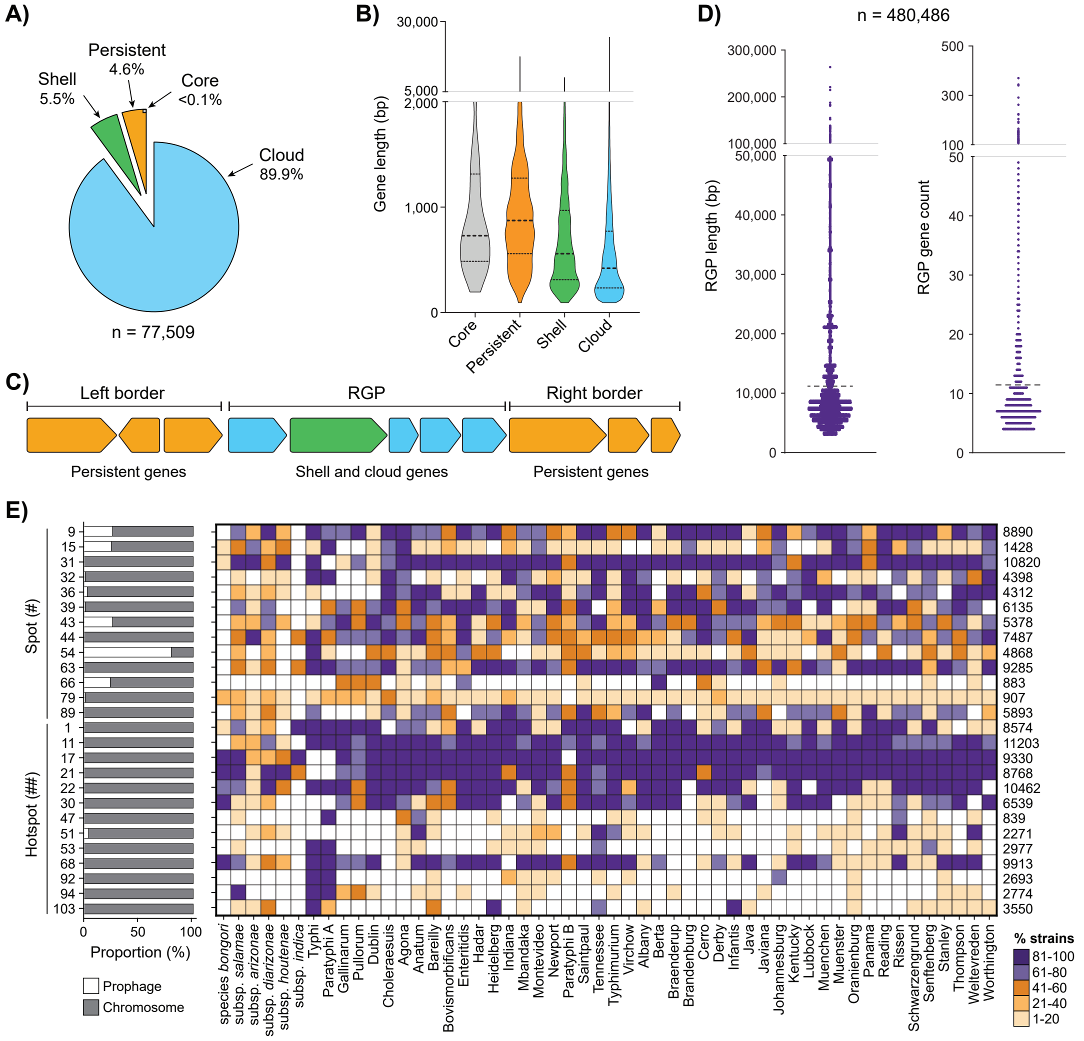


Figure 2.5 Pangenome analysis of 12,244 *Salmonella* genomes. A) Partition of *Salmonella* gene families by PPanGGOLiN, based on their conservation across strains. Core, conserved in all genomes; persistent, conserved in almost all genomes; shell, moderately conserved; cloud, poorly conserved. **B)** Length of core, persistent, shell, and cloud genes. **C)** Schematic representation of a region of genomic plasticity (RGP), consisting of shell and cloud genes, identified in between conserved border genes. **D)** Length and gene count in the identified RGP. **E)** Distribution of 26 integration spots across chromosomes and prophages and their prevalence across *Salmonella* subspecies and serovars. The 26 spots correspond to those identified in >1% strains of *Salmonella* containing pathogenicity determinants. Spots (#) are characterised by the presence of > 1 and ≤ 100 unique genes, and hotspots (##) by the presence of > 100 genes.

response to stimuli (Kirkconnell *et al.*, 2017). This observation is consistent with the role of accessory shell and cloud genes, which are likely to confer fitness benefits under specific environmental and stress conditions.

Clusters of accessory shell and cloud genes form RGP, primarily originating from horizontal gene transfer events. These RGP can be grouped into specific insertion spots based on the presence of conserved flanking persistent genes (**Fig. 2.5C**). Our analysis identified a total of 673,113 RGP, among which 71.4% (480,486) were clustered in 1,345 spots. These RGP have an average length of 11,182 bp and an average of 11 genes per RGP (**Fig. 2.5D**). The majority of the RGP (96.5%) is located in the bacterial chromosome, and the remaining 3.5% are prophages (i.e. border genes of the spot corresponding to those of the prophage) (**Supplementary Table 2.10**). Out of the 1,345 spots, 74.65% (1,004) were specific to a single type of RGP, while the remaining spots exhibited the potential to harbour a diverse array of RGP families with diverse gene content. Importantly, 1.64% (22) of these spots could harbour >100 distinct RGP families (**Supplementary Table 2.11**), suggesting higher rates of gene acquisition and underscoring these regions as hotspots for gene integration (Bazin *et al.*, 2020).

We screened all spots for the presence of virulence genes, antibiotic resistance genes, stress resistance genes, and defence systems (**Supplementary Table 2.12**). Among the resulting 266 spots, we selected those with variable content present in at least 1% of the strains, yielding 26 spots (#) (13 of which are hotspots, ##) for further analysis. Some spots were relatively specific to certain serovars, such as spot ##47 in serovars Paratyphi A, Anatum, Agona and Rissen, spot ##92 in Typhi, Paratyphi A, Johannesburg, or spot ##94 in host-specific serovars. In contrast, other spots were widely distributed, such as spots #9 or ##22 (**Fig. 2.5E**).

When examining the gene content of these spots related to virulence, stress, antibiotic resistance, and defence systems, we can observe that genes with specific functions show a clear propensity to congregate within particular locations (**Fig. 2.6A**). For instance, *lfp* genes preferentially localise in hotspot ##30, while *fae* genes predominantly localise in spot #36 (**Fig. 2.6B, Supplementary Table 2.12**). The gene cluster conferring tolerance to gold (*gol*) distinctly favours spot ##17; the absence of this spot in serovars Typhi and Paratyphi A leads to the absence of gold resistance (**Fig. 2.3**). However, spot ##17 is present in a few strains of *S. enterica* subsp. *houtenae*, where gold resistance is lacking, indicating that the presence of this spot does not consistently correlate with gold resistance. Similarly, the gene cluster conferring arsenic resistance (*ars*) predominantly localises in spot ##103, which is prevalent in serovars Heidelberg and Infantis, the strains of which display the highest prevalence of arsenic resistance. However, spot ##103 is also frequently found in serovar Typhi, where arsenic resistance is absent. Collectively, these findings underscore that the presence of a spot where a specific gene cluster predominantly localises does not unequivocally signify the presence of said gene cluster; conversely, the absence of the site often corresponds to the absence of the specific gene cluster. In cases of the former, other influencing factors may contribute to the selection for the presence of such gene clusters, potentially spurred by environmental pressures.

Another important example of gene cluster preference for distinct spots involves the type I-E CRISPR-Cas system, predominantly found in spot ##22. Spot ##22 is ubiquitously present across all species except for serovars Brandenburg, Java, Javiana, Johannesburg, Lubbock, Mbandaka, Panama, Reading, and Worthington (**Supplementary Table 2.12**). In these serovars, the CRISPR-Cas system is either entirely absent or present in only a limited number of strains (**Fig. 2.3**). The well-conserved nature of the type I-E CRISPR-Cas system in *Salmonella* (Shariat, Timme, Pettengill, Barrangou, & Dudley, 2015; Kushwaha, Bhavesh, Abdella, Lahiri, & Marathe, 2020) seems intrinsically tied to the widespread prevalence of spot ##22 (90% of all strains). Additionally, RM types I and IV have a strong preference for spot ##68, which is present in over 80% of the strains, thus accounting for the wide prevalence of RM systems in *Salmonella*.

On a broader scale, we also observe a general inclination for gene clusters with related functions to cluster within the same spots. This is well demonstrated by particular spots housing diverse anti-phage defences, functioning as hotspots for variable defence systems (e.g., ##11, #39, #43, and #63). For example, spot #63 contains an array of defence systems, with Dpd, Druantia type III, Mokosh type II, and Wadjet type III present in more than 80% of instances (**Fig. 2.6, Supplementary Table 2.12**).

In summary, our findings reveal the preference of gene clusters to integrate into specific spots within the *Salmonella* genome, with the presence of these spots indicating the potential presence of these gene clusters. These dynamic regions play a critical role in bacterial adaptability and fitness, as evidenced by the exclusive association of the type I-E CRISPR-Cas system with serovars containing spot ##22.

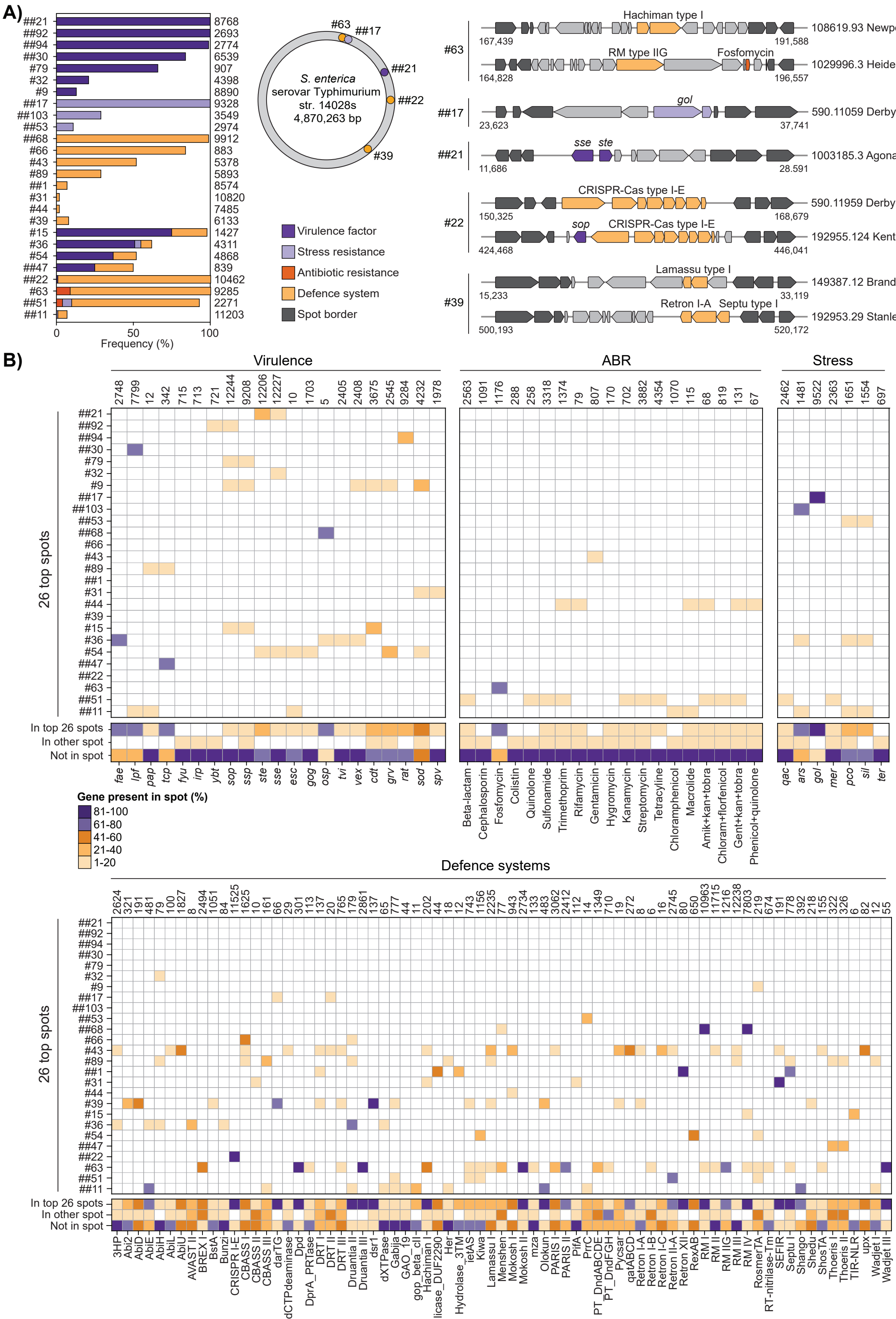


Figure 2.6 Distribution of gene content in the spots of *Salmonella* genomes. A) Distribution frequency of pathogenicity determinants across spots identified in at least 1% strains of *Salmonella.* Mapping of spots present on the reference strain *S. enterica* subsp. *enterica* serovar Typhimurium str. 14028s, with examples illustrating various gene arrangements. **B)** Frequency at which a specific gene cluster appears outside a spot, in one of the 26 most abundant spots, or in other spots. Only gene clusters found in spots are depicted.

2.3.7 Spot flanking genes are likely determinants of gene cluster preference

We investigated whether the preference of the gene clusters for particular spots was influenced by the gene neighbourhood, particularly the highly conserved genes flanking the spots. To accomplish this, we examined the genomic locations of the 26 prevalent spots identified in our study (an interactive visualisation of the gene content of the spots can be found at the associated GitHub, see Data Availability).

While some flanking genes have unknown functions, their predicted roles suggest potential connections to spot functionality. For instance, spots ##1 and #63, which seem to be preferred by anti-phage defence systems, are associated with helix-turn-helix (HTH)-type transcriptional regulators (*ecpR* and *gntR*, respectively) as border genes (**Supplementary Table 2.13**). Gene *ecpR* in spot ##1 is negatively regulated by H-NS and positively regulated by itself and the integration host factor (IHF), a protein involved in various phage-related processes such as integration and propagation (Zablewska & Kur, 1995). The regulatory interactions involving *ecpR* and IHF suggest a potential influence of the first on phage-related processes. Similarly, spot #63 features the flanking gene *gntR*, which influences cell wall permeability and bacterial motility, both factors known to affect phage infectivity (An *et al.*, 2011). Additionally, spot #63 contains a flanking gene encoding protein L-threonine 3-dehydrogenase, and L-threonine has been observed to impact phage infection in *E. coli* (L. Cheng *et al.*, 2020). Spot ##22 houses the CRISPR-Cas type I-E defence system and is adjacent to the *cysD* gene involved in cysteine biosynthesis (Malo & Loughlin, 1990). Notably, the regulation of the *cysD* gene involves the *cnpB* gene, which participates also in CRISPR-Cas regulation (Zhang, Yang, & Bai, 2018). This co-localisation suggests potential coordination between these genes for regulatory purposes.

Spot ##17 is a hotspot for gold resistance and is flanked by the *oprM* gene. *oprM* encodes an outer membrane protein that functions as an antibiotic and metal pump (Masi & Pagès, 2013), enabling the bacterium to defend against the toxicity of antibiotics and metals. This suggests that the presence of *oprM* in the vicinity of spot ##17 contributes to gold resistance by facilitating the efflux of gold ions from the bacterial cell. Similarly, spot ##53, associated with copper and silver resistance genes, is adjacent to the *uspB* gene encoding a universal stress response protein (Liu *et al.*, 2007). UspB promotes cell survival and protects against stress-induced damage, potentially aiding the bacterium in coping with copper and silver stress.

Spot ##21 contains the *ste* and *see* genes involved in *Salmonella* enterotoxin production and is flanked by the *mdoD* gene, which encodes glucan biosynthesis protein D. This protein is essential for the synthesis of osmoregulated periplasmic glucans, which contribute to the stability and integrity of the bacterial cell envelope (Debarbieux, Bohin, & Bohin, 1997). Although no direct connection between periplasmic glucans and enterotoxin production has been reported, it is possible that glucan production and the secretion systems responsible for enterotoxin export indirectly influence each other through broader cellular processes or regulatory networks. Spot ##30 contains the *lpf* gene cluster responsible for the production of long polar fimbriae, which facilitates bacterial adherence and colonisation of host cells and tissues (Doughty *et al*., 2002; R. A. Cheng & Wiedmann, 2020). This spot is flanked by the *eptB* gene encoding Kdo (2)-lipid A phosphoethanolamine 7''-transferase, an enzyme that modifies the lipopolysaccharide (LPS) lipid A portion, contributing to bacterial resistance against cationic antimicrobial peptides (Wang, Quinn, & Yan, 2015). *eptB* may support the survival of *Salmonella* adhering to host cells *via lpf* by protecting the bacterial cells from the antimicrobial peptides produced by endothelial cells, particularly in environments like the gastrointestinal tract.

Finally, spot ##51, which harbours several antibiotic resistance genes, is flanked by the *yidC* and *mdtL* genes. The *yidC* gene encodes the membrane protein YidC, crucial for the insertion and folding of membrane proteins in bacteria. YidC is implicated in the proper folding and assembly of essential membrane proteins associated with antibiotic resistance mechanisms and has been proposed as a potential antibiotic target (Tzeng *et al*., 2020; Dalbey, Kaushik, & Kuhn, 2023). On the other hand, the *mdtL* gene encodes a multidrug efflux transporter protein responsible for exporting a wide range of drugs and toxic compounds out of the bacterial cell, contributing to antibiotic resistance.

In conclusion, our investigation revealed potential relationships between spot functionality and the genes in their vicinity. The identification of specific flanking genes suggests their involvement in various processes related to phage defence, metal resistance, stress response, and antibiotic resistance. These spatial arrangements provide insights into potential coordination, regulatory connections, and adaptive mechanisms within bacterial genomes.

2.4 Discussion

The genetic landscape of *Salmonella* is a mosaic shaped by various factors, with RGP acting as significant contributors. These dynamic genomic segments house diverse gene clusters that hold the potential to dictate the genetic makeup of *Salmonella* strains. In the traditional view, gene distribution within a genome was often perceived as a stochastic process, but recent insights have challenged this view by revealing that genes linked to specific functions tend to cluster in certain regions (Schmidt & Hensel, 2004; Makarova, Wolf, Snir, & Koonin, 2011). The extent and implications of this phenomenon for bacterial evolution and adaptation have remained largely unexplored.

Our findings revealed a distinctive pattern of non-random integration of gene clusters into specific RGP of *Salmonella*. Exploring the prevalence of certain RGP across diverse lineages of *Salmonella* revealed their pivotal role in shaping genetic content and, thus, the pathogenicity and survival strategies of each lineage. Noteworthy examples include the presence of SPI-2 in one strain of *S. bongori*, indicating it may have developed the ability to infect warm-blooded hosts. This divergence challenges established notions of gene distribution and exemplifies how RGP can redefine our understanding of gene presence and absence across lineages. Furthermore, the association between the absence of type I-E CRISPR-Cas systems and the lack of spot #22 provides a rationale for the previous observation of the missing type I-E CRISPR-Cas system in specific *Salmonella* serovars (Gupta *et al*., 2019).

The mobility of gene clusters across genomes has raised questions about their integration without guaranteed regulation of expression upon insertion (Nitschké *et al*., 1998; Overbeek, Fonstein, D'Souza, Pusch, & Maltsev, 1999; Pellegrini, Marcotte, Thompson, Eisenberg, & Yeates, 1999; Galperin & Koonin, 2000). Potential problems include situations where the cluster relies on regulatory interactions that are not present in the new host, genes fail to express correctly, or auxiliary interactions and dependencies on the host come into play (Fischbach & Voigt, 2010). However, our study demonstrates a non-random integration pattern of RGP and their associated gene clusters, suggesting a purposeful selection of locations rather than randomness. Bacterial genomes are often organised in gene clusters regulated by shared regulators (Lawrence, 2002), supporting the idea that RGP are placed in specific spots primarily due to the benefit of co-regulation. This suggests that genes flanking certain genomic spots might dictate the integration of particular RGP. For instance, the strategic insertion of genes responsible for long polar fimbriae production in regions flanked by antimicrobial peptide resistance genes suggests functional synergy, potentially aiding the survival of *Salmonella* by protecting it from host-produced antimicrobial peptides during invasion. This underscores the likely coordination of gene expression among co-localised genes. Similarly, positioning stress resistance genes near specific RGP, harbouring metal resistance genes (##17, #53) might reflect a fine-tuned regulatory network to efficiently counter stressors. Notably, hotspots associated with antibiotic resistance genes (e.g., #51) are flanked by genes implicated in antibiotic resistance mechanisms, while certain spots with anti-phage defence systems (#1, #63) are flanked by HTH transcriptional regulators linked to phage-related processes. These preferences for genomic locations might be driven by selective pressures or other factors that ensure co-expression and coordinated functionality, contributing to the intricate landscape of bacterial adaptation and evolution. Examining these potential functional links can unveil novel pathogenicity traits and gene interaction networks crucial for understanding *Salmonella* pathogenicity.

Unsurprisingly, our findings underscored the prominent role of plasmids in influencing ABR patterns within *Salmonella*. Notably, a substantial proportion (78%) of ABR genes were housed within these MGEs, particularly those of the IncN, IncA/C, IncHI1 and IncI1 plasmid groups. Noteworthy variations were observed across subspecies and serovars, with ABR genes predominantly concentrated in *S. enterica* subsp. *enterica*, raising concerns over the potential amplification of ABR due to human antibiotic usage. An especially troubling discovery was the presence of colistin resistance genes within serovars associated with human outbreaks, such as Saintpaul, Cholerasuis, and Paratyphi B. Colistin, designated as a crucial antibiotic by the World Health Organization (Vázquez *et al.*, 2021), serves as a last-resort defence against life-threatening infections caused by multidrug-resistant Gram-negative bacteria (Vázquez *et al.*, 2021). The occurrence of plasmid-borne colistin resistance within these outbreak-causing serovars (Lima, Domingues, & Da Silva, 2019) carries the risk of propagation to other bacteria, including those with substantial clinical relevance.

But the role of plasmids extends beyond ABR. The IncF group favours virulence factors, aligning with previous reports (McMillan, Jackson, & Frye, 2020). Plasmids from the IncI1 and IncA/C groups are key vectors for *qac* gene dissemination, associated with antimicrobial and biocide resistance, and also carry mercury and tellurite resistance determinants. Plasmids affiliated with the IncHI2 group and of *Shigella flexneri* preferentially carry stress resistance determinants (*qac, mer, sil* and *ter*), while those from the pBBSB1 and IncHI groups emerge as prominent bearers of defence systems. The differential prevalence of these traits in *Salmonella* can be attributed to the distinct plasmid types prevalent in each lineage. For instance, IncA/C is mostly found in *S*. Typhimurium and confers resistance against mercury, tetracycline and sulfonamide, while IncHI1 and IncN, found primarily in *S*. Typhi, exhibit resistance against tetracycline and beta-lactams (Holt *et al*., 2007).

Our study also reveals the involvement of prophages in contributing to pathogenicity-associated gene patterns within *Salmonella*, especially in the case of anti-phage defence systems. Moreover, specific phage genera are linked to the dissemination of other factors, with *Brunovirus* and *Peduovirus* frequently carrying virulence genes, *Traversvirus* carrying stress resistance genes, and *Xuanwuvirus* harbouring some ABR genes. Overall, our findings underscore the multifaceted contributions of plasmids and prophages in shaping the pathogenicity of diverse *Salmonella* lineages.

In the broader context, these findings offer a novel perspective on deciphering the evolutionary trajectory of *Salmonella*. By uncovering the complex relationships between gene clusters, RGP, and pathogenic attributes, we gain deeper insight into mechanisms driving the emergence of diverse *Salmonella* lineages. This knowledge not only enriches our understanding of the evolution of *Salmonella* but also holds promise for predicting its future adaptations and developing targeted interventions to combat infections.

Chapter 3

Studying the phylogenomics to understand the CRISPR-Cas diversity in *Salmonella*

Publications from this objective-

1. Kushwaha SK, Bhavesh NLS, Abdella B, Lahiri C, Marathe SA. The phylogenomics of CRISPR-Cas system and revelation of its features in *Salmonella*. Scientific Report. 2020 Dec 3;10(1):21156. DOI: 10.1038/s41598-020-77890-6. PMID: 33273523.
2. Kushwaha SK, Kumar AA, Gupta H, Marathe SA. The Phylogenetic Study of the CRISPR-Cas System in *Enterobacteriaceae*. Current Microbiology. 2023 Apr;80(6):196. DOI: 10.1007/s00284-023-03298-w. PMID: 37118221.

3.1 Introduction

Genus *Salmonella* is classified into two species, *Salmonella enterica (S. enterica)* and *S. bongori*. *S. enterica* evolved into six subspecies (subsp.), namely, *enterica*, *salamae*, *arizonae*, *diarizonae*, *houtenae* and *indica* (Lamas *et al.,* 2018). The host range for serovars of *S. enterica* subsp. *enterica* vary from broad-host-range to host-adapted and host-restricted (Gao *et al.,* 2017), pertinent to within-host evolution (Ilyas, Tsai, & Coombes, 2017). Before divergence, *S. bongori* and *S. enterica* acquired *Salmonella* pathogenicity island 1 (SPI-1) (Gal-Mor, 2019), and later, *S. enterica* laterally acquired SPI-2, thereby enhancing its virulence potential (Gal-Mor, 2019). As per the adopt-adapt model of bacterial speciation (Sheppard, Guttman, & Fitzgerald, 2018), the adopted lateral gene(s) divert the evolutionary path, promoting bacterial adaptation and consequently increasing its fitness (Brooks, Turkarslan, Beer, Lo, & Baliga, 2011). Over time, both species horizontally acquired multiple virulence factors, progressively enhancing their pathogenicity (Ilyas *et al*., 2017).

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and a set of CRISPR-associated (*cas*) genes are suggested to be acquired by horizontal gene transfer (HGT) event (Touchon & Rocha, 2010; McDonald, Regmi, Morreale, Borowski, & Boyd, 2019). The Cas1 and Cas2 proteins are essential for spacer acquisition from invading mobile genetic elements (MGE) (Lamas *et al*., 2018), while all Cas proteins participate in primed adaptation to update the invaders' memory (Krivoy *et al*., 2018). The newly acquired spacers are added at the leader proximal end of the CRISPR array (Lamas *et al*., 2018). The Cas proteins work in conjunction with the CRISPR-RNA to carry out the interference step (Gao *et al*., 2017).

S. enterica possesses a type I–E CRISPR system comprising a *cas* operon and two CRISPR arrays, CRISPR1 and CRISPR2 (Karimi, Ahmadi, Najafi, & Ranjbar, 2018), separated by ~16 kb (Shariat, Timme, Pettengill, Barrangou, & Dudley, 2015). The *cas* operon in proximity to the CRISPR1 array (Koonin & Makarova, 2019) contains eight *cas* genes. Two distinct *cas* gene profiles have been observed with reported incongruence between the *cas* and whole genome phylogeny (Pettengill *et al*., 2014). Similar nonconformity is noted for the CRISPR array (Timme *et al*., 2013). Contrarily, a phylogenetic congruence of the CRISPR loci and whole genome was obtained for strains of *S. enterica* serovar Gallinarum biovar Pullorum (Xie *et al*., 2017). Fricke *et al.,* observed a partial correlation between the

CRISPR arrays and the phylogeny of *S. enterica* isolates (Fricke *et al*., 2011). Studies on the phylogeny of CRISPR-Cas system have been done in other bacteria as well, suggesting its role in shaping the accessory genome (van Belkum *et al*., 2015). The phylogenetic analysis of *Shigella* and *E. coli* indicates a similarity in the terminal repeats between the two species (Yang *et al*., 2015). The number of CRISPR arrays is negatively correlated with the pathogenic potential of *Escherichia coli,* where the reduction in CRISPR activity is proposed to promote HGT, favouring its evolution (García-Gutiérrez, Almendros, Mojica, Guzmán, & García-Martínez, 2015). Conversely, some reports have demonstrated a positive correlation between CRISPR and pathogenicity owing to the virulence genes regulation (Sampson & Weiss, 2014; R. Li *et al*., 2016; Cui *et al*., 2020).

To test the association of the CRISPR-Cas system with the serovar host/habitat diversity, we studied the evolutionary pattern of the CRISPR-Cas system across strains of *Salmonella*. The strains assorted into two groups with respect to the CRISPR1-leader and *cas* operon features. This divergence was analysed in comparison to multi-locus sequence typing (MLST) based on the seven housekeeping genes. Spacer versatility was assessed with respect to the protospacer source. Additionally, we studied the evolution of the CRISPR-Cas system within *Enterobacteriaceae* by analysing the CRISPR-Cas phylogeny among six genera. The phylogenetics of all CRISPR-Cas components was investigated and compared with that of the housekeeping gene, *gyrB.* Further insights into the evolution of the CRISPR-Cas system were obtained by mapping the protospacer sources of the CRISPR spacers.

3.2 Materials and Methods

3.2.1 Sequence data collection

Our study comprises 133 strains belonging to two species, *S. bongori* and *S. enterica,* including 22 serovars and three subspecies. These samples were primitively isolated from multiple sources, including primates, poultry, swine, cattle, food specimens, and the natural environment (GenBank database). The complete genome sequences for all these annotated strains were obtained from the GenBank database. Only experimentally validated sequences were considered to ensure the legitimacy of the data being used. For MLST, sequences of seven housekeeping genes, namely*, purE, hemD, aroC, dnaN, hisD, thrA* and *sucA* were retrieved from BIGSdb software (Jolley, Bray, & Maiden, 2018), and the unannotated ones were extracted from the genome's annotation files using a customised written bash script. The composite sequence tags were allocated for the allelic profiles of these seven genes.

The genome sequences of 146 strains comprising six *Enterobacteriaceae* species-*Salmonella, Escherichia*, *Klebsiella*, *Shigella, Citrobacter,* and *Cronobacter* were obtained from the National Center for Biotechnology Information database. The completeness of these sequences was verified using BUSCO analysis (Manni, Berkeley, Seppey, Simão, & Zdobnov, 2021). Genomes with BUSCO scores greater than 95% were considered. These genome sequences were annotated using Prokka (Seemann, 2014), and the *gyrB* sequences were extracted. Further analysis was carried out on 39 shortlisted strains.

3.2.2 Analysis of the CRISPR-Cas components

The CRISPR and *cas* loci of all the strains were obtained in the correct orientation after retrieving the data from GenBank and CRISPR-Cas++ database (Couvin *et al*., 2018) and verified using the CRISPR-Cas Typer (Russel, Pinilla-Redondo, Mayo-Muñoz, Shah, & Sørensen, 2020). The upstream and downstream regions of these arrays were aligned with the leader sequences previously reported by (Couvin *et al*., 2018) to know the correct sequence of the CRISPR array. The arrays were then classified as CRISPR1 and CRISPR2 after verifying the leader sequence and its position with respect to the *cas* operon.

To create spacer maps of the CRISPR arrays, the spacers were aligned, and similarity was calculated. The intra- and inter-serovar spacer conservation was estimated using Python scripts. A similarity of 90% was considered to maximise their homology to construct the spacer map. The orientation of the individual *cas* genes was traced, and the sequence similarity was calculated using a custom Python script. The amino acid sequences of Cse1 and the essential domains of Cas3 protein (HD domain, helicase C terminal domain, and the DEAD-box) of *Salmonella* were extracted from the UniProt database and aligned with the reported sequences of *E. coli* using the tool Clustal Omega.

Most strains of *S. enterica* subsp. *enterica* had both the CRISPR arrays. However, all the analysed strains of *S. enterica* subsp. *enterica* serovar Heidelberg, a few strains of serovar Typhimurium, and one strain of serovar Tennessee are reported to harbour more than two CRISPR arrays (Couvin *et al*., 2018). Instead, our analysis confirmed that the CRISPR1 array of serovars Typhimurium and Heidelberg were divided into two parts by a stretch of 74 nucleotides consisting of two truncated spacers and a direct repeat (DR). The two parts of the CRISPR1 array taken together in concatenation aligned well with the intact CRISPR1 array of other strains of serovar Typhimurium. Similarly, the CRISPR1 array of serovar Tennessee strain (str.) CFSAN070645 was divided into three parts (containing 19, 24, and 16 spacers) and the CRISPR2 into two parts (consisting of 10 and 11 spacers) due to the presence of mutated DRs rendering a stretch of 91 bp undetectable as a part of the CRISPR array. Therefore, we considered the concatenated forms of these CRISPR arrays as a single unit for further analysis.

Our analysis also indicated the occurrence of the CRISPR1 array with two spacers each in the serovars Dublin, Gallinarum, Pullorum, and Gallinarum/Pullorum. However, neither of these CRISPR arrays was described as valid in the CRISPR-Cas++ database, and the CRISPRCasFinder software allocated 27 bp long DRs and 34 bp long spacer sequences. Likewise, the CRISPR2 arrays of serovar Typhi and serovar Pullorum str. S06004 identified through our analysis was not detectable by this database. The CRISPR2 array of serovar Typhi possessed only one erratic spacer and that of serovar Pullorum str. S06004 had two spacers. We considered all these strains and their respective CRISPR-Cas systems in our analysis.

The sequence logo for the CRISPR leader and DR sequences was generated using the tool WEBLOGO ver. 2.8.2 (Crooks, Hon, Chandonia, & Brenner, 2004). The MGEs were manually checked 50 kb upstream and downstream of each CRISPR loci using the annotated GenBank files. Further, the GC content of the CRISPR-Cas components and the whole genome was computed using Python script.

3.2.3 Phylogenetic analysis

For the CRISPR leader and *cas3* (for *Salmonella* and inter-genus analysis), *cas* operon (for *Salmonella*), CRISPR1 consensus DR sequences, *cas1-2* (for inter-genus analysis), and *gyrB* multiple sequence alignment was performed on the aforesaid sequences by MUSCLE version 3.6 with default parameters (Edgar, 2004) integrated into Molecular Evolutionary Genetics Analysis version 10 (MEGA X) (Kumar, Stecher, Li, Knyaz, & Tamura, 2018). All positions with alignment gaps and missing data were excluded (complete deletion option). The resulting alignments of respective groups of sequences was used to construct each phylogenetic tree using the Maximum Likelihood (ML) method

(Jin, Nakhleh, Snir, & Tuller, 2007) guided by the most suitable evolutionary model proposed by the Bayesian approach (Tamura, 1992). The trees were given confidence with a bootstrap value of 1000 iterations. The substitution models and the parameters used for the reconstructed trees were the Tamura-Nei model with Gamma distribution for MLST, Tamura 3-parameter model for CRISPR1-leader and CRISPR2-Leader and Kimura2-parameter model along with gamma distribution for concatenated *cas* genes, *cas1-2* genes, *cas3* and *gyrB* gene. The Newick format of the trees was used for further visualisation and analyses through MEGA X. All trees were drawn to scale, and the branch lengths were calculated as the number of substitutions per site.

The phenograms for the CRISPR1 and CRISPR2 spacers (for *Salmonella*) were constructed based on the presence-absence matrix. The spacers for each strain were considered present if they had 90% sequence similarity. Using this, a Jaccard similarity matrix was created. The Jaccard distance was computed based on this matrix, and the phenogram was created using the neighbour-joining method in MEGAX (Kumar *et al.*, 2018). A pairwise distance matrix was constructed for the CRISPR1 consensus DR (for inter-genus analysis), and the phylogenetic trees were built using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) method. The bootstrap confidence level was 1000 iterations. The trees were visualised and annotated using the R package ggtree (Yu, 2020). High-resolution images were obtained using the tool Interactive Tree of Life (iTOL) (Letunic & Bork, 2019).

3.2.4 Protospacer Analysis

The spacer sequences for a particular serovar were extracted from the CRISPR-Cas++ database in the fna format, and the data of all the strains were combined to obtain a unique set of spacers. The files were then uploaded to the CRISPRTarget tool (Biswas, Gagnon, Brouns, Fineran, & Brown, 2013) to get the protospacer target hits. The data was extracted from Genbank Phage, RefSeq-Plasmid and IMGVR databases. The parameters for the initial BLAST screen in CRISPRTarget were kept default. The output obtained gave the accession number of the protospacer sources corresponding to these spacers. The hits obtained for Genbank Phage and RefSeq-Plasmid had accession numbers corresponding to NCBI. While the accession number for the hits obtained from the IMGVR database corresponds to the IMG/VR viral resource. Hits were chosen with the percentage identity

>95% and the bit score >50. The accession numbers of the protospacer hits obtained were matched across serovars using a customised Python script. Based on these matches, a heat map was created using GraphPad Prism v9.2.0.

3.3 Results

3.3.1 Diversity of the CRISPR arrays in *Salmonella*

We extracted all possible CRISPR1 and CRISPR2 arrays in the correct orientation for 133 *Salmonella* strains. *S. bongori* and *S. enterica* subsp. *enterica* contained both CRISPR arrays while subsp. *arizonae* and *diarizonae,* had only one array. One of the six examined strains of subspecies *arizonae* had an intact CRISPR array. We mapped the spacer sequences (**Fig. 3.1, see Appendix-II**) of all strains, illustrating the blueprint of spacer conservation among the strains within and across the serovars. The acquisition of spacers is in a precise fashion with the conservation of spacer arrangement for a specific serovar. However, a few spacers are absent from the CRISPR array(s) of some strains. The spacers of serovars Enteritidis, Heidelberg, and Typhi are highly conserved among their respective strains, whereas the serovars Typhimurium, Newport, Anatum, Montevideo, and Tennessee had significant variability in the spacer composition (**Fig. 3.1**). Among all strains, we identified 440 and 330 unique spacers within the 2,221 and 2,211 spacers of CRISPR1 and CRISPR2 arrays, respectively. The average abundance of spacers for CRISPR1 and CRISPR2 is 15.3 and 12.6, respectively (**Table 3.1**). CRISPR1 array of serovar Tennessee str. ATCC 10722 (63 spacers) and CRISPR2 array of serovar Typhimurium str. USDA-ARS-USMARC-1880 (35 spacers) are the largest. CRISRP1 array of serovar Dublin, Gallinarum, Pullorum and, Gallinarum/Pullorum (two spacers), and CRISPR2 array of serovars Sendai and Typhi (one spacer) are the shortest (**Fig. 3.1**). We observed duplication and triplication of spacer(s) in some serovars (**Fig. 3.2**).

Strikingly, the analysis of the CRISPR arrays in serovars Montevideo and Saintpaul separated the respective strains into two groups, each with two distinct sets of unique and conserved spacers. For serovar Montevideo, the two groups comprised eight (later defined as Montevideo-STM) and nine strains (later defined as Montevideo-STY). However, CRISPR arrays of all the analysed strains of serovar Saintpaul (that we define as Saintpaul-STM), except strain SARA26 (an outlier, defined as Saintpaul-STY), had similar spacer compositions. This suggests that the serovars Montevideo and Saintpaul could be

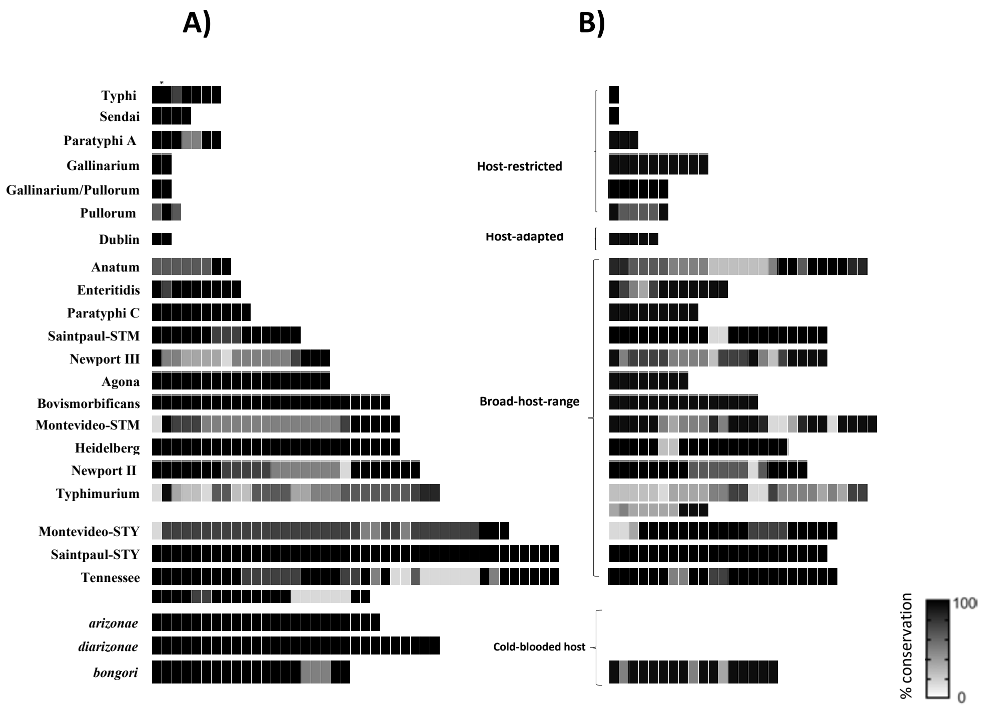


Figure 3.1 Graphic map of spacer conservation in A) CRISPR1 and B) CRISPR2 array for *Salmonella* serovars. The shades of grey represent the conservation percentage of a given spacer in all the strains of the respective serovar, where the darker box indicates the presence of a spacer in most of the strains (black: 100%), while the lighter box indicates the presence of spacer in a few strains. * indicates the merging of two spacers in a few strains of serovar Typhi.

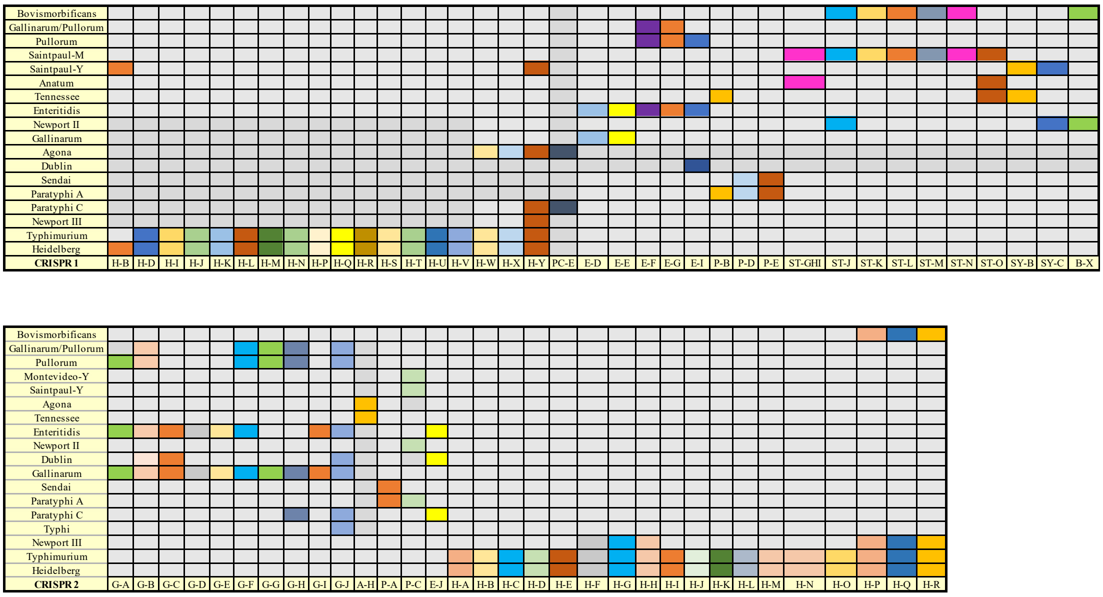


Figure 3.2 Inter-serovar spacer conservation of various serovars of *Salmonella* in the CRISPR1 and CRISPR2 arrays. The colour code for a particular column represents spacer sequences with greater than 90% nucleotide similarity. The DRs have been eliminated for simplicity.

Table 3.1 The statistics of the spacer index for the serovars under consideration

			No. of strains analysed	CRISPRI*			CRISPR2*		
				Minimum	Maximum	Average	Minimum	Maximum	Average
Salmonella enterica subsp. *enterica*	Host-restricted	Typhi	7	6	7	6.14	1	1	1
		Paratyphi A	3	5	7	6.33	3	3	3
		Gallinarum	2	2	2	2	10	10	10
		Pullorum	2	2	2	2	2	6	4
		Gallinarum/Pullorum	2	2	2	2	6	6	6
		Sendai	1	4	4	4	1	1	1
	Host adapted	Dublin	1	2	2	2	5	5	5
	Broad-host range	Anatum	13	2	8	4.77	8	26	20.25
		Paratyphi C	1	10	10	10	9	9	9
		Heidelberg	6	25	25	25	16	18	17.67
		Newport II	10	17	26	24.4	12	19	18.3
		Newport III	6	4	18	11.33	10	20	16.83
		Enteritidis	20	8	9	8.95	8	12	11.35
		Typhimurium	23	8	28	17.95	15	35	25.39
		Bovismorbificans	1	24	24	24	15	15	15
		Tennessee	7	41	63	52.43	21	23	22.14
		Montevideo	17	4	36	23.59	16	25	19.94
		Saintpaul	5	11	4 1	19.4	20	22	20.8
		Agona	1	18	18	18	8	8	8
	Cold-blooded hosts	*Salmonella enterica* subsp. *arizonae*	1	23	23	23	-	-	-
		Salmonella enterica subsp. *diarizonae*	1	29	29	29	-	-	-
		Salmonella bongori	3	20	20	20	17	17	17
Total						15.29			12.58

polyphyletic with respect to CRISPR1 loci, like that reported for serovar Newport (Zheng *et al.*, 2017). The broad-host-range serovars have multiple spacers, while the host-specific serovars have few spacers (**Fig. 3.1**).

The DR sequence is conserved within the respective array across all the serovars- 5' GTGTTCCCCGCGCCAGCGGGGATAAACCG 3' except for a few SNPs like the CRISPR1-STM contains C/T at the 14th position. The last DR is degenerated (Richter, Chang, & Fineran, 2012) with significant variation near the 3′ end.

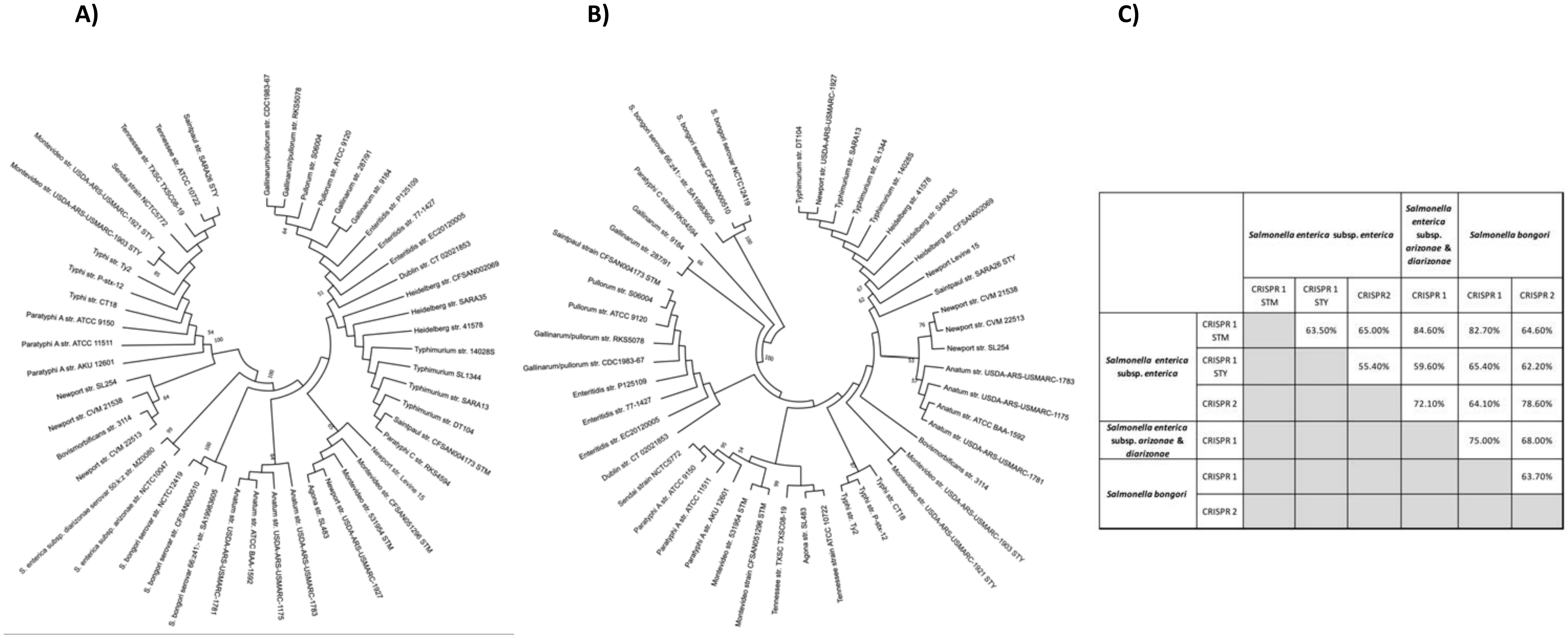
3.3.2 Phylogeny and classification of the CRISPR loci

Further analysis was performed on 49 shortlisted strains representing different species, subspecies and serovars with varied host ranges. A minimum number of strains of each serovar were chosen to represent almost all combinations of the spacers. To understand the evolutionary pattern of *Salmonella* serovars concerning the CRISPR loci, we generated phylogenetic trees for the leader sequences and spacers.

3.3.2.1 Evolutionary studies of the CRISPR leader

For the leader phenogram, the topology has been observed in most clades and sub-clades, as evidenced by their high confidence level from either the bootstrap values or the aLRT (approximate likelihood ratio test) scores. The CRISPR1-leader tree had two distinct clades, comprising typhoidal and non-typhoidal *Salmonella* serovars (Pettengill *et al.*, 2014) (**Fig. 3.3A**). Thus, we classified the corresponding CRISPR loci as CRISPR1-STM and CRISPR1-STY. The strains of serovars Saintpaul and Montevideo harbouring these loci were accordingly defined as Saintpaul-STM/Montevideo-STM and Saintpaul-STY/Montevideo-STY. The CRISPR1-STM clade included strains that are host-adapted, host-restricted or have a broad host range (**Fig. 3.3A**) (Anderson & Kendall, 2017). The CRISPR1-STY/*cas*-STY clade also contains the serovars Montevideo, Newport-II, Tennessee, Bovismorbificans and Saintpaul having broad-host-range (Jones *et al.*, 2008; Andino & Hanning, 2015) and association with outbreaks of human salmonellosis (Sheth *et al.*, 2011; Brandwagt *et al.*, 2018; Plumb *et al.*, 2019).

In the CRISPR2-leader phenogram (**Fig. 3.3B**), *S. bongori* emerged as an outgroup for the entire tree, and serovar Paratyphi C seems to have evolved distinctly from other serovars of *S. enterica* subsp. *enterica*. The topology and the sub-lineage were very distinct


A)
B)
C)
Saintpaul str. SARA26 STY
Tennessee str. ATCC 10722
Tennessee str. TXSC TXSC08-19
Sendai strain NCTC5772
Montevideo str. USDA-ARS-USMARC-1921 STY
Montevideo str. USDA-ARS-USMARC-1903 STY
Typhi str. Ty2
Typhi str. P-stx-12
Typhi str. CT18
Paratyphi A str. ATCC 9150
Paratyphi A str. ATCC 11511
Paratyphi A str. AKU 12601
Newport str. SL254
Newport str. CVM 21538
Bovismorbificans str. 3114
Newport str. CVM 22513
S. enterica subsp. diarizonae serovar 50:k:z str. MZ0080
S. enterica subsp. arizonae str. NCTC10047
S. bongori serovar str. NCTC12419
S. bongori serovar str. CFSAN000510
S. bongori serovar 66:z41:- str. SA19983605
Anatum str. USDA-ARS-USMARC-1781
Anatum str. ATCC BAA-1592
Anatum str. USDA-ARS-USMARC-1175
Anatum str. USDA-ARS-USMARC-1783
Agona str. SL483
Newport str. USDA-ARS-USMARC-1927
Montevideo str. 531954 STM
Montevideo str. CFSAN051296 STM
Newport str. Levine 15
Paratyphi C str. RKS4594
Saintpaul str. CFSAN004173 STM
Typhimurium str. DT104
Typhimurium str. SARA13
Typhimurium SL1344
Typhimurium str. 14028S
Heidelberg str. 41578
Heidelberg str. SARA35
Heidelberg str. CFSAN002069
Dublin str. CT 02021853
Enteritidis str. EC20120005
Enteritidis str. 77-1427
Enteritidis str. P125109
Gallinarum str. 9184
Gallinarum str. 287/91
Pullorum str. ATCC 9120
Pullorum str. S06004
Gallinarum/pullorum str. RKS5078
Gallinarum/pullorum str. CDC1983-67
100
54
100
64
95
100
94
62
51
64
85
S. bongori serovar 66:z41:- str. SA19983605
S. bongori serovar CFSAN000510
S. bongori serovar NCTC12419
Paratyphi C strain RKS4594
Gallinarum str. 9184
Gallinarum str. 287/91
Saintpaul strain CFSAN004173 STM
Pullorum str. S06004
Pullorum str. ATCC 9120
Gallinarum/pullorum str. RKS5078
Gallinarum/pullorum str. CDC1983-67
Enteritidis str. P125109
Enteritidis str. 77-1427
Enteritidis str. EC20120005
Dublin str. CT 02021853
Sendai strain NCTC5772
Paratyphi A str. ATCC 9150
Paratyphi A str. ATCC 11511
Paratyphi A str. AKU 12601
Montevideo str. 531954 STM
Montevideo strain CFSAN051296 STM
Tennessee str. TXSC TXSC08-19
Agona str. SL483
Tennessee strain ATCC 10722
Typhi str. Ty2
Typhi str. P-stx-12
Typhi str. CT18
Montevideo str. USDA-ARS-USMARC-1921 STY
Montevideo str. USDA-ARS-USMARC-1903 STY
Bovismorbificans str. 3114
Anatum str. USDA-ARS-USMARC-1781
Anatum str. ATCC BAA-1592
Anatum str. USDA-ARS-USMARC-1175
Anatum str. USDA-ARS-USMARC-1783
Newport str. SL254
Newport str. CVM 22513
Newport str. CVM 21538
Saintpaul str. SARA26 STY
Newport Levine 15
Heidelberg str. CFSAN002069
Heidelberg str. SARA35
Heidelberg str. 41578
Typhimurium str. 14028S
Typhimurium str. SL1344
Typhimurium str. SARA13
Newport str. USDA-ARS-USMARC-1927
Typhimurium str. DT104
100
100
66
95
54
99
87
76
53
57
62
52

		Salmonella enterica subsp. *enterica*			*Salmonella enterica* subsp. *arizonae & diarizonae*	*Salmonella bongori*	
		CRISPR 1 STM	CRISPR 1 STY	CRISPR2	CRISPR 1	CRISPR 1	CRISPR 2
Salmonella enterica subsp. *enterica*	CRISPR 1 STM		63.50%	65.00%	84.60%	82.70%	64.60%
	CRISPR 1 STY			55.40%	59.60%	65.40%	62.20%
	CRISPR 2				72.10%	64.10%	78.60%
Salmonella enterica subsp. *arizonae & diarizonae*	CRISPR 1					75.00%	68.00%
Salmonella bongori	CRISPR 1						63.70%
	CRISPR 2						

Figure 3.3 The phylogeny and conservation of CRISPR-leaders. A) CRISPR1 and **B)** CRISPR2 across *Salmonella* serovars. **C)** A matrix depicting the inter-species and inter-subspecies conservation of the leader sequence of both the CRISPR arrays. The values represent the percentage nucleotide identity with respect to the entire query cover.

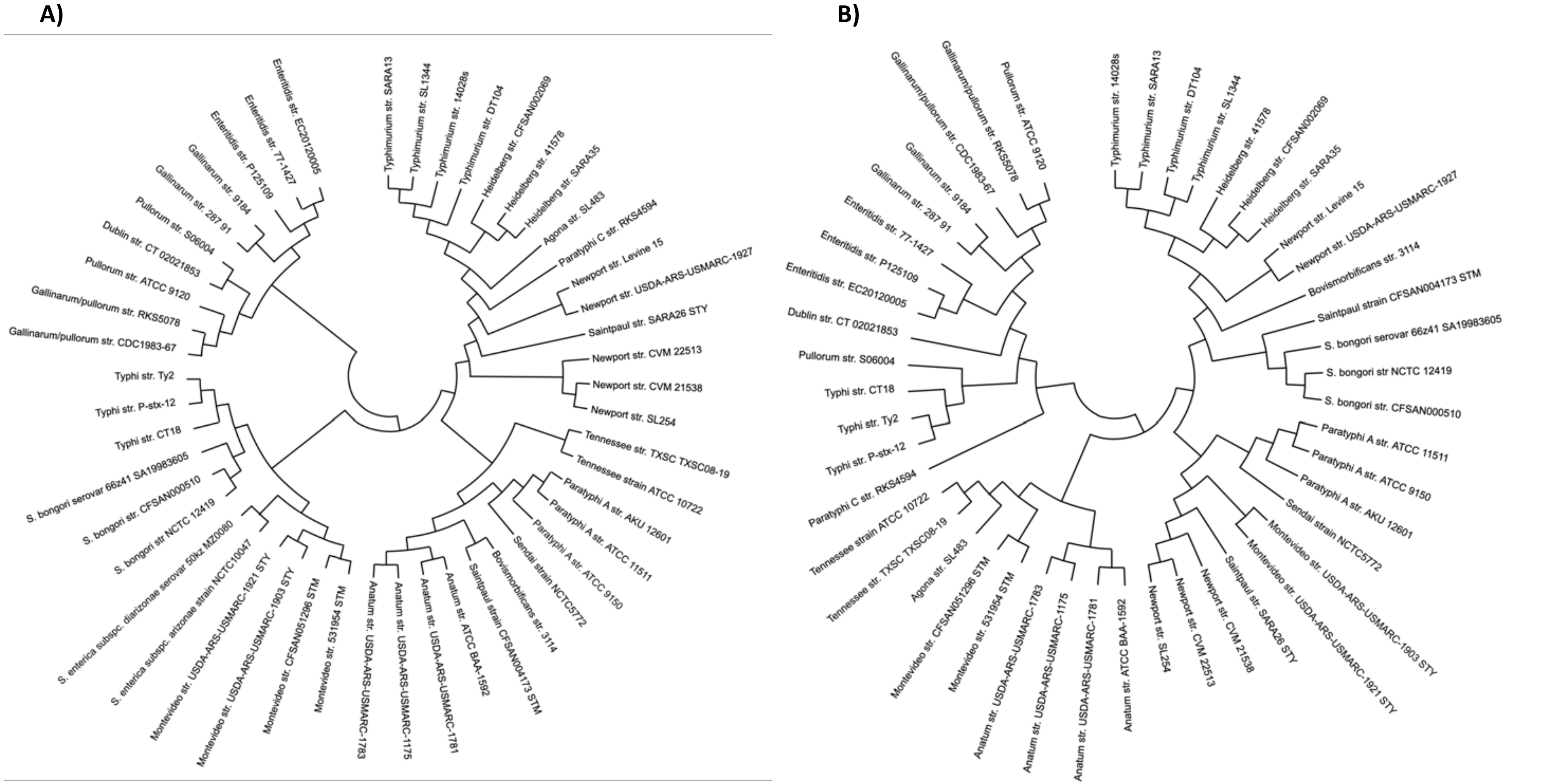

A)
Typhimurium str. SARA13
Typhimurium str. SL1344
Typhimurium str. 14028s
Typhimurium str. DT104
Heidelberg str. CFSAN002069
Heidelberg str. 41578
Heidelberg str. SARA35
Agona str. SL483
Paratyphi C str. RKS4594
Newport str. Levine 15
Newport str. USDA-ARS-USMARC-1927
Saintpaul str. SARA26 STY
Newport str. CVM 22513
Newport str. CVM 21538
Newport str. SL254
Tennessee str. TXSC TXSC08-19
Tennessee strain ATCC 10722
Paratyphi A str. AKU 12601
Paratyphi A str. ATCC 11511
Paratyphi A str. ATCC 9150
Sendai strain NCTC5772
Bovismorbificans str. 3114
Saintpaul strain CFSAN004173 STM
Anatum str. ATCC BAA-1592
Anatum str. USDA-ARS-USMARC-1781
Anatum str. USDA-ARS-USMARC-1175
Anatum str. USDA-ARS-USMARC-1783
Montevideo str. 531954 STM
Montevideo str. CFSAN051296 STM
Montevideo str. USDA-ARS-USMARC-1903 STY
Montevideo str. USDA-ARS-USMARC-1921 STY
S. enterica subspc. arizonae strain NCTC10047
S. enterica subspc. diarizonae serovar 50kz MZ0080
S. bongori str NCTC 12419
S. bongori str. CFSAN000510
S. bongori serovar 66z41 SA19983605
Typhi str. CT18
Typhi str. P-stx-12
Typhi str. Ty2
Gallinarum/pullorum str. CDC1983-67
Gallinarum/pullorum str. RKS5078
Pullorum str. ATCC 9120
Dublin str. CT 02021853
Pullorum str. S06004
Gallinarum str. 287 91
Gallinarum str. 9184
Enteritidis str. P125109
Enteritidis str. 77-1427
Enteritidis str. EC20120005
B)
Typhimurium str. 14028s
Typhimurium str. SARA13
Typhimurium str. DT104
Typhimurium str. SL1344
Heidelberg str. 41578
Heidelberg str. CFSAN002069
Heidelberg str. SARA35
Newport str. Levine 15
Newport str. USDA-ARS-USMARC-1927
Bovismorbificans str. 3114
Saintpaul strain CFSAN004173 STM
S. bongori serovar 66z41 SA19983605
S. bongori str NCTC 12419
S. bongori str. CFSAN000510
Paratyphi A str. ATCC 11511
Paratyphi A str. ATCC 9150
Paratyphi A str. AKU 12601
Sendai strain NCTC5772
Montevideo str. USDA-ARS-USMARC-1903 STY
Montevideo str. USDA-ARS-USMARC-1921 STY
Saintpaul str. SARA26 STY
Newport str. CVM 21538
Newport str. CVM 22513
Newport str. SL254
Anatum str. ATCC BAA-1592
Anatum str. USDA-ARS-USMARC-1781
Anatum str. USDA-ARS-USMARC-1175
Anatum str. USDA-ARS-USMARC-1783
Montevideo str. 531954 STM
Montevideo str. CFSAN051296 STM
Agona str. SL483
Tennessee str. TXSC TXSC08-19
Tennessee strain ATCC 10722
Paratyphi C str. RKS4594
Typhi str. P-stx-12
Typhi str. Ty2
Typhi str. CT18
Pullorum str. S06004
Dublin str. CT 02021853
Enteritidis str. EC20120005
Enteritidis str. P125109
Enteritidis str. 77-1427
Gallinarum str. 287 91
Gallinarum str. 9184
Gallinarum/pullorum str. CDC1983-67
Gallinarum/pullorum str. RKS5078
Pullorum str. ATCC 9120

Figure 3.4 The phylogeny of CRISPR spacers the A) CRISPR1 and B) CRISPR2 array. Serovars Heidelberg, Newport III and Typhimurium; Paratyphi A and Sendai; and Dublin, Enteritidis, Gallinarum, Pullorum and Gallinarum/Pullorum club together in both trees and are named as HNT, PS and DEGP clade.

from that of the CRISPR1-leader tree with intermixing of serovars of the two distinct clades. For example, serovar Saintpaul-STY grouped with serovars Typhimurium, Newport-III, and Heidelberg, whereas Sendai and Paratyphi A grouped with Montevideo-STM while Newport-II clubbed with Anatum. This suggests different evolutionary trajectories of both CRISPR loci.

3.3.2.2 Categorisation of the leader sequence in the light of CRISPR leader phylogeny

The leader sequence analysis suggests serovars of *S. enterica* subsp. *enterica* have two distinct types of CRISPR1-leaders (**Fig. 3.3A**), justifying their divergence in two clades. One of the leader sequences is identical to that of Newport-II (Shariat *et al.*, 2015) and is present in all the serovars of the CRISPR1-STY clade. Serovars Enteritidis, Gallinarum, Pullorum and Gallinarum/Pullorum have <98% leader identity and, thus, cluster in the CRISPR1-leader tree (**Fig. 3.3A**). On similar grounds, other serovars cluster or separate from each other. The CRISPR1-leader of *S. bongori* and *S. enterica* subsp. *arizonae* and subsp. *diarizonae* maximally matched with that of CRISPR1-STM (**Fig. 3.3C**) and hence grouped in the CRISPR1-STM clade.

The CRISPR2-leader sequence is highly conserved (with a few SNPs) among all the serovars of *S. enterica* subsp. *enterica* (**Fig. 3.3B**) justifying their segregation from *S. bongori.* The variations due to SNPs explain the serovar clustering in the CRISPR2-leader tree. For instance, the leaders of serovars Paratyphi A and Typhi having 94% sequence similarity segregated into separate clades, while the serovars Paratyphi A and Sendai clubbed together with 100% similarity.

3.3.2.3 Evolutionary study of CRISPR arrays

The phylogeny of CRISPR arrays was studied with respect to the spacer content. Only ~8.6-9.6% of unique spacers (37/440: CRISPR1 and 32/330: CRISPR2) were shared by two or more serovars (**Fig. 3.2**). Thus, the spacer trees were constructed based on the presence-absence matrix. In both the CRISPR1- and CRISPR2- spacer trees, serovars Enteritidis, Dublin, Gallinarum, Gallinarum/Pullorum and Pullorum formed one clade (clade-DEGP) while the other serovars formed the second (**Fig. 3.4**). In CRISPR2-spacer tree, serovar Typhi and Paratyphi C grouped with clade-DEGP sharing anchor spacer with these serovars (**Fig. 3.2 & 3.4B**). The second clade had three distinct subclades with serovar composition of two (named HNT and PS) was partially constant: serovars Heidelberg, Newport-III, and Typhimurium in clade-HNT and serovars Paratyphi A and Sendai in clade-PS. Serovars within these clades (clade-DEGP) and sub-clade (clade-HNT and clade-PS) share many spacers of both arrays (**Fig. 3.2**). However, the other serovars show spacer matches with random serovars (**Fig. 3.2 & 3.4**) and cluster differently in both spacer trees. *S. enterica* subsp. *arizonae* and *diarizonae* (both possessing only CRISPR1 array) and *S. bongori* associated with poikilotherms do not form a separate clade but intermix with the serovars of *S. enterica* infecting endotherms.

In the CRISPR1-spacer tree, serovars Agona, Newport-II, Paratyphi C and Saintpaul-STY grouped with clade-HNT as they share anchor spacer with these serovars (**Fig. 3.2**). Serovars Anatum, Bovismorbificans, Saintpaul-STM and Tennessee clubbed with clade-PS, while serovars Typhi and Montevideo grouped with the species/subspecies associated with poikilotherms (**Fig. 3.4A**). In the CRISPR2-spacer tree, *S. bongori*, serovar Bovismorbificans and Saintpaul-STM grouped with clade-HNT while serovars Newport-II, Saintpaul-STY and Montevideo-STY with clade-PS as they share anchor spacer with Paratyphi A (**Fig. 3.2**). Serovars Agona, Montevideo-STM, Anatum and Tennessee formed a separate sub-clade. Serovars Anatum and Tennessee grouped in both the trees but had different relationships with other clades (**Fig. 3.4B**).

3.3.2.4 MLST phenogram and its association with the CRISPR array

MLST is a robust and widely accepted phylogenetic reflection of the species taxonomy (Pérez-Losada, Arenas, & Castro-Nallar, 2018). Hence, we generated a reference MLST tree for the shortlisted strains using concatenated allelic data of seven housekeeping genes (**Fig. 3.5**). *S. bongori* separated as a distinct clade from other *S. enterica* serovars. All other serovars formed lineages within a serovar-specific cluster depicted to have evolved together as an individual taxon, except serovar Saintpaul and Newport. Serovar Saintpaul str. SARA26 separated from all serovars of subspecies *enterica* and str. CFSAN004173 clustered with Typhimurium/Heidelberg/Newport-II group. In this light, serovar Saintpaul turns out to be polyphyletic like serovar Newport (Porwollik *et al.*, 2004). Serovar Paratyphi A is closer to serovar Typhimurium with 98.8% similarity in the seven genes than serovar Typhi (98.6% similarity). The CRISPR and MLST phenograms are discordant with respect to their topologies, thereby signify a differential evolutionary path

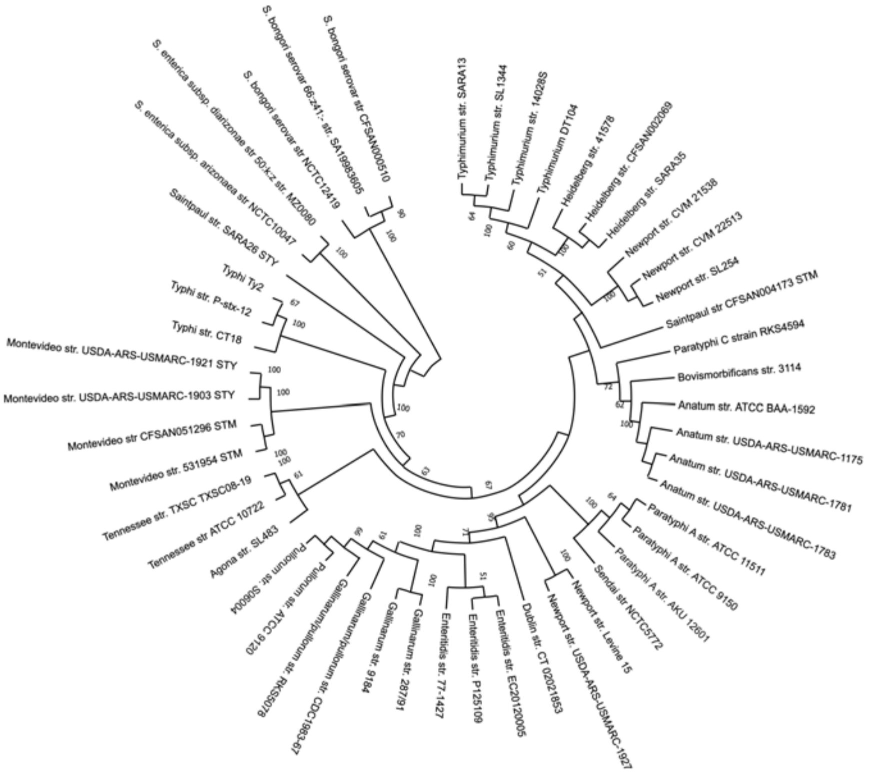

S. bongori serovar str CFSAN000510
S. bongori serovar 66:z41:- str SA19983605
S. bongori serovar str NCTC12419
S. enterica subsp. diarizonae str 50:k:z str. MZ0080
S. enterica subsp. arizonaea str NCTC10047
Saintpaul str. SARA26 STY
Typhi Ty2
Typhi str. P-stx-12
Typhi str. CT18
Montevideo str. USDA-ARS-USMARC-1921 STY
Montevideo str. USDA-ARS-USMARC-1903 STY
Montevideo str CFSAN051296 STM
Montevideo str. 531954 STM
Tennessee str. TXSC TXSC08-19
Tennessee str ATCC 10722
Agona str. SL483
Pullorum str. S06004
Pullorum str. ATCC 9120
Gallinarumpullorum str. RKS5078
Gallinarumpullorum str. CDC1983-67
Gallinarum str. 9184
Gallinarum str. 287/91
Enteritidis str. 77-1427
Enteritidis str. P125109
Enteritidis str. EC20120005
Dublin str. CT 02021853
Newport str. USDA-ARS-USMARC-1927
Newport str. Levine 15
Sendai str NCTC5772
Paratyphi A str. AKU 12601
Paratyphi A str. ATCC 9150
Paratyphi A str. ATCC 11511
Anatum str. USDA-ARS-USMARC-1783
Anatum str. USDA-ARS-USMARC-1781
Anatum str. USDA-ARS-USMARC-1175
Anatum str. ATCC BAA-1592
Bovismorbificans str. 3114
Paratyphi C strain RKS4594
Saintpaul str CFSAN004173 STM
Newport str. SL254
Newport str. CVM 22513
Newport str. CVM 21538
Heidelberg str. SARA35
Heidelberg str. CFSAN002069
Heidelberg str. 41578
Typhimurium DT104
Typhimurium str. 14028S
Typhimurium str. SL1344
Typhimurium str. SARA13

Figure 3.5 The MLST phylogeny. The phylogenetic tree was constructed using the concatenated sequences of seven housekeeping genes- *purE, hemD, aroC, dnaN, hisD, thrA*, and *sucA*.

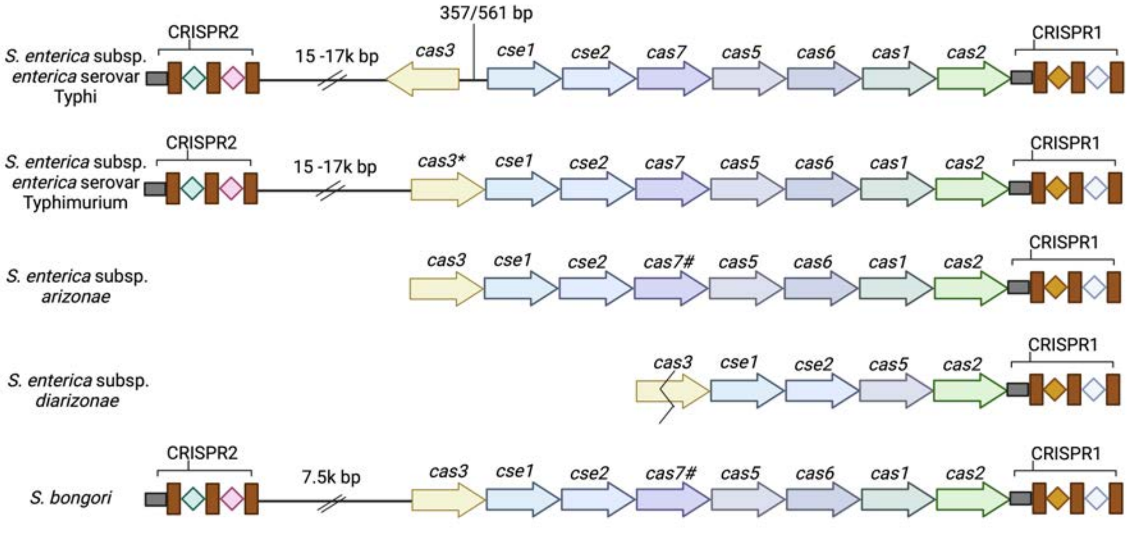


Figure 3.6 Orientation of the CRISPR array and the *cas* operon in *Salmonella*. Five types of arrangements were evident in *Salmonella*. The *cas*-STY: in strains of CRISPR1-STY clade. The *cas*-STM type operon was subdivided into four types - *S. enterica* subsp. *enterica* (*cas*-STM), *S. bongori* (*cas*-STM.B), *S. enterica* subsp. *enterica*, subsp. *arizonae* (*cas*- STM.A) and subsp. *diarizonae* (*cas*-STM.D). * indicates all the strains of serovar Montevideo-STM and *S. enterica* subps. *diarizonae* str. MZ0080 (used in our study) contain a non-sense mutation in *cas3*. # indicates *Salmonella bongori* str. SA19983605 (used in our study) does not contain the *cas7* gene and thunderbolt indicates all strains of *S. enterica* subsp. *arizonae* contain a stop codon in the *cas3* operon.

of the CRISPR loci (possibly due to a plausible acquisition of CRISPR loci through HGT) than that of the housekeeping genes. Serovars Montevideo-STM and Montevideo-STY possess the same housekeeping genes but differ in CRISPR arrays, segregating them into two groups in CRISPR phenograms.

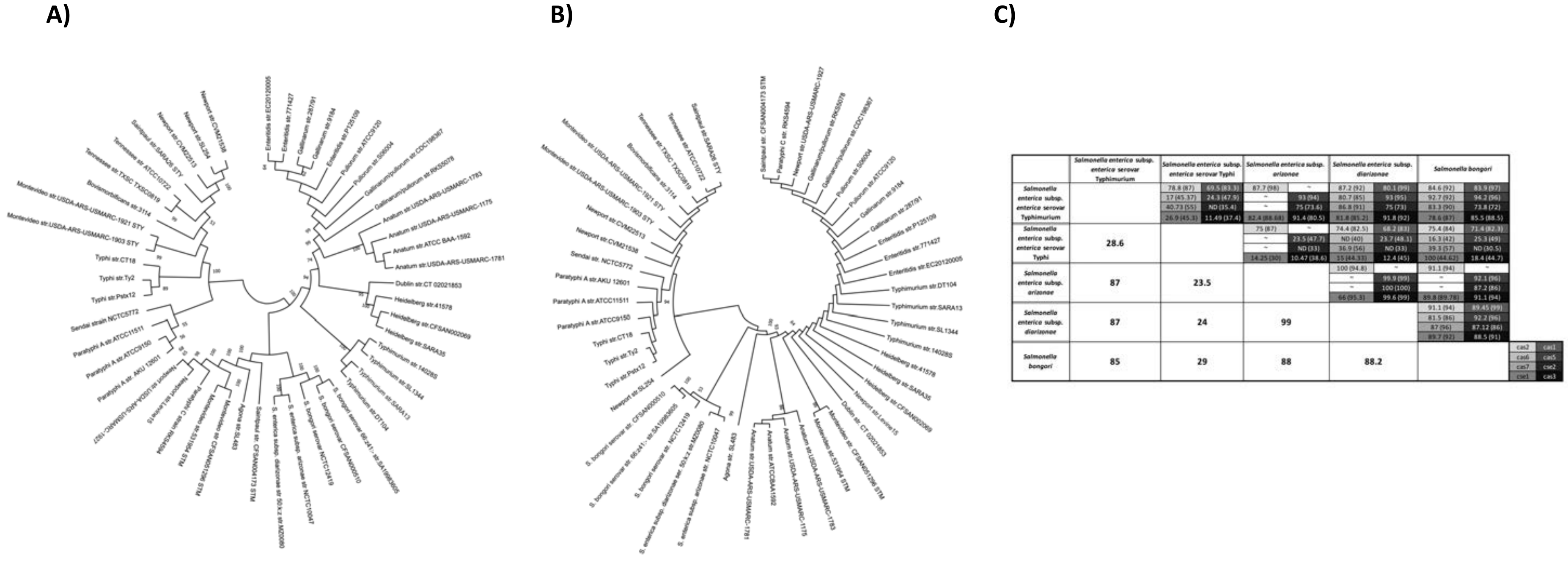
3.3.3 Phylogeny and classification of the *cas* operon

3.3.3.1 Diversification of *cas* operon and its association with the CRISPR1 array

Two distinct *cas* gene arrangements were obtained for the strains comprising CRISPR1-STY and CRISPR1-STM clades. Thus, the *cas* operon of the respective categories were denoted as *cas*-STY and *cas*-STM. For *cas*-STY, the *cas3* gene is present as a complement and is singled out from the other *cas* genes by a gap of 357 bp (561 for serovar Montevideo-STY) (**Fig. 3.6**). For *cas*-STM, the *cas* genes are contiguous but the *cas3* gene of serovar Montevideo-STM and *S. enterica* subsp. *arizonae* is degenerate, having a premature stop codon. Moreover, we noticed structural heterogeneity within the *cas*-STM operon across CRISPR1-STM strains with respect to its position in both the CRISPR loci and the *cas* gene composition (**Fig. 3.6**). The *cas* operon of *S. bongori, S. enterica* subsp. *enterica*, subsp. *arizonae* and subsp. *diarizonae* were termed as *cas*-STM.B, *cas*-STM.E, *cas*-STM.A, and *cas*-STM.D, respectively.

3.3.3.2 Evolutionary studies and conservation of *cas* operon in *Salmonella*

The *cas* operon's heterogeneity was further assessed through phylogenetic analysis of the *cas3* gene and the entire *cas* operon (**Fig. 3.7**). Two clades and the clustering of serovars obtained in both phenograms are far more analogous with the CRISPR1- leader phenogram. To gain insights into the serovar clustering in *cas* genes, we performed a detailed comparative analysis of *cas* operon. The analysis of all *cas* genes considered in concatenation revealed the highest nucleotide similarity (99%) between subspecies *arizonae* and *diarizonae* and the lowest (28.6%) between the *cas*-STM and *cas*-STY groups (**Fig. 3.7C**). Between the latter groups, *cas1* shares the highest similarity (74.4%-78.8% nucleotide and 82.5%-87% amino acid match), while *cse2* shares the lowest similarity (no significant nucleotide match and 35% amino acid identity) (**Fig. 3.7C**). The Cas3 nuclease of *cas*-STM showed poor nucleotide (10.5-18.4%) and amino acid (37.4-45%) match with the *cas*-STY category. However, the functionally important domains- HD domain (~ 48%),


A)
B)
C)
Typhi str.CT18
Typhi str.Ty2
Typhi str.Pstx12
Sendai strain NCTC5772
Paratyphi A str.ATCC11511
Paratyphi A str.ATCC9150
Paratyphi A str. AKU 12601
Dublin str.CT 02021853
Heidelberg str.41578
Heidelberg str.CFSAN002069
Heidelberg str.SARA35
Typhimurium str.14028S
Typhimurium str.SL1344
Typhimurium str.SARA13
Typhimurium str.DT104
Enteritidis str.EC20120005
Enteritidis str.771427
Enteritidis str.P125109
Gallinarum str.287/91
Gallinarum str.9184
Pullorum str.ATCC9120
Pullorum str.S06004
Gallinarum/pullorum str.RKS5078
Gallinarum/pullorum str.CDC1983-67
Anatum str.USDA-ARS-USMARC-1783
Anatum str.USDA-ARS-USMARC-1175
Anatum str.ATCC BAA-1592
Anatum str.USDA-ARS-USMARC-1781
Agona str.SL483
Newport str.SL254
Newport str.CVM22513
Newport str.CVM21538
Newport str.Levine15
Montevideo str.USDA-ARS-USMARC-1921 STY
Montevideo str.USDA-ARS-USMARC-1903 STY
Bovismorbificans str.3114
Tennessee str.TXSC TXSC08-19
Tennessee str.ATCC10722
Saintpaul str.SARA26 STY
Saintpaul str.CFSAN004173 STM
Montevideo str.CFSAN051296 STM
Montevideo str.531954 STM
Paratyphi C strain RKS4594
S. bongori serovar CFSAN000510
S. bongori serovar 66:z41:- str.SA19983605
S. bongori serovar NCTC12419
S. enterica subsp. arizonae str.NCTC10047
S. enterica subsp. diarizonae ser.50:k:z str.MZ0080

C)

	Salmonella enterica subsp. enterica serovar Typhimurium	*Salmonella enterica subsp. enterica serovar Typhi*	*Salmonella enterica subsp. arizonae*	*Salmonella enterica subsp. diarizonae*	*Salmonella bongori*
Salmonella enterica subsp. enterica serovar Typhimurium		78.8 (87) 69.5 (83.3) 17 (45.37) 24.3 (47.9) 40.73 (55) ND (35.4) 26.9 (45.3) 11.49 (37.4)	87.7 (98) ~ ~ 93 (94) ~ 75 (73.6) 82.4 (88.68) 91.4 (80.5)	87.2 (92) 80.1 (99) 80.7 (85) 93 (95) 86.8 (91) 75 (73) 81.8 (85.2) 91.8 (92)	84.6 (92) 83.9 (97) 92.7 (92) 94.2 (96) 83.3 (90) 73.8 (72) 78.6 (87) 85.5 (88.5)
Salmonella enterica subsp. enterica serovar Typhi	28.6		75 (87) ~ ~ 23.5 (47.7) ~ ND (33) 14.25 (30) 10.47 (38.6)	74.4 (82.5) 68.2 (83) ND (40) 23.7 (48.1) 36.9 (56) ND (33) 15 (44.33) 12.4 (45)	75.4 (84) 71.4 (82.3) 16.3 (42) 25.3 (49) 39.3 (57) ND (30.5) 100 (44.62) 18.4 (44.7)
Salmonella enterica subsp. arizonae	87	23.5		100 (94.8) ~ ~ 99.9 (99) ~ 100 (100) 66 (95.3) 99.6 (99)	91.1 (94) ~ ~ 92.1 (96) ~ 87.2 (86) 89.8 (89.78) 91.1 (94)
Salmonella enterica subsp. diarizonae	87	24	99		91.1 (94) 89.45 (99) 81.5 (86) 92.2 (96) 87 (96) 87.12 (86) 89.7 (92) 88.5 (91)
Salmonella bongori	85	29	88	88.2	

cas2	cas1
cas6	cas5
cas7	cse2
cse1	cas3

Figure 3.7 The phylogeny and conservation of *cas* genes. A & B Phylogeny of *cas* genes across *Salmonella* serovars for the **A)** entire *cas* operon and the **B)** *cas3* gene. **C)** Conservation of all the individual *cas* gene and Cas protein sequences. The amino acid percentage conservation is depicted in parenthesis. The term 'ND' represents no nucleotide sequence similarity based on the default parameter of the tool Nucleotide-BLAST. The values in the lower diagonal of the matrix indicate the percentage nucleotide match of the entire *cas* operon between the categories.

helicase C-terminal domain (~77%), and the DEAD-box (~81%) were similar. The *cse1* gene was quite distinct between the *cas*-STM and *cas*-STY categories. The functionally important residues of Cse1 from *E. coli* include Gly (157), glycine-loop residues (159-161), Lys (268), Asn (353), Glu (354) and Ala (355) required for the recognition of PAM sequences (Hayes *et al*., 2016) and lysine residues (289-290) for recruiting Cas3 protein (Hayes *et al*., 2016). Most of these residues are conserved across the *cas*-STM and *cas*-STY categories, indicating that although the Cse1 and Cas3 differ significantly between these serovars, their functionality remains conserved.

3.3.4 Inter-genus analysis of the CRISPR-Cas system

Next, we performed comparative sequence analysis and phylogenetics across six *Enterobacteriaceae* species: *Escherichia, Citrobacter, Cronobacter, Klebsiella, Salmonella*, and *Shigella*.

3.3.4.1 Phylogeny of *cas3*

The evolutionary relationship concerning the *cas3* gene of strains belonging to six *Enterobacteriaceae* species was analysed for 146 strains (**Fig. 3.8**). As per the phenogram, the strains were sorted into two distinct clades: (i) consisting of *Escherichia* and *Shigella* (labelled as clade ES_h), and (ii) *Cronobacter*, *E. coli* YSP8-1, *Klebsiella*, *Citrobacter* and *Salmonella*. The second clade is further segregated into two sub-clades comprising (a) *Cronobacter*, *E. coli* YSP8-1 and *S. enterica* subsp. *enterica* serovar Typhi str. CT18 and (b) *Citrobacter* and *Salmonella*.

Further analysis was performed on 39 shortlisted strains representing different species. A minimum number of strains of each species was chosen for ease of handling and interpretation but represented all clades of the phenogram. As expected, the composition of the clades remained the same except that the *Cronobacter* grouped with the *Klebsiella* and *Citrobacter* sub-clade (**Fig. 3.9A**). The clades displayed mixed arrangements, forming different sub-lineages. For example, *Cronobacter* developed a separate sub-cluster (labelled as sub-clade C_R), and all strains belonging to *K. pneumoniae* clubbed together (labelled sub-clade K_P). However, *K. oxytoca* str. AR0028*, K. michiganensis* str. K518*,* and *Klebsiella* sp. STW0522-44 clustered with *Citrobacter and S.* Typhimurium (labelled as sub-clade KC_lS_{TM}). The percentage of *cas3* nucleotide sequence

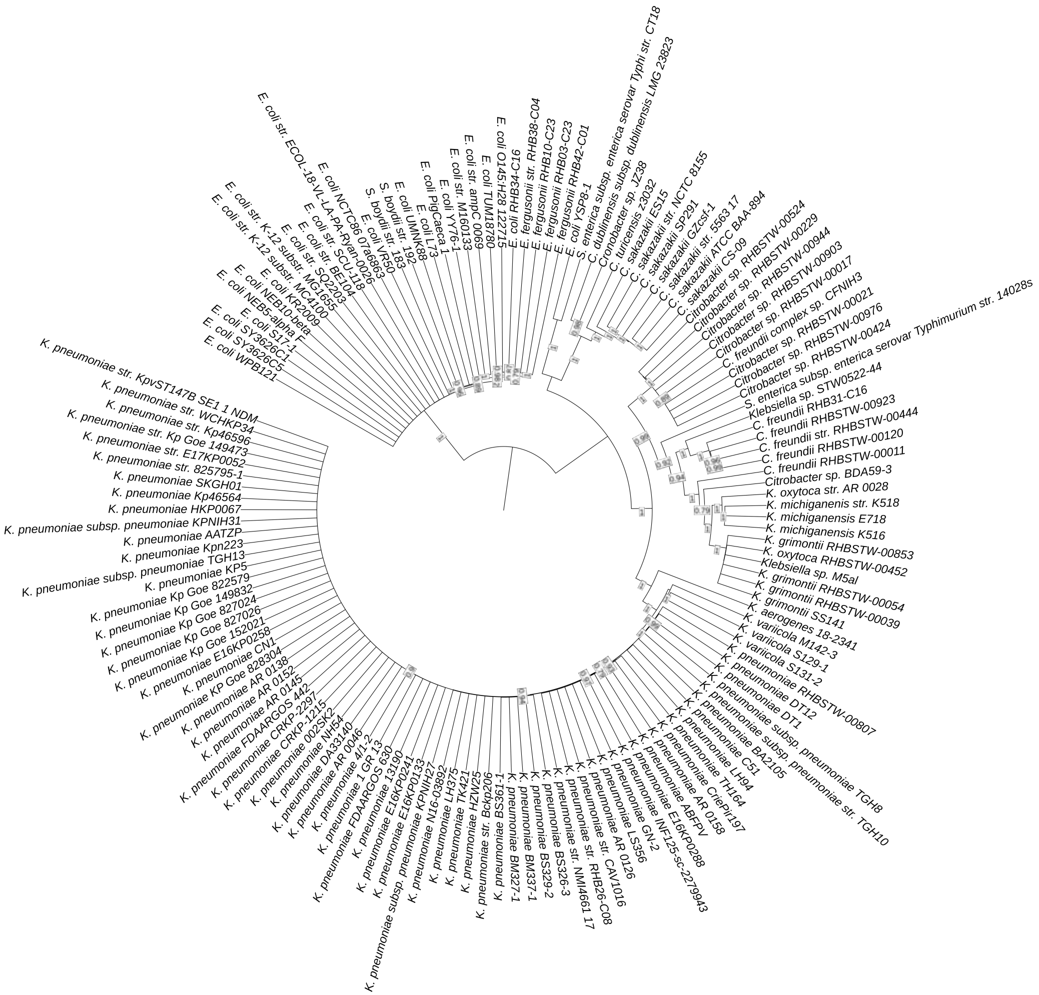


Figure 3.8 The phylogeny of *cas3* across *Enterobacteriaceae*. The CRISPR-leader sequences were aligned using MUSCLE, and the phylogenetic tree was constructed using ML.

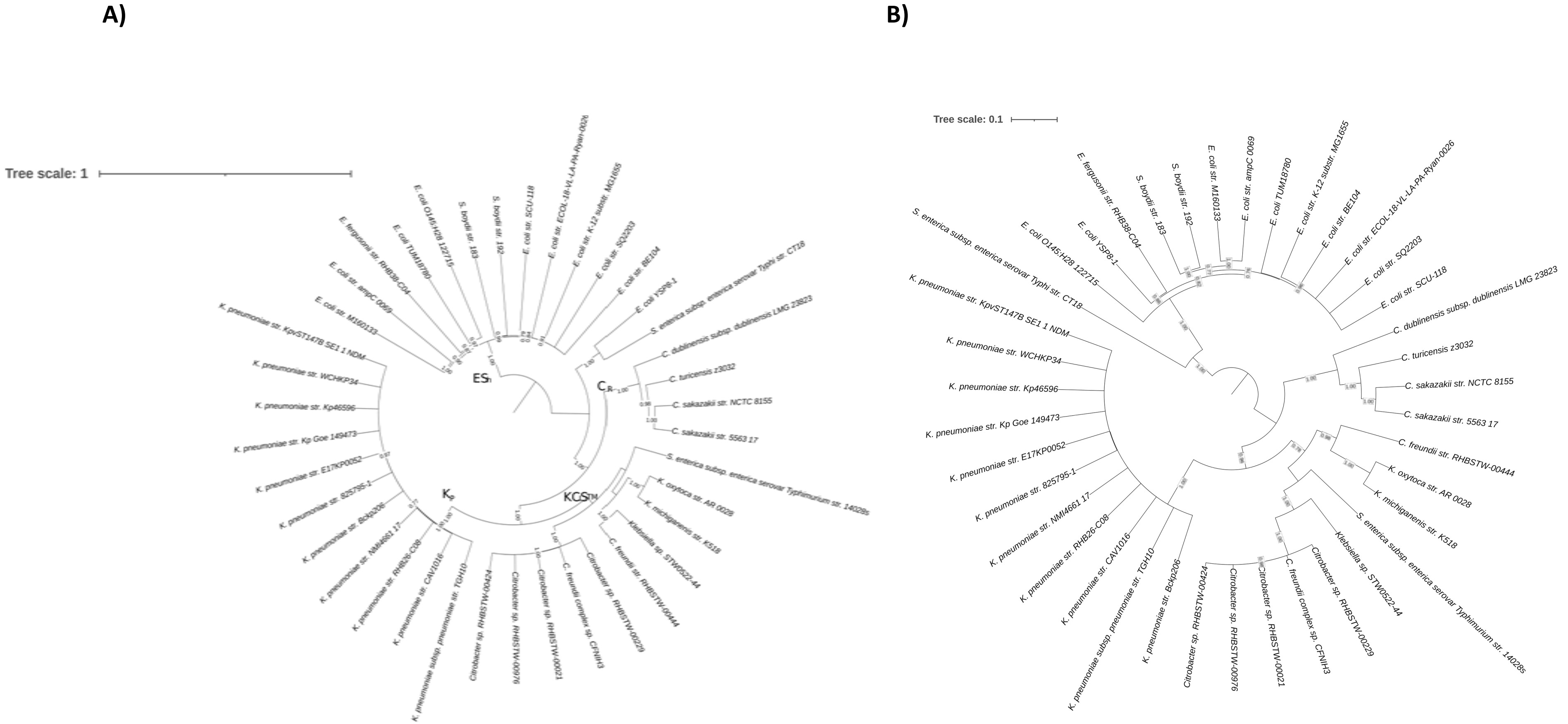

A)
Tree scale: 1
E. fergusonii str. RHB38-C04
E. coli TUM18780
E. coli O145:H28 122715
S. boydii str. 183
S. boydii str. 192
E. coli str. SCU-118
E. coli str. ECOL-18-VL-LA-PA-Ryan-0026
E. coli str. K-12 substr. MG1655
E. coli str. SQ2203
E. coli str. BE104
E. coli YSP8-1
S. enterica subsp. enterica serovar Typhi str. CT18
C. dublinensis subsp. dublinensis LMG 23823
C. turicensis z3032
C. sakazakii str. NCTC 8155
C. sakazakii str. 5563 17
S. enterica subsp. enterica serovar Typhimurium str. 14028s
K. oxytoca str. AR 0028
K. michiganensis str. K518
Klebsiella sp. STW0522-44
C. freundii str. RHBSTW-00444
Citrobacter sp. RHBSTW-00229
C. freundii complex sp. CFNIH3
Citrobacter sp. RHBSTW-00021
Citrobacter sp. RHBSTW-00976
Citrobacter sp. RHBSTW-00424
K. pneumoniae subsp. pneumoniae str. TGH10
K. pneumoniae str. CAV1016
K. pneumoniae str. RHB26-C08
K. pneumoniae str. NMI4661 17
K. pneumoniae str. Bckp206
K. pneumoniae str. 825795-1
K. pneumoniae str. E17KP0052
K. pneumoniae str. Kp Goe 149473
K. pneumoniae str. Kp46596
K. pneumoniae str. WCHKP34
K. pneumoniae str. KpvST147B SE1 1 NDM
E. coli str. M160133
E. coli str. ampC 0069
ESh
CR
KCSTM
Kp
B)
Tree scale: 0.1
E. fergusonii str. RHB38-C04
S. boydii str. 183
S. boydii str. 192
E. coli str. M160133
E. coli str. ampC 0069
E. coli TUM18780
E. coli str. K-12 substr. MG1655
E. coli str. BE104
E. coli str. ECOL-18-VL-LA-PA-Ryan-0026
E. coli str. SQ2203
E. coli str. SCU-118
C. dublinensis subsp. dublinensis LMG 23823
C. turicensis z3032
C. sakazakii str. NCTC 8155
C. sakazakii str. 5563 17
C. freundii str. RHBSTW-00444
K. oxytoca str. AR 0028
K. michiganensis str. K518
S. enterica subsp. enterica serovar Typhimurium str. 14028s
Klebsiella sp. STW0522-44
Citrobacter sp. RHBSTW-00229
C. freundii complex sp. CFNIH3
Citrobacter sp. RHBSTW-00021
Citrobacter sp. RHBSTW-00976
Citrobacter sp. RHBSTW-00424
K. pneumoniae str. Bckp206
K. pneumoniae subsp. pneumoniae str. TGH10
K. pneumoniae str. CAV1016
K. pneumoniae str. RHB26-C08
K. pneumoniae str. NMI4661 17
K. pneumoniae str. 825795-1
K. pneumoniae str. E17KP0052
K. pneumoniae str. Kp Goe 149473
K. pneumoniae str. Kp46596
K. pneumoniae str. WCHKP34
K. pneumoniae str. KpvST147B SE1 1 NDM
S. enterica subsp. enterica serovar Typhi str. CT18
E. coli O145:H28 122715
E. coli YSP8-1

Figure 3.9 The phylogeny of A) *cas3* and B) *cas1-cas2* across *Enterobacteriaceae*. The CRISPR-leader sequences were aligned using MUSCLE, and the phylogenetic tree was constructed using ML.

similarity between strains justifies their coherence and segregation within the *cas3* phenogram. The *cas3* sequences of *S. enterica* serovar Typhi and Typhimurium sequences have poor (9.36%) sequence identity; even though they belong to the same species, they are segregated into separate clades. Similar is the case for *E. coli* YSP8-1 and *E. coli* str. SQ2203. Instead, the *cas3* sequences of *E. coli* YSP8-1 and *S. enterica* subsp. *enterica* serovar Typhi str. CT18 shows 55.10% identity, justifying their clubbing and segregation of *E. coli* YSP8-1 from other strains of the clade ES_h. The *cas3* gene is also singled out from the other *cas* genes for these two strains and is present on the complementary strand (**Fig. 3.6**).

3.3.4.2 Phylogeny of *cas1-cas2*

We constructed a phenogram for the *cas1* and *cas2* genes involved in DNA recognition and spacer acquisition (Xue & Sashital, 2019). The phenogram suggests that these genes are highly conserved within the strains of the same species except for *Salmonella* (**Fig. 3.9B**). The phylogenetic tree has two distinct clades with *Escherichia*, *Shigella* and *S. enterica* subsp. *enterica* serovar Typhi str. CT18, while the other clade consists of *Klebsiella, Citrobacter, Cronobacter,* and *S. enterica* serovar Typhimurium. The topology of the phenogram correlates with their *cas1-cas2* nucleotide sequence similarity. The sequence similarity was >90% within the strains of each clade. In the phenogram, *K. michiganenins* str. K518 and *K. oxytoca* str. AR0028 are closer to the *Citrobacter* strains, while *Salmonella* serovars split into two clades. The *cas1-cas2* genes of *S. enterica* subsp. *enterica* serovar Typhi str. CT18 and *S. enterica* subsp. *enterica* serovar Typhimurium str. 14028s have 74.71% identity. While *cas1-cas2* genes of *S. enterica* serovar Typhi str. CT18 has 77.29% identity with *E. coli* str. SQ2203 and *S. enterica* serovar Typhimurium str. 14028s have 82.32% identity with *Klebsiella* sp. STW0522-44.

3.3.4.3 Phylogeny and characterisation of the CRISPR loci

Phylogeny of consensus DR and last DR

The CRISPR loci are defined by their DR sequences generally conserved within species (Horvath *et al.*, 2008). The CRISPR1-DR sequences for strains in our database are 29 bp long. The phylogenetic tree of the consensus DR sequences was constructed based on a pairwise distance matrix (**Fig. 3.12A**). Strain *C. sakazakii* str. NCTC 8155 emerged as an outgroup for the tree and has ~79-83% sequence identity with other *Cronobacter* strains. The other strains diverged into two major groups. One group had *Shigella* and *Escherichia* strains like that observed for the clade ES_h in the *cas3* and CRISPR1 phenogram. Nevertheless, *Shigella* strains and *E. coli* YSP8-1 formed a separate sub-clade within the clade ES_h. *E. coli* YSP8-1 was closer to the *Shigella* strains, with their CRISPR1-DR having 93% sequence identity. The second group contained *Salmonella, Klebsiella, Citrobacter*, and *Cronobacter* species. Even though there was no similar spacer between *S. enterica* serovar Typhi and Typhimurium, the CRISPR1-DRs are highly conserved. CRISPR1-DR of these serovars are 93% identical. The consensus CRISPR1-DR sequence 5' GTGTTCCCCGCGCCAGCGGGGATAAACCG 3' was mostly conserved across the analysed species except for *Cronobacter.*

The CRISPR DRs within each array are generally conserved, but the last DR (Horvath *et al*., 2008) is heterogeneous. Thus, we also performed a phylogenetic analysis of the terminal DR in *Enterobacteriaceae* species. The terminal DR phenogram is discrepant with that obtained for other CRISPR-Cas components. We did not observe a species-specific distribution of strains in the phenogram (**Fig. 3.12B**). For example, *E. fergusonii* clubbed with *Citrobacter; S. enterica* subsp. *enterica* serovar Typhi str. CT18 with *Klebsiella* sp. STW0522; and *E. coli* TUM18780, 0145:H28 with other *Klebsiella* strains. *C. dublinesis* subsp. *dublinesis* LMG 23823 formed an outgroup for this last DR tree. Similar to earlier reports, the terminal DR is truncated, and degeneracy was observed near the 3' end.

Evolutionary study of CRISPR1 leader

More than 600 *Escherichia, Shigella*, and *Klebsiella* strains have the CRISPR/Cas system that matches with CRISPR1-STM/*cas*-STM. Nevertheless, a few strains of the *Enterobacteriaceae* family (*Klebsiella* and *Citrobacter*) contain CRISPR1-STM and CRISPR1-STY array and *cas* operon. As some strains have more than one CRISPR array, we defined the CRISPR array proximal to the *cas* operon as CRISPR1 for consistency.

The phylogeny of the CRISPR1 leader sequence showed two significant clades (**Fig. 3.10A**). The first clade consisted of strains from clade ES_h, *E. coli* YSP8-1 and *Salmonella* strains, while the second clade comprised the remaining strains. *C. dublinesis* subsp. *dublinesis* LMG 23823 separated in the second clade having significantly less sequence similarity with *Citrobacter* sp. RHBSTW-00229 and 72.5% sequence similarity with *C.*

sakazakii str. NCTC 815. The clade ES_h and sub-clade K_P had compositions similar to the *cas3* phenogram. In contrast to the *cas3* phylogeny, *S. enterica* subsp. *enterica* serovar Typhimurium str. 14028s clubbed with the clade ES_h as their CRISPR1 leaders show a similarity of 69.5%. *E. coli* YSP8-1 grouped with the other *E. coli* strains as the CRISPR1 leader sequences of all the *E. coli* strains are ~100% identical.

Spacer conservation of CRISPR1 array

We also inspected the heterogeneity of the CRISPR1 array across the shortlisted strains by analysing its spacer conservation within and across species (query cover >90% and percent similarity >90%).

Intra-species spacer conservation: 932 spacers were detected across the 39 strains, with 606 unique spacers. The conservation of the spacer sequences and their arrangement was variable for the species analysed in this work (**Fig. 3.11**). The array was conserved in both *Shigella boydii* strains. A significant number of spacers are conserved within most strains of *Citrobacter, E. coli,* and *K. pneumonia,* while some spacer deletions/additions are observed in some strains of these species. Many strains of *K. pneumonia* have conservation of leader-distal spacers but not the proximal ones. *C. freundii* str. RHBSTW-00444 and *Citrobacter* sp. RHBSTW-00229 have unique sets of spacers; nevertheless, one spacer (spacer 22) of *C. freundii* str. RHBSTW-00444 is identical to that (spacer 29) of *Citrobacter* sp. RHBSTW-00424 and *C. freundii* complex sp. CFNIH3. Spacer duplication is observed in *Citrobacter* sp. RHBSTW-00229 (spacer 12 and 13) and *Klebsiella* (str. CAV1016- spacer 31 and 59 and str. RHB26-C08- spacer 17 and 26; 18 and 27; 19 and 28; 20 and 29) (**Fig. 3.11**). No strains of *Cronobacter* and *Salmonella* showed intra-species spacer conservation among the strains studied. The unique set of spacers was observed in some strains of *Escherichia* and *Klebsiella,* like *E. coli* O145:H28 122715DNA, *E. fergusonii* strain RHB38-C04, *K. pneumoniae* str. Bckp206, *K. oxytoca* str. AR_0028 etc. However, the first spacer of *E. coli* O145:H28 122715DNA is identical to that of *E. coli* TUM187180.

Inter-species spacer conservation: No inter-species spacer conservation was observed except between *Escherichia* and *Shigella.* Spacers E and F of *Shigella boydii* matched with spacers L and M of *E. coli* YSP8-1, respectively. This indicates that the CRISPR spacers are species-specific and suggest different protospacer sources for the studied species.

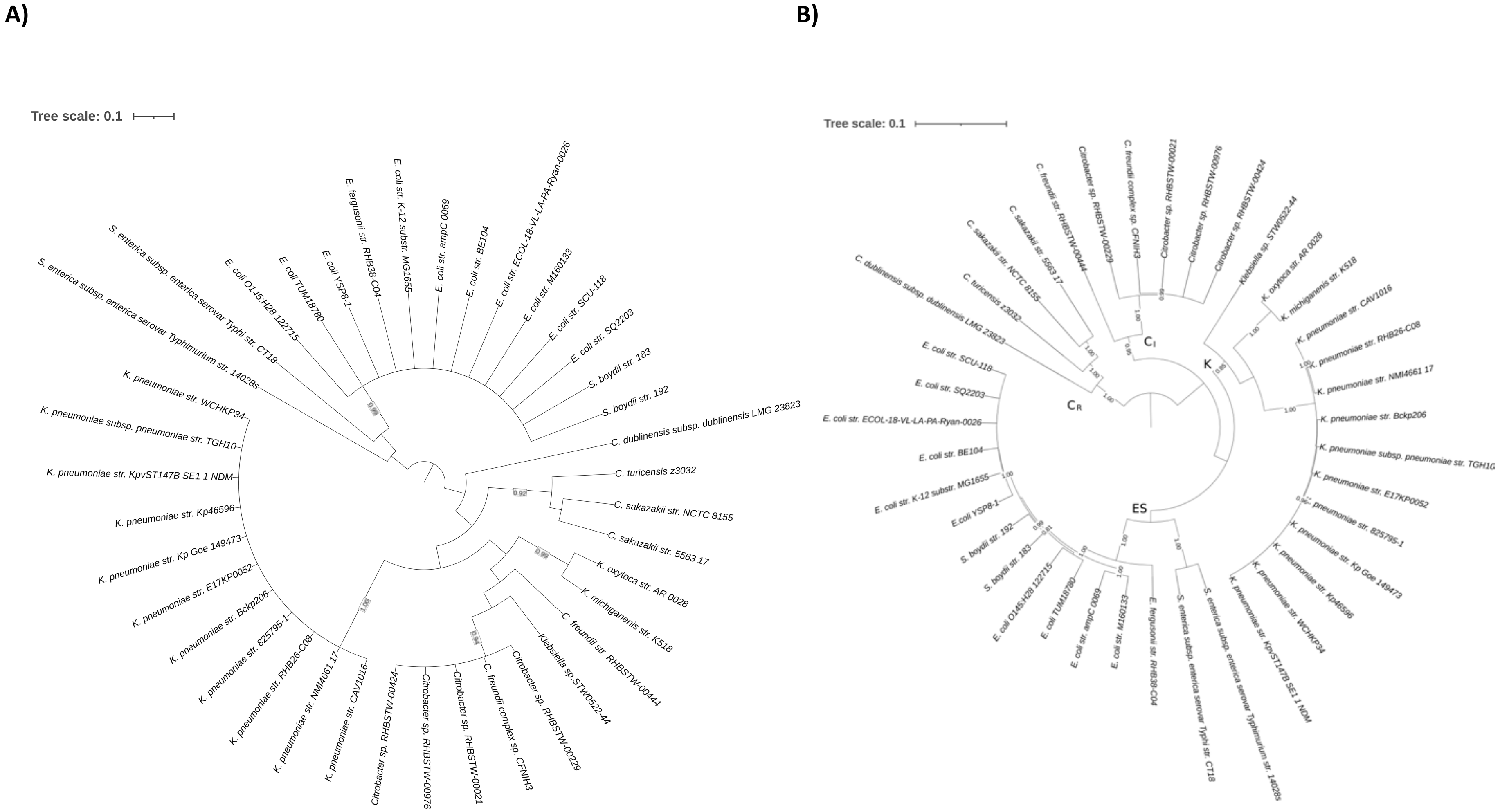

A)
B)
Tree scale: 0.1
S. enterica subsp. enterica serovar Typhimurium str. 14028s
S. enterica subsp. enterica serovar Typhi str. CT18
E. coli O145:H28 122715
E. coli TUM18780
E. coli YSP8-1
E. fergusonii str. RHB38-C04
E. coli str. K-12 substr. MG1655
E. coli str. ampC 0069
E. coli str. BE104
E. coli str. ECOL-18-VL-LA-PA-Ryan-0026
E. coli str. M160133
E. coli str. SCU-118
E. coli str. SQ2203
S. boydii str. 183
S. boydii str. 192
C. dublinensis subsp. dublinensis LMG 23823
C. turicensis z3032
C. sakazakii str. NCTC 8155
C. sakazakii str. 5563 17
K. oxytoca str. AR 0028
K. michiganenis str. K518
C. freundii str. RHBSTW-00444
Klebsiella sp.STW0522-44
Citrobacter sp. RHBSTW-00229
C. freundii complex sp. CFNIH3
Citrobacter sp. RHBSTW-00021
Citrobacter sp. RHBSTW-00976
Citrobacter sp. RHBSTW-00424
K. pneumoniae str. CAV1016
K. pneumoniae str. NMI4661 17
K. pneumoniae str. RHB26-C08
K. pneumoniae str. 825795-1
K. pneumoniae str. Bckp206
K. pneumoniae str. E17KP0052
K. pneumoniae str. Kp Goe 149473
K. pneumoniae str. Kp46596
K. pneumoniae str. KpvST147B SE1 1 NDM
K. pneumoniae subsp. pneumoniae str. TGH10
K. pneumoniae str. WCHKP34
0.99
0.92
0.99
1.00
0.94
Tree scale: 0.1
C. dublinensis subsp. dublinensis LMG 23823
C. turicensis z3032
C. sakazakii str. NCTC 8155
C. sakazakii str. 5563 17
C. freundii str. RHBSTW-00444
Citrobacter sp. RHBSTW-00229
C. freundii complex sp. CFNIH3
Citrobacter sp. RHBSTW-00021
Citrobacter sp. RHBSTW-00976
Citrobacter sp. RHBSTW-00424
Klebsiella sp. STW0522-44
K. oxytoca str. AR 0028
K. michiganenis str. K518
K. pneumoniae str. CAV1016
K. pneumoniae str. RHB26-C08
K. pneumoniae str. NMI4661 17
K. pneumoniae str. Bckp206
K. pneumoniae subsp. pneumoniae str. TGH10
K. pneumoniae str. E17KP0052
K. pneumoniae str. 825795-1
K. pneumoniae str. Kp Goe 149473
K. pneumoniae str. Kp46596
K. pneumoniae str. WCHKP34
K. pneumoniae str. KpvST147B SE1 1 NDM
S. enterica subsp. enterica serovar Typhimurium str. 14028s
S. enterica subsp. enterica serovar Typhi str. CT18
E. fergusonii str. RHB38-C04
E. coli str. M160133
E. coli str. ampC 0069
E. coli TUM18780
E. coli O145:H28 122715
S. boydii str. 183
S. boydii str. 192
E.coli YSP8-1
E. coli str. K-12 substr. MG1655
E. coli str. BE104
E. coli str. ECOL-18-VL-LA-PA-Ryan-0026
E. coli str. SQ2203
E. coli str. SCU-118
C_I
K
C_R
ES
1.00
0.95
0.86
1.00
1.00
1.00
1.00
1.00
1.00
1.00
1.00
1.00
1.00
1.00

Figure 3.10 The Phylogeny of the A) CRISPR1-leader sequence B) *gyrB* gene of species of *Enterobacteriaceae* family. The CRISPR1-leader sequences were aligned using MUSCLE, and the phylogenetic tree was constructed by ML. The bootstrap values are indicated at each node.

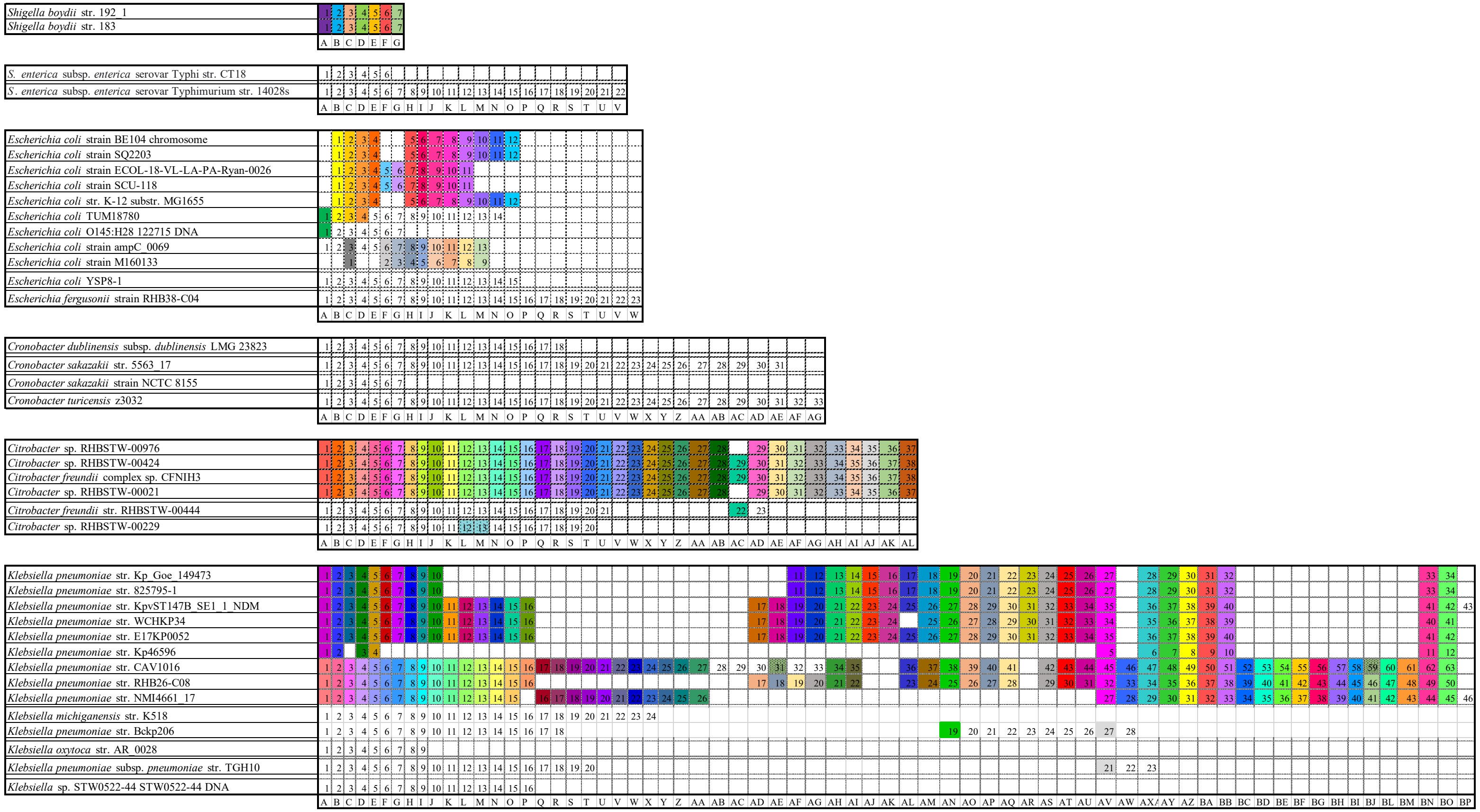

Shigella boydii str. 192_1
Shigella boydii str. 183
S. enterica subsp. enterica serovar Typhi str. CT18
S. enterica subsp. enterica serovar Typhimurium str. 14028s
Escherichia coli strain BE104 chromosome
Escherichia coli strain SQ2203
Escherichia coli strain ECOL-18-VL-LA-PA-Ryan-0026
Escherichia coli strain SCU-118
Escherichia coli str. K-12 substr. MG1655
Escherichia coli TUM18780
Escherichia coli O145:H28 122715 DNA
Escherichia coli strain ampC_0069
Escherichia coli strain M160133
Escherichia coli YSP8-1
Escherichia fergusonii strain RHB38-C04
Cronobacter dublinensis subsp. dublinensis LMG 23823
Cronobacter sakazakii str. 5563_17
Cronobacter sakazakii strain NCTC 8155
Cronobacter turicensis z3032
Citrobacter sp. RHBSTW-00976
Citrobacter sp. RHBSTW-00424
Citrobacter freundii complex sp. CFNIH3
Citrobacter sp. RHBSTW-00021
Citrobacter freundii str. RHBSTW-00444
Citrobacter sp. RHBSTW-00229
Klebsiella pneumoniae str. Kp_Goe_149473
Klebsiella pneumoniae str. 825795-1
Klebsiella pneumoniae str. KpvST147B_SE1_1_NDM
Klebsiella pneumoniae str. WCHKP34
Klebsiella pneumoniae str. E17KP0052
Klebsiella pneumoniae str. Kp46596
Klebsiella pneumoniae str. CAV1016
Klebsiella pneumoniae str. RHB26-C08
Klebsiella pneumoniae str. NMI4661_17
Klebsiella michiganensis str. K518
Klebsiella pneumoniae str. Bckp206
Klebsiella oxytoca str. AR_0028
Klebsiella pneumoniae subsp. pneumoniae str. TGH10
Klebsiella sp. STW0522-44 STW0522-44 DNA

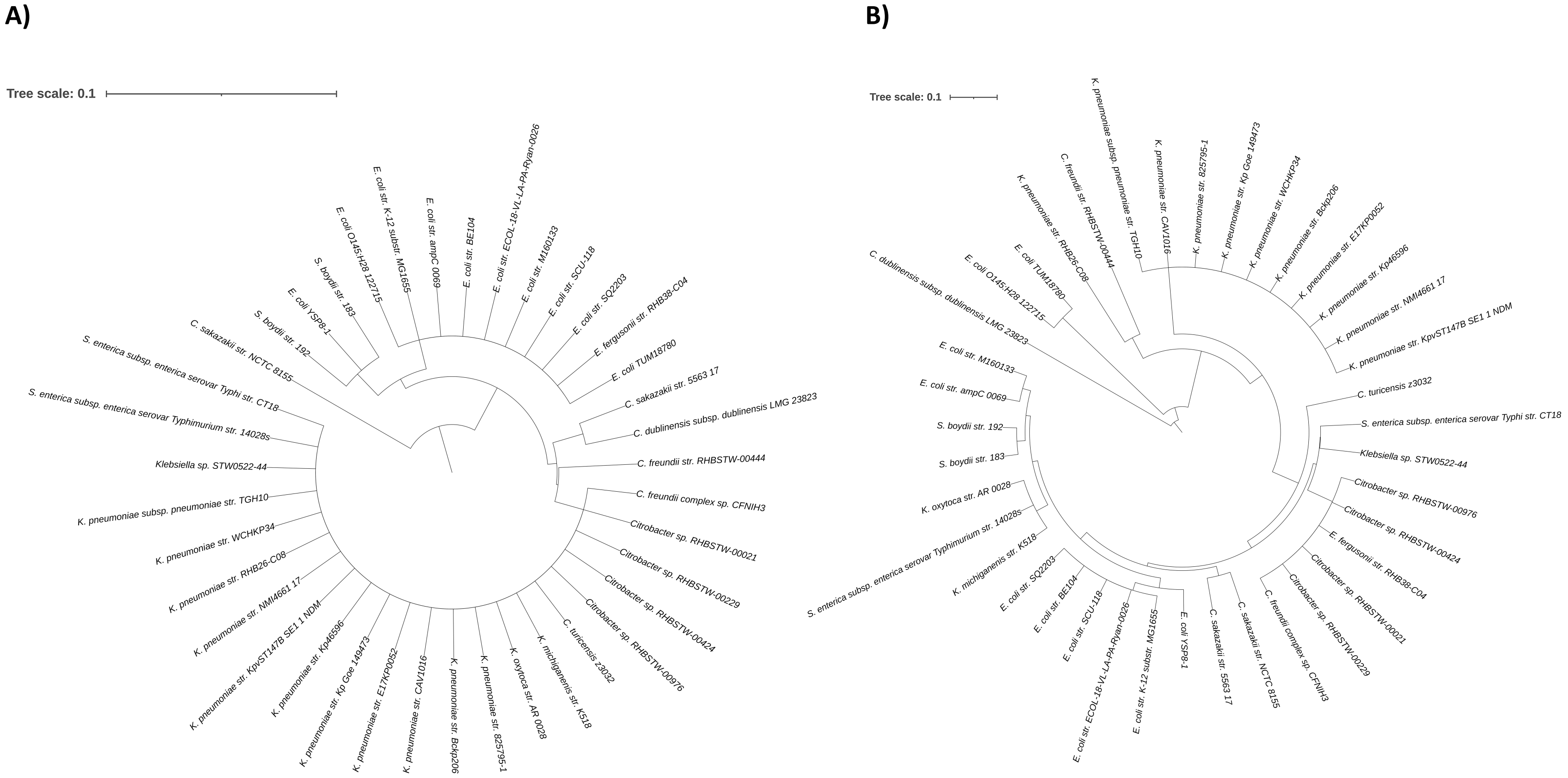
Figure 3.11 Spacer conservation across *Enterobacteriaceae* species. The diagram represents spacer maps for *Shigella, Citrobacter, Cronobacter, Escherichia, Salmonella* and *Klebsiella*. The colour code for a particular column represents spacer sequences with greater than 90% nucleotide similarity. The white colour code for the same columns shows no similarity. The number denotes the position of the spacer from the leader sequence. The DR have been eliminated for simplicity. The duplication (*Citrobacter* sp. RHBSTW-00229 -Spacer 12 and 13 and *Klebsiella* str. CAV1016- spacer 31 and 59 and str. RHB26-C08- spacer 17 and 26; 18 and 27; 19 and 28; 20 and 29) is depicted as a pattern in the coloured box.

3.3.4.4 Phylogeny of *gyrB*

The phylogenetic impression of the species' taxonomy can be deciphered by inspecting the phylogeny of their housekeeping genes. Thus, we studied the evolutionary history of selected *Enterobacteriaceae* species using the DNA gyrase B (*gyrB*) gene (Fukushima, Kakinuma, & Kawaguchi, 2002). The *gyrB* tree had two major clades containing (i) *Cronobacter* strains (clade C_R) and (ii) *Salmonella, Escherichia*, *Klebsiella*, *Shigella,* and *Citrobacter* strains (**Fig. 3.10B**). The second clade further contained subclades (i) *Citrobacter* strains (clade C_I), (ii) Klebsiella strains (clade K), and (iii) *Escherichia, Shigella* and *Salmonella* strains (clade ES). *Escherichia* and *Shigella* strains formed a consolidated group, indicating the recent emergence of the *Shigella* strains (Fukushima *et al*., 2002). Overall, the phenogram showed species-specific clustering. The *gyrB* tree shows incongruent relationships with the CRISPR-Cas trees, indicating the complex evolutionary history of the CRISPR-Cas system, including convergent evolution and HGT.

3.3.5 CRISPR-Cas system in *Salmonella* genome is flanked with MGE

To decipher the probable involvement of HGT, we screened the presence of the signature MGE, namely, helicase, transposase, and integrase (Deng *et al*., 2019; McDonald *et al*., 2019) in the proximity of the CRISPR-Cas region of *Salmonella*. To this end, we also analysed the GC content of this region in comparison to the whole genome. We found that 18 out of 20 serovars (with representative strains of each considered) showed truncated/probable transposase at a position 30 kb upstream of the CRISPR1 loci (**Table 3.2**). The transposable elements are not uniformly found within ±30 kb of any region in the genome (**Fig. 3.13**), suggesting CRISPR could have been possibly acquired *via* transposition. The GC content of the CRISPR arrays for most of the serovars was higher than the GC content of the whole genome except for a few serovars with smaller arrays with lower GC content due to the AT-rich leader sequence (**Table 3.2**). A transposase gene was also upstream of the CRISPR2 array in serovars Paratyphi A and Typhi. Moreover, a helicase gene was downstream of the CRISPR2 array in the serovars Typhi and Typhimurium.



A)
Tree scale: 0.1
E. coli str. K-12 substr. MG1655
E. coli str. ampC 0069
E. coli str. BE104
E. coli str. ECOL-18-VL-LA-PA-Ryan-0026
E. coli str. M160133
E. coli str. SCU-118
E. coli str. SQ2203
E. fergusonii str. RHB38-C04
E. coli TUM18780
C. sakazakii str. 5563 17
C. dublinensis subsp. dublinensis LMG 23823
C. freundii str. RHBSTW-00444
C. freundii complex sp. CFNIH3
Citrobacter sp. RHBSTW-00021
Citrobacter sp. RHBSTW-00229
Citrobacter sp. RHBSTW-00424
Citrobacter sp. RHBSTW-00976
C. turicensis z3032
K. michiganensis str. K518
K. oxytoca str. AR 0028
K. pneumoniae str. 825795-1
K. pneumoniae str. Bckp206
K. pneumoniae str. CAV1016
K. pneumoniae str. E17KP0052
K. pneumoniae str. Kp Goe 149473
K. pneumoniae str. Kp46596
K. pneumoniae str. KpvST147B SE1 1 NDM
K. pneumoniae str. NMI4661 17
K. pneumoniae str. RHB26-C08
K. pneumoniae str. WCHKP34
K. pneumoniae subsp. pneumoniae str. TGH10
Klebsiella sp. STW0522-44
S. enterica subsp. enterica serovar Typhimurium str. 14028s
S. enterica subsp. enterica serovar Typhi str. CT18
C. sakazakii str. NCTC 8155
S. boydii str. 192
E. coli YSP8-1
S. boydii str. 183
E. coli O145:H28 122715
B)
Tree scale: 0.1
K. pneumoniae subsp. pneumoniae str. TGH10
K. pneumoniae str. CAV1016
K. pneumoniae str. 825795-1
K. pneumoniae str. Kp Goe 149473
K. pneumoniae str. WCHKP34
K. pneumoniae str. Bckp206
K. pneumoniae str. E17KP0052
K. pneumoniae str. Kp46596
K. pneumoniae str. NMI4661 17
K. pneumoniae str. KpvST147B SE1 1 NDM
C. turicensis z3032
S. enterica subsp. enterica serovar Typhi str. CT18
Klebsiella sp. STW0522-44
Citrobacter sp. RHBSTW-00976
Citrobacter sp. RHBSTW-00424
E. fergusonii str. RHB38-C04
Citrobacter sp. RHBSTW-00021
Citrobacter sp. RHBSTW-00229
C. freundii complex sp. CFNIH3
C. sakazakii str. NCTC 8155
C. sakazakii str. 5563 17
E. coli YSP8-1
E. coli str. K-12 substr. MG1655
E. coli str. ECOL-18-VL-LA-PA-Ryan-0026
E. coli str. SCU-118
E. coli str. BE104
E. coli str. SQ2203
K. michiganensis str. K518
S. enterica subsp. enterica serovar Typhimurium str. 14028s
K. oxytoca str. AR 0028
S. boydii str. 183
S. boydii str. 192
E. coli str. ampC 0069
E. coli str. M160133
C. dublinensis subsp. dublinensis LMG 23823
E. coli O145:H28 122715
E. coli TUM18780
K. pneumoniae str. RHB26-C08
C. freundii str. RHBSTW-00444

Figure 3.12 The phylogeny of A) DR sequences and B) the last DR sequence across *Enterobacteriaceae*. The neighbour-joining tree was constructed based on distance matrix analysis of the consensus DR sequences.

Table 3.2 MGE candidates flanking the CRISPR-Cas system

	Genome location (Loci start and Loci end)			MGE	Percentage GC Content				
	CRISPR2	*cas*	CRISPR1	Transposase/ Helicase	CRISPR2 loci	*cas* operon	CRISPR1 loci	CRISPR-Cas	Whole genome
Paratyphi A str. AKU_12601	2902111-2902322	2885645-2894598	2885105-2885560	2856664-2857091 and 3007579-3008037	51.3*	49.9	53.2*	49	52.18
Newport II str. SL254	3073142-3074328	3056558-3064452	3054859-3056473	3024723-3025160	54.2	50.4	55.6	51	52.22
Newport III str. USDA-ARS-USMARC-1927	975001-975639	983339-991789	991886-993012	1020178-1020414	51.3	52	57.2	51	52.18
Heidelberg str. SL476	3069137-3070263	3052976-3061435	3051217-3052879	3022734-3022874	56.3	53.1	56.2	53	52.07
Enteritidis str. EC20121175	2967508-2968269	2951453-2959906	2950779-2951356	2923727-2923816	55.8	52.8	55.7	51	52.17
Gallinarum str. 9184	1224776-1225415	1233429-1241471	1241469-1241716	1271625-1271848	56.8	53.3	47.6*	51	52.2
Pullorum str. ATCC 9120	3871330-3871722	3855356-3863734	3855036-3855283	3827981-3828070	53.6*	52.9	48.4*	51	52.19
Gallinarum/Pullorum str. CDC1983-67	2947545-2947937	2931570-2939948	2931250-2931497		53.6*	52.9	48.4*	51	52.23
Montevideo-STY str. USDA-ARS-USMARC-1900	1003049-1004420	1011949-1021097	1021182-1023406	1050505-1050672	57.7	50.4	57	52	52.35
Montevideo-STM str. CDC 2010K-0257	992948-994014	1010602-1010052	1010149-1011641	1038743-1038910	56.6	52.4	55.9	51	52.21
Bovismorbificans str. 3114	2976895-2977839	2960410-2969354	2958839-2960331		54	50	55.6	50	52.16
Anatum str. CDC 06-0532	970192-971440	979153-987612	987709-988225	1018009-1018429	56.2	52.8	55.5	51	52.18
Tennessee str. ATCC 10722	963889-965320	972914-981858	981943-985815	1015542-1015960	59.3	50.4	57.1	52	52.23
Saintpaul-STM str. CFSAN004173	946615-947863	955563-964025	964122-965066	993235-993458	57.2	52	57.4	51	52.21
Saintpaul-STY str. SARA26	944744-946114	953779-962723	962808-965337	993600-993740	58.2	50.3	56.5	51	52.05
Dublin str. CT_02021853	3137409-3137742	3121348-3129807	3121100-3121350	3090967-3091107	51.6*	53	44.6*	51	52.18
Agona str. SL483	3005517-3006033	2989328-2997778	2988105-2989231	2956649-2956789 and 2975984-2977192	54.2	52	56.3	51	52.08
Typhimurium str. CFSAN001921	473159-474652	482228-490687	490784-492564	522104-522291 and 423309-425144	55.9	50.4	56.9	51	52.17
Typhi str. CT18	2943208-2943716	2926652-2935104	2926182-2926567	2898592-2899034 and 3013615-3014073	39.8*	50.4	57.3	50	52.05
Subsp. *arizonae*		962736-971156	971253-972682			52.9	57.4	52	51.38
Subsp. *diarizonae*		1126557-1134989	1135086-1136883			52.9	56.7	53	51.54
S. *bongori*	922139-923204	930621-939089	939186-940434		53.4	52.4	54.5	50	51.33

*The lower GC content of CRISPR arrays due to the AT-rich leader sequences are represented by asterisks

3.3.6 Mapping protospacer sources of CRISPR spacers

We mapped the protospacer sources (plasmids, phages, and viruses) using the CRISPRTarget tool (Biswas *et al.*, 2013) and compared them across serovars (**Fig. 3.14**). We identified protospacers for CRISPR1 in a range of 44% (Agona and Enteritidis) to 100% (Gallinarum/Pullorum and Sendai). Similarly, for CRISPR2, the percentage of protospacers range from 40% (Dublin) to 100% (Typhi and Sendai). Common protospacer sources were observed, majorly for the serovars sharing spacers with each other. For example, serovars Heidelberg and Typhimurium shared sufficiently high protospacer sources compared to other serovar pairs. Thus, even though the serovars inhabit/infect similar habitats/hosts, e.g., serovar Enteritidis and Typhimurium, they differ in their protospacer sources. Protospacers were not traced for a substantial proportion of CRISPR1 (~36% ±14.8) and CRISPR2 (~36% ±15.6) spacers. No correlation was observed between the number of spacers and protospacers, especially for arrays with high spacer content.

Further, we mapped the protospacer sources for the *Enterobacteriaceae* species to explore commonalities across species and trace their evolutionary pathway. For *Citrobacter* and *Klebsiella,* 17 and 75 protospacers mapped to genomes of the Myophage, Siphophage, and Podophage groups of bacteriophages, respectively. Spacers of *Cronobacter* species mapped to genomes of Siphophage bacteriophage and *Salmonella* phages. Six spacers for *Escherichia* matched their usual MGE protospacer targets including phages infecting the *Enterobacteriaceae* family. One spacer of *E. fergusonii* targets the Stx1a-converting phage that codes Shiga toxin 1 protein.

3.4 Discussion

The evolutionary mechanisms in bacteria are highly complex, with environmental factors intricately modulating the genome architecture and functionality. Further, HGT and recombination events significantly influence the evolutionary framework of the bacteria. Our study probes into the evolution of *Salmonella* with respect to the CRISPR-Cas system that influences the genome evolution (Nguyen *et al.*, 2018) and bacterial virulence (Cui *et al.*, 2020). We categorised the CRISPR-Cas system into two types, namely, CRISPR1-STM/*cas*-STM and CRISPR1-STY/*cas*-STY, based on the phylogenetic segregation and differences in the CRISPR1-leader and *cas* genes features of the strains studied.

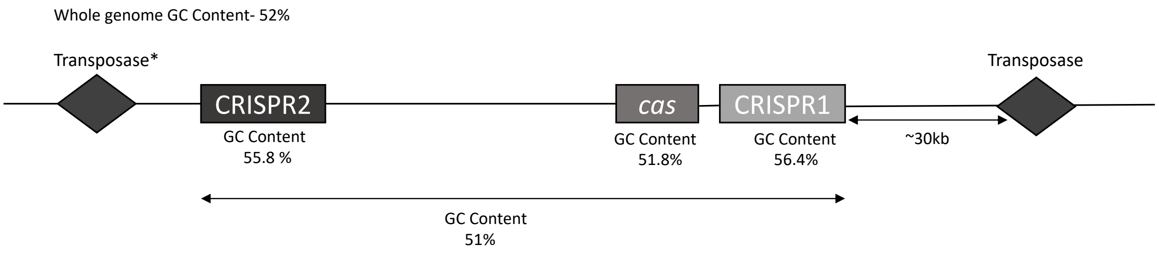


Figure 3.13 A generalised representation of the signature genes involved in horizontal gene transfer. All *Salmonella* serovars except serovars Bovismorbificans and Gallinarum/Pullorum contain the transposase gene upstream of CRISPR1 loci.

* - transposase upstream of CRISPR2 is present only in serovars Typhi and Typhimurium.

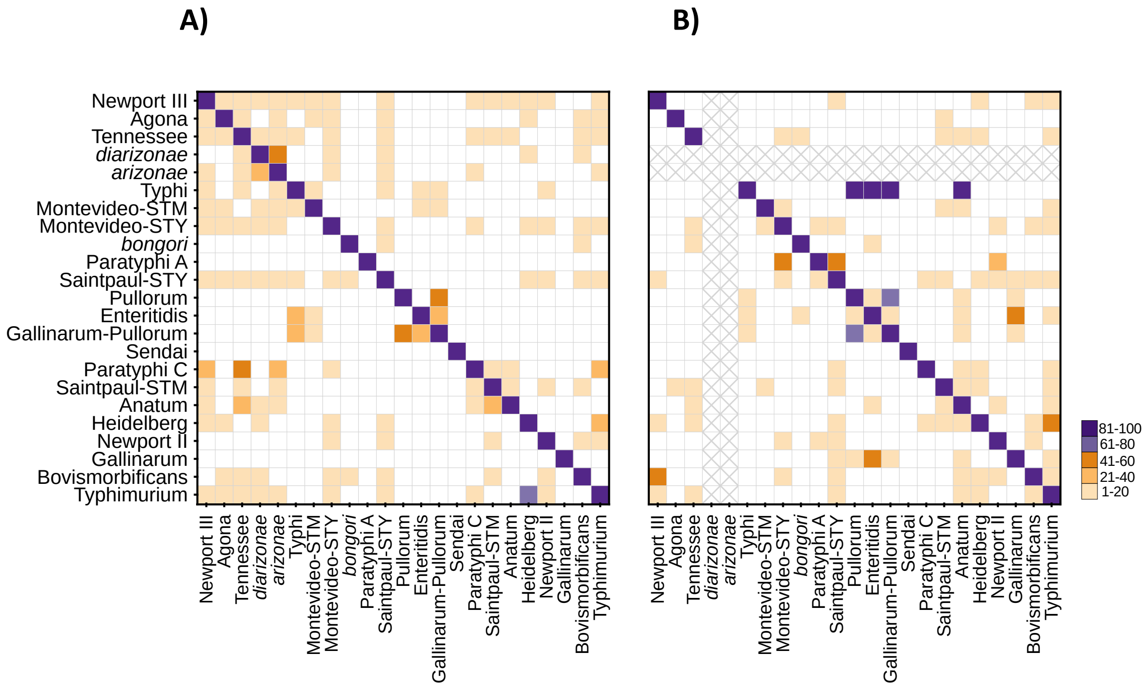


Figure 3.14 Heat map for sharing of protospacer source by pairs of serovars for spacers belonging to A) CRISPR1 array and B) CRISPR2 array.

The CRISPR-Cas evolution is portrayed as complex, having modular character hindering its forthright categorisation based on the serovar host range and geographical location. The serovars, Newport-II and Newport-III, infect primates, reptiles, and Aves (Ferrari *et al*., 2019) but are still segregated into two separate clades in the CRISPR1-leader tree. Serovar Typhimurium strain SARA13 and Saintpaul SARA26 were isolated from the same geographic location, France (GenBank database), but segregated into CRISPR1-STM and CRISPR1-STY clades, respectively. The conservation of the array within strains of all the serovars, irrespective of the geographic location, suggests CRISPR acquisition to be a primaeval event.

The chronicles of battles between the bacteria and the invading MGE are registered as spacers in the CRISPR arrays. The spacer conservation was weak across the serovars but significant within themselves except for those of serovars Montevideo, Newport, and Saintpaul. However, spacer variability was observed within a few serovars like Typhimurium and Newport-III, showing some variations in their CRISPR1-spacer composition. Thus, the acquisition of the spacers could be a primitive event, with different selection pressures operating on different serovars to maintain the spacer composition. One elucidation is the spacer composition of the system could potentially leverage protection against invading MGE (Nguyen *et al*., 2018) or pathogenic potential, possibly through endogenous gene regulation (R. Li *et al*., 2016; Bozic, Repac, & Djordjevic, 2019; Cui *et al*., 2020) as implicated elsewhere (R. Li *et al*., 2016; Nguyen *et al*., 2018; Bozic *et al*., 2019; Cui *et al*., 2020), thereby resulting in the spacers preservation. This polymorphism of spacers across serotypes finds utility in serotyping (Fabre *et al*., 2012; Thompson *et al*., 2018).

The CRISPR1- and CRISPR2- spacer trees were distinct from each other. However, some serovars (clade-HNT, clade-PS, and clade-DEGP) were consistently grouped in all the CRISPR and *cas* trees implying a highly conserved CRISPR-Cas system within the serovar-group. For example, serovar Heidelberg has 66% of CRISPR1- and 100% of CRISPR2-spacers identical to the serovar Typhimurium. This may indicate a recent divergence of these serovars in the evolutionary timeline of *Salmonella*. Notably, some serovars like Bovismorbificans, Anatum, Saintpaul, Montevideo, and Typhi grouped differently in CRISPR-leader and -spacer phenograms. This indicates random spacer acquisition/loss or multiple HGT events in these serovars. Further, spacer tree analyses suggest that the grouping and segregation of the serovars are independent of host-specificity and their habitat. For example, a primate-specific serovar Typhi clubbed with bird/cattle-specific serovars. Moreover, the serovars with similar host ranges or habitats largely have non-overlapping protospacer sources (comprising MGE).

The anchor spacer gives an indirect correlation of the last common ancestor (LCA) for the array and is generally conserved for a particular serovar (Shariat *et al.*, 2015). Many serovars of the clades in the spacer tree share the anchor spacer (**Fig. 3.2 & 3.4**), thereby suggesting an LCA for the array in each clade. However, for some serovars, other spacers, but not the anchor spacer, are shared. For instance, the serovar Gallinarum shares CRISPR1 spacers with Enteritidis but not the anchor spacer, implicating the loss of some common spacers, including the anchor spacer. Serovar Bovismorbificans share five CRISPR1 spacers with serovar Saintpaul-STM and anchor spacer with serovar Newport-II, indicating divergence from Newport-II and recombination with Saintpaul-STM.

The *cas* genes of the strains in the *cas*-STM and *cas*-STY categories are highly similar within each category but differ from the other, except for the *cas1* and *cas2* genes required for the spacer acquisition (Nuñez *et al.*, 2014). However, the key residues of Cse1 and the functional domains of Cas3 are conserved, indicating the conservation of their functionality. The strains comprising *cas*-STM and *cas*-STY are identical to CRISPR1-STM and CRISPR1-STY, respectively. This is empirical, as the CRISPR1 array and the *cas* operon are juxtaposed. Furthermore, the CRISPR1-STY/*cas*-STY category strains showed higher substitutions per sequence site, implying the plasticity for new alterations.

The size of the spacer set for a given serovar is proportional to its host range (**Fig. 3.1**). Ubiquitous serovars like Typhimurium, Newport-II, Tennessee, and Heidelberg have huge spacer sets, while host-specific/adapted serovars like Typhi, Sendai, Gallinarum, Dublin possesses a few spacers. Considering the role of spacers in regulating endogenous genes (Wimmer & Beisel, 2019) and preventing invading MGE (Nguyen *et al.*, 2018), we put forward two possible hypotheses. The spacer versatility in broad-host-range serovars can be due to exposure to a wide range of environments, and/or it permits the regulation of different genes. In both cases, the bacteria possibly gain the advantage of adapting to multiple stress factors like attack by MGE and hostile host conditions. All the spacers of the host-specific serovars Gallinarum, Pullorum, and Gallinarum/Pullorum are present in serovar Enteritidis (a broad-host-range serovar) along with some additional spacers further testifying the hypotheses. The protospacers (MGE) sources among these serovars are reasonably common (**Fig. 3.14**). Moreover, even though serovar Enteritidis (Suar *et al.*, 2006) is a broad-host-range serovar and shares the habitats (e.g., mammalian gut) with that of serovar Typhimurium (Suar *et al.*, 2006) and Heidelberg (Foley, Johnson, Ricke, Nayak, & Danzeisen, 2013) they hardly have common protospacer source. Further, the binding of Cascade complex along with endogenous crRNA to >100 chromosomal targets in *E. coli* (Cooper, Stringer, & Wade, 2018) and *S. enterica* subsp. *enterica* serovar Typhimurium indicates the regulation of gene expression by the CRISPR-Cas system. The results of Cui *et al.*, further support endogenous gene regulation, showing modulation of virulence and biofilm genes by CRISPR-Cas system.

Among the host-specific/adapted serovars, the primate-specific serovars, namely Typhi, Paratyphi A, and Sendai, have a CRISPR1-STY/*cas*-STY system. The remaining four serovars are specific to poultry or cattle containing the CRISPR1-STM/*cas*-STM system. We propose that the CRISPR1-STY/*cas*-STY system may provide some advantage to serovars of the CRISPR1-STY clade. This would either prevent MGE invasion or regulate endogenous genes in the primate (a restricted host for typhoidal serovars) gut. Nevertheless, the serovars do not have a common protospacer source, possibly indicating some advantage in endogenous gene regulation. However, in-depth analyses and further research are warranted to understand any advantage of having a CRISPR1-STY/*cas*-STY system in these serovars.

The incongruence in CRISPR and *cas* trees with the MLST tree implies a plausible event of HGT. Similar incongruency with the CRISPR-Cas system of whole genome phylogeny is also reported elsewhere (Timme *et al.*, 2013; Pettengill *et al.*, 2014). A truncated transposase, ~30 kb upstream of the CRISPR1 array and a high GC content of the CRISPR array possibly hint at the occurrence of an HGT event (Daubin, Lerat, & Perrière, 2003; Ravenhall, Škunca, Lassalle, & Dessimoz, 2015). Further support is evidenced through the histone-like nucleoid-structuring protein (H-NS) mediated regulation of *cas* operon in *S. enterica* subsp. *enterica* serovar Typhi (Medina-Aparicio *et al.*, 2011). H-NS is associated with HGT, acting as a transcriptional silencer of horizontally acquired genes by binding to the AT-rich DNA and blocking the RNA polymerase (Ilyas *et al.*, 2017). One may argue the regulation of the CRISPR array by H-NS through its AT-rich leader, as reported for *E. coli* (Pul *et al.*, 2010; Ilyas *et al.*, 2017). Thus, H-NS could have initially silenced the CRISPR-Cas system and later evolved to regulate the functioning of *cas* operon and the CRISPR arrays. However, validation of such mechanism in other strains of *Salmonella* needs further accreditation.

It was found that the leader sequences are generally conserved throughout (with a few SNPs) the strains of the same species. The analysis of leader sequences across species showed that the conserved region usually lies toward the distal end of the CRISPR array. Could this region be a core leader sequence that is critical for the functionality of the CRISPR array? Further studies in this direction are needed to understand this better. Moreover, in *C. freundii* complex sp. CFNIH3, a truncated transposase, was found 30 kb upstream of the CRISPR1 loci. The region between transposase and CRISPR1 shared 40% similarity with that of *S. enterica* subsp. *enterica* serovar Typhimurium, indicating an occurrence of an HGT event. The split of *Salmonella* serovars into two separate clades and clubbing of serovar of CRISPR1-STM with *Shigella* and *E. coli* was also observed in the Cas1 phylogram reported by Touchon *et al.,* 2010, thus conforming to our results.

With the comprehensive analysis of all the results, we put forward the following hypotheses for the evolution of the CRISPR-Cas system in *Salmonella*. Given that a good proportion of *Escherichia, Shigella,* and *Klebsiella* strains harbour CRISPR1-STM/*cas*-STM type leader and operon, we hypothesise that the LCA of the array for *Enterobacteriaceae* family could have been CRISPR1-STM/*cas*-STM type. Moreover, after the divergence from these genera, *Salmonella* could have laterally acquired its CRISPR2 array, as there exists no similarity in their leader sequences. In contrast, leaders of *S. enterica* and *S. bongori* are 78% identical and well-conserved *S. enterica* subsp. *arizonae* and subsp. *diarizonae* do not have a CRISPR2 array, which could have been lost in evolution. Many strains of subsp. *arizonae* do not contain the CRISPR1 array, suggesting its loss as well. We also observed substantial conservation of CRISPR2-leader throughout *S. enterica subsp. enterica*. With this background, we propose the following. Apparently, one, few or all the serovars belonging to the CRISPR1-STY/*cas*-STY clade could have acquired CRISPR1-STY leader and *cas*-STY operon from an unknown source, possibly by HGT event in the gut of primates, subsequently transmitting amongst other *Salmonella* strains or other genera whereas the CRISPR2 leader remained unaffected. However, one cannot rule out a similar possibility for a CRISPR1-STM/*cas*-STM type system. The inheritance of the CRISPR1-STY/*cas*-STY system perhaps renders a competitive advantage in primate gut to the strains possessing it in terms of its pathogenicity and enhanced survival in hostile conditions. For better insights, we investigated the CRISPR-Cas evolution across the *Enterobacteriaceae* family.

The comparative analyses of the phylogenetic trees across the *Enterobacteriaceae* family for the CRISPR-Cas across components highlight a distinct pattern of evolution among CRISPR-Cas systems but present strong evidence of coevolution overall. *cas1* and *cas2* genes are known to be highly conserved and are reflected through the *cas1-cas2* phenogram. We found that the *cas* genes and the CRISPR1 locus are highly variable within closely related *Enterobacteriaceae* species, even among serovars/strains of the same species like *E. coli* and *Salmonella*. However, the leader sequences showed conservation within strains of the same species, with a few minor differences. Interestingly, the DR sequences of the CRISPR array were highly conserved among the six species, indicating a common ancestral origin for the CRISPR array in these bacteria (Díez-Villaseñor, Almendros, García-Martínez, & Mojica, 2010; Bernick, Cox, Dennis, & Lowe, 2012). This suggests a common ancestor for the CRISPR array for these species. Intriguingly, we found that some species share their protospacer sources, possibly because they inhabit similar habitats.

The phenogram of the housekeeping gene, *gyrB,* for the six *Enterobacteriaceae* species, depicts a consistent grouping of the closely related strains belonging to the same species. However, the phenograms of different CRISPR-Cas components were incongruous with the *gyrB* phenogram. The topologies of CRISPR-Cas phenograms showed intermixing of some strains of different species, indicating species relatedness or HGT across strains. This indicates a different evolutionary path of the CRISPR-Cas system to that of the housekeeping genes. Further, as these species have ecological equivalence, insights from our study may hint at a shared evolutionary history of the CRISPR-Cas system in these species.

Chapter 4

Analysing self-targeting CRISPR spacers in *Salmonella* to understand their role in endogenous gene regulation

4.1 Introduction

The CRISPR-Cas system, comprised of clustered regularly interspaced short palindromic repeats and CRISPR-associated proteins, is an adaptive immune mechanism in bacteria. Its primary function is to safeguard against invading bacteriophages and plasmids by integrating new spacers that correspond to the fragments from the genetic material (protospacers) of these intruding entities (Brouns *et al*., 2008). Nevertheless, the discovery by Zegans *et al.,* introduced a paradigm shift, raising questions about the role of CRISPR in regulating endogenous genes within the bacterial genome (Zegans *et al*., 2009). They demonstrated that the CRISPR-Cas system PA14 of *Pseudomonas aeruginosa* controls the DMS3 prophage-dependent inhibition of the biofilm and swarming motility (Zegans *et al*., 2009). Further, Vercose *et al.,* discovered self-targeting spacers (STS) in the type I-F CRISPR-Cas system of *Pectobacterium atrosepticum* (Vercoe *et al*., 2013).

Over the years, various instances have highlighted the involvement of the CRISPR-Cas system in non-canonical functions in various bacteria (Aklujkar & Lovley, 2010; Hale *et al*., 2012; Sampson & Weiss, 2013). Numerous instances shed light on the role of type I-E Cas3 in gene regulation, like those responsible for biofilm formation and fluoride resistance in *Streptococcus* (Tang *et al*., 2019) and virulence in *Porphyromonas gingivalis* (Solbiati, Duran-Pinedo, Godoy Rocha, Gibson, & Frias-Lopez, 2020). A computational study by Bozic *et al.,* revealed that in *E. coli*, the type I-E CRISPR-Cas system spacers primarily target the host genome rather than bacteriophage genomes. The analysis indicates a non-random distribution of hits in the host genome, with a preference for the reverse strand and regions associated with transcription or its regulation (Bozic, Repac, & Djordjevic, 2019).

In *Salmonella*, three studies have elucidated the association of the type I-E CRISPR-Cas system with its physiology. In *S*. Enteritidis, the deletion of *cas3* influenced quorum-sensing (QS) genes, type three secretion systems (T3SS), *Salmonella* pathogenicity island-1 (SPI-1), and genes associated with flagella formation (Cui *et al*., 2020). Deleting the type I-E CRISPR-Cas system in *S*. Typhi reduced the expression of the outer membrane proteins thereby impacting oxidative stress response, bile salt resistance, osmotic balance, chemotaxis, and virulence (Medina-Aparicio *et al*., 2021). Sharma *et al.,* found that the CRISPR-Cas system of *S.* Typhimurium enhances surface-attached biofilm formation while inhibiting pellicle-biofilm formation (Sharma, Das, Raja, & Marathe, 2022).

In this study, our objective is to thoroughly analyse the CRISPR-Cas system within different *Salmonella* serovars, with a specific emphasis on its prevalence, diversity, and potential functions beyond adaptive immunity. Our dataset comprises information extracted from 12,244 *Salmonella* strains, revealing the predominant presence of the type I-E CRISPR-Cas system with very few protospacers.

To delve deeper into the intricacies of the *Salmonella* CRISPR-Cas system, we narrowed our focus to three specific serovars: Enteritidis, Typhimurium, and Typhi. Our analysis focused on CRISPR spacers and their respective gene targets, assessing the functional implications of these interactions. Noteworthy findings emerged, highlighting the roles of the CRISPR-Cas system in regulating key metabolic genes and underscoring the multifaceted nature of the CRISPR-Cas system in *Salmonella*.

4.2 Materials and Methods

4.2.1 Sequence data collection and identification of plasmids and prophages

A comprehensive dataset of 16,506 *Salmonella* genomes was downloaded from the PathoSystems Resource Integration Center (PATRIC) (Gillespie *et al.*, 2011) and NCBI genome databases in May 2021. After removing the duplicates and assessing the genome quality using BUSCO (Manni, Berkeley, Seppey, Simão, & Zdobnov, 2021) and FastANI (Jain, Rodriguez-R, Phillippy, Konstantinidis, & Aluru, 2018), we retained 12,244 genomes. The coding region of the bacterial genome was obtained using the tool Prodigal (Hyatt *et al.*, 2010).

Prophage regions were detected using Phigaro version 2.2.6 (Starikova *et al.*, 2020) in default mode. We determined host specificity through a literature review and classified the *Salmonella* serovars as host-specific, host-adapted, or with a broad host range, while serovars with limited data were marked as having an unknown host range (Eswarappa, Karnam, Nagarajan, Chakraborty, & Chakravortty, 2009; Andino & Hanning, 2015).

4.2.2 Analysis of the CRISPR-Cas components and protospacers target hits

The CRISPRCasTyper version 1.6.4 (Russel, Pinilla-Redondo, Mayo-Muñoz, Shah, & Sørensen, 2020) was employed to detect the CRISPR-Cas genes and arrays in the dataset of 12,244 genomes. The correct orientation of the CRISPR array was determined with respect to the orientation of the *cas* operon. Various aspects of the spacers, including spacer count per array per bacterial strain, spacer length, spacer position in the CRISPR array, and intra-serovar spacer conservation, were analysed using custom Python scripts.

The plasmid database was downloaded from the PLSDB (Schmartz *et al.*, 2022), and the bacteriophage database was sourced from the NCBI bacteriophage database as of September 2023. The obtained CRISPR spacers were aligned against the plasmid and bacteriophage sequences using nucleotide BLAST with criteria set at 80% query cover and 90% identity.

4.2.3 Detection of self-targeting spacers

The CRISPR spacers of *Salmonella* serovars Enteritidis, Typhimurium and Typhi were subjected to nucleotide BLAST against their target genes with criteria set at word size of 5 bp and E-value of 1. Interactions between the spacer and the target occurring at the DNA level (binding to non-coding strand) were categorised as "RNA-" interactions, while interactions at the anti-sense strand/mRNA level (binding to coding strand/mRNA) were categorised as "RNA+" interactions. These interactions were analysed using custom Python scripts, considering the orientation of the gene, the CRISPR array in the genome and the BLAST.

Assuming that crRNA binds to the target RNA the overall binding energy of crRNA binding to its target RNA was assessed using the IntaRNA tool (Mann, Wright, & Backofen, 2017) while considering the accessibility and seeding of potential interaction sites within the RNA molecules.

4.2.4 Annotation of the genome, analysis of the target genes and their pathways

Functional annotation of the genes was conducted using Prokka version 1.14.6 (Seemann, 2014) and eggNOG classification (Huerta-Cepas *et al.*, 2019). We identified candidate genes for Enteritidis, Typhimurium and Typhi serovars for further analysis. The genes were selected if they were targeted by any spacer in at least 1% of the strains and created a network illustrating these interactions using Cytoscape (Shannon *et al.*, 2003). The associated pathways for these genes were determined using the Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa, Goto, Kawashima, Okuno, & Hattori, 2004).

4.2.5 Analysis of putative Protospacer Adjacent Motifs

Putative PAM sequences were analysed by inspecting the three-nucleotide segment preceding the spacer's alignment with the target DNA using custom Python scripts. For the type I-E CRISPR-Cas system, the established and well-recognised PAM sequence is AWG. Additionally, a study by Stringer *et al.,* has provided experimental evidence of the binding of Cas5 to chromosomal regions, and the consensus PAM sequences were AWG, AWA, AWC, and TTR (Stringer, Baniulyte, Lasek-Nesselquist, Seed, & Wade, 2020). Computational analysis by Nobrega *et al*., showed the consensus PAM sequence for type I-E to be AWG, AGG and GAG (Nobrega, Walinga, Dutilh, & Brouns, 2020). A study by Fineran *et al.,* in *E. coli* depicted 29 PAMs that cause the binding of the Cascade complex. These include (WWR, RWR, RRR, GRW, WWG, WWA, WWY, RWY, and WCA) (W corresponds to A and T; R corresponds to G and A; Y corresponds to C and T) (Fineran *et al*., 2014). Hence to ensure the comprehensiveness of our study, we scrutinised all 64 potential combinations and compared our findings with the available literature.

4.2.6 Identification of anti-CRISPR proteins

The amino acid sequences of identified anti-CRISPRs (Acrs) within the type I-E CRISPR-Cas system were employed for a similarity search against the unique genes of *S.* Enteritidis, Typhimurium, and Typhi strains using protein BLAST with an E-value of 1e^{-3}.

4.2.7 Statistical Analyses

Simple linear regression and Pearson correlation analysis were performed on the count of CRISPR spacers per genome versus genome size, GC content of the genome, prophage count and length of prophages in the genome using GraphPad Prism 9.2.0. The significance level was set at a two-tailed P-value with a confidence interval of 95%.

4.3 Results

4.3.1 Type I-E CRISPR-Cas is a predominant defence system within *Salmonella* with variations in the CRISPR attributes

We extracted the CRISPR-Cas system in 12,244 *Salmonella* strains comprising two species and six subspecies of *Salmonella*, with 46 serovars of *Salmonella enterica* subsp. *enterica.* Of the 12,244 strains examined, most (~94%, 11,525) strains contain the type I-

E CRISPR-Cas system, with exceptions found exclusively within strains of *S. enterica* subsp. *enterica* serovars Brandenburg, Lubbock, Reading, Panama, Mbandaka, Johannesburg, Javiana, and Worthington (**Fig. 4.1A**). Notably, the CRISPR-Cas system was absent in many strains belonging to species and subspecies, except subsp. *arizonae* that infect cold-blooded hosts.

Our analysis primarily centred on strains featuring less than 10 CRISPR arrays, CRISPR arrays with fewer than 70 spacers, and spacers shorter than 70 bp. This is because larger arrays and spacers are infrequent in providing a reliable statistical analysis (Nobrega *et al.*, 2020). On average, most strains from serovars with a broad host range contained two CRISPR arrays, except for Indiana, which contained a single CRISPR array. In contrast, strains from host-restricted or host-adapted serovars like Typhi, Pullorum, Dublin, and Choleraesuis generally possessed only one array. Only nine out of 2,440 strains in Typhi, two out of five strains in Pullorum, two out of 133 strains in Dublin, and two out of seven in Choleraesuis contain more than one CRISPR array. Serovar Paratyphi A and Gallinarum have a median of two CRISPR arrays (**Fig. 4.1B**). The median count of CRISPR spacers was higher in broad-host range serovars and lower in host-restricted and host-adapted serovars (**Fig. 4.1C**). The median length of all spacers was 32 bp, except for serovars Gallinarum, Pullorum, Dublin, Mbandaka, Johannesburg, and Schwarzengrund, having a median length of 34 bp, and Enteritidis, which featured a median length of 33 bp (**Fig. 4.1D**).

4.3.2 Most of the CRISPR protospacers are not within phages and plasmids

We analysed the spacer counts within the genomes, assessing their correlation with genome size, GC content, prophage count and length of the prophage content. Our findings reveal that while an increase in spacer count exhibits a moderately significant correlation (r-value: 0.452) with the increase in genome size, we did not see any statistically significant association with the GC content, prophage count and length of the prophage content (**Fig. 4.2**).

Next, we inspected the putative spacer targets within the reported plasmid and phage genomes. Our analysis involved mapping a unique dataset of *Salmonella* spacers, encompassing 7,624 distinct sequences, against two databases: plasmids containing 34,513 sequences and phages containing 8,750 sequences. The result of this analysis show

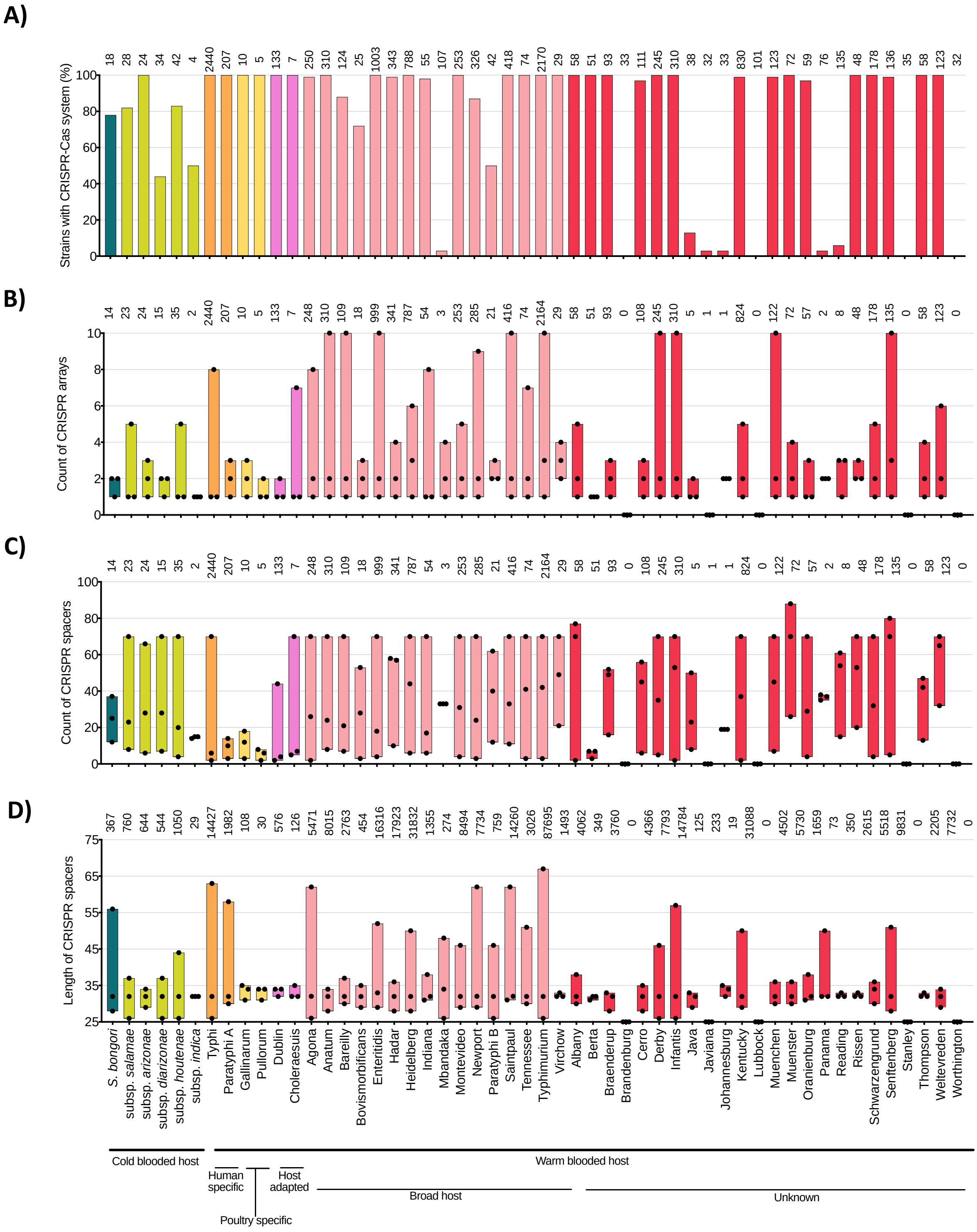


Figure 4.1 Analysis of the CRISPR Cas system in *Salmonella*. A) Percent of strains with the CRISPR-Cas system. **B)** Count of the CRISPR array. **C)** Count of CRISPR spacers. **D)** Length of CRISPR spacers. The X-axis displays the diversity of *Salmonella enterica*, encompassing two species, six subspecies of *S. enterica*, and 46 serovars within *Salmonella enterica* subsp. *enterica*. These serovars are colour-coded to indicate their host specificity, categorised as host-specific, host-adapted, broad host range, or unknown host range. The N at the top of the bar represents the sample size for each analysis. The three spheres inside the bars represent the maximum, median, and minimum counts for each parameter.

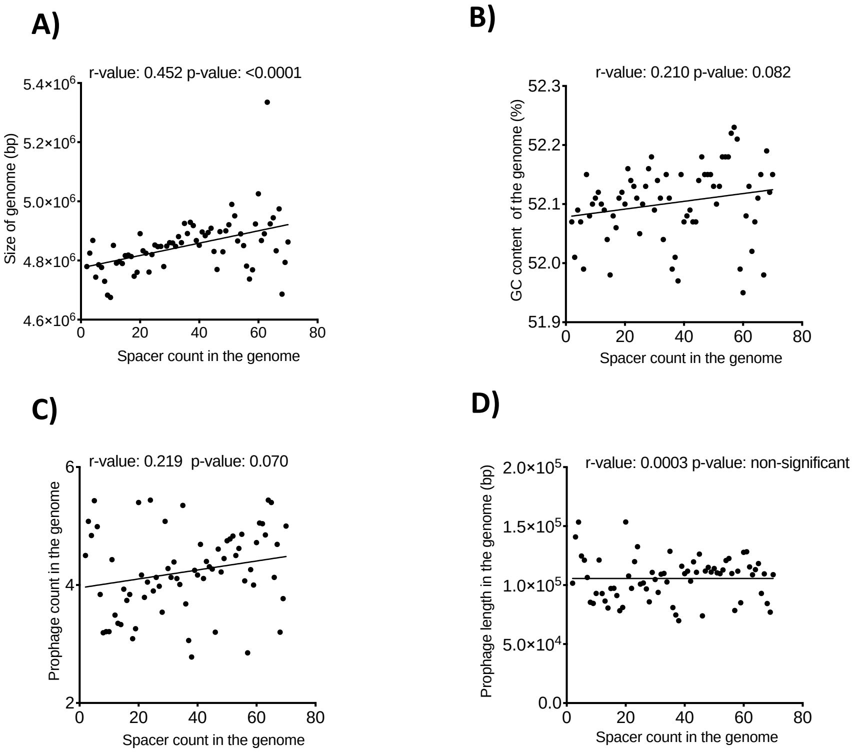


Figure 4.2 Relationship between the spacer count in the genome versus A) Size of the genome, B) GC content of the genome, C) Prophage count in the genome and D) Prophage length in the genome. The X-axis values are in the increasing number of spacers.

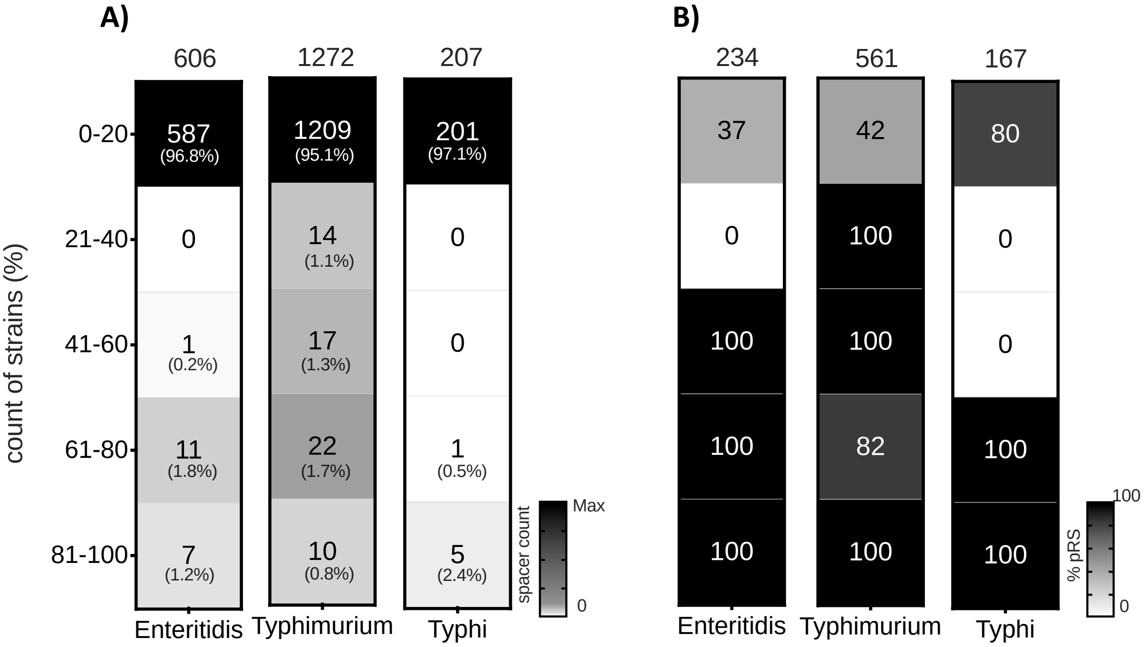


Figure 4.3 Prevalence of spacer conservation and potential regulatory spacers (PRS) conservation among the strains of *S.* Enteritidis, Typhimurium and Typhi. A) Spacer conservation- The Y-axis represents the percentage of strains in the genome. The N at the top of each bar represents the count of unique spacers. The values and the percentage within the box indicate the count of unique spacers in the specified spacer conservation range. **B)** PRS conservation- The percentage inside the box shows the percentage of PRS to the total count of unique spacers within the designated range of spacer conservation.

that 365 spacers displayed significant matches against 2,674 plasmids predominantly belonging to the *Enterobacteriaceae* family. Additionally, we observed 43 spacers with matches against 21 phages, primarily associated with *Salmonella* phages, especially Gifsy-1 and Gifsy-2.

4.3.3 A significant proportion of CRISPR spacers show a partial match with endogenous genes

In **Chapter 3,** we identified two distinct CRISPR systems specific to typhoidal and non-typhoidal *Salmonella* strains. So, for further analysis, we selected representative serovars, Enteritidis, Typhimurium and Typhi, from these two categories and also based on the availability of literature reporting CRISPR-Cas's role in *Salmonella* physiology (Cui *et al*., 2020; Medina-Aparicio *et al*., 2021; Sharma *et al*., 2022). In our genome dataset of 12,244 *Salmonella* strains, 999, 2164, and 2,440 strains of serovar Enteritidis, Typhimurium and Typhi, respectively, contain CRISPR-Cas system (**Fig. 4.1**). On average, they have two, three and one CRISPR arrays, respectively. Detailed scrutiny, revealed that in serovar Typhimurium, the CRISPR array adjacent to the *cas* operon was detected as two separate arrays owing to the presence truncated spacer and direct repeat. Thus, on average, serovar Typhimurium contains two CRISPR arrays.

Serovar Enteritidis, Typhimurium and Typhi contain 606, 1272, and 207 unique spacers, respectively, of which 7, 10 and 5 are conserved in more than 80% of the strains in respective serovars (**Fig. 4.3A**). Even though the spacers are highly conserved they occupy random positions in the array of different strains. Next, we aligned the spacer sequences with the coding regions of the respective bacterial genomes and identified the potential spacer targets. As per the literature, the Cascade complex of the type I-E CRISPR-Cas system can bind to the target DNA with as little as 5 bp complementarity between the crRNA and target RNA (Cooper, Stringer, & Wade, 2018). Therefore, to identify the spacer targets we performed a nucleotide BLAST with a word size criterion of 5 bp. The spacers yielding gene hits were denoted as potential regulatory spacers (PRS) due to their potential to regulate gene expression, possibly by interfering with the transcription/translation process. The data indicates that spacers found in less than 20% of strains, except for serovar Typhi can occasionally act as PRS, whereas spacers present in more than 80% of strains are surely PRS (**Fig. 4.3B**).

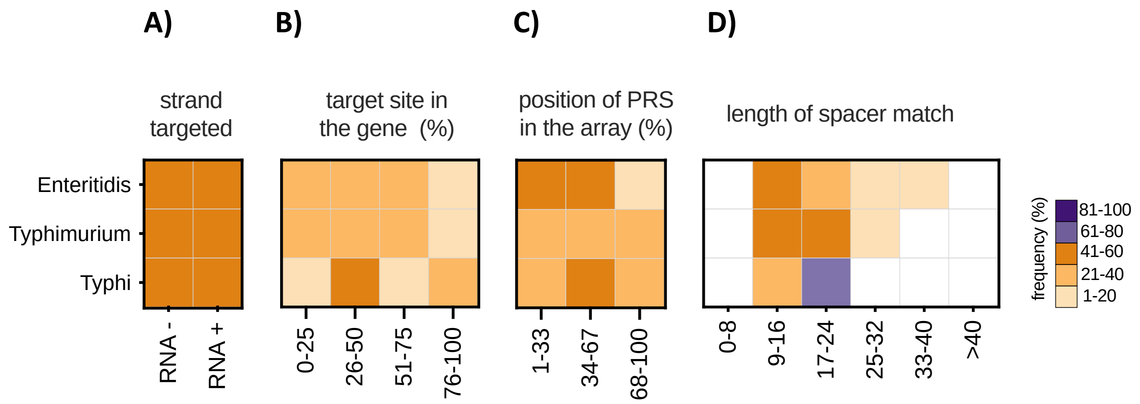


Figure 4.4 Attributes of potential regulatory spacers (PRS) spacers in *S.* Enteritidis, Typhimurium and Typhi. A) Preference of PRS for targeting sense or anti-sense strands. **B)** Preference of PRS for locations in the gene. The reading frame of the genes was segmented into quartiles, ranging from 0-25%, 26-50%, 51-75%, and 76-100%, and the location of the spacer sequence match was categorised accordingly. **C)** The position of the PRS in the CRISPR array. The directional orientation of the CRISPR array is defined with respect to its proximity to the leader. The CRISPR array was segmented into thirds – 1-33%, 34-67% and 68-100%. The position of the spacer was classified accordingly. **D)** The length of the PRS targeting the genes. The length was categorised into 6 quartiles, spanning 8 bp each.

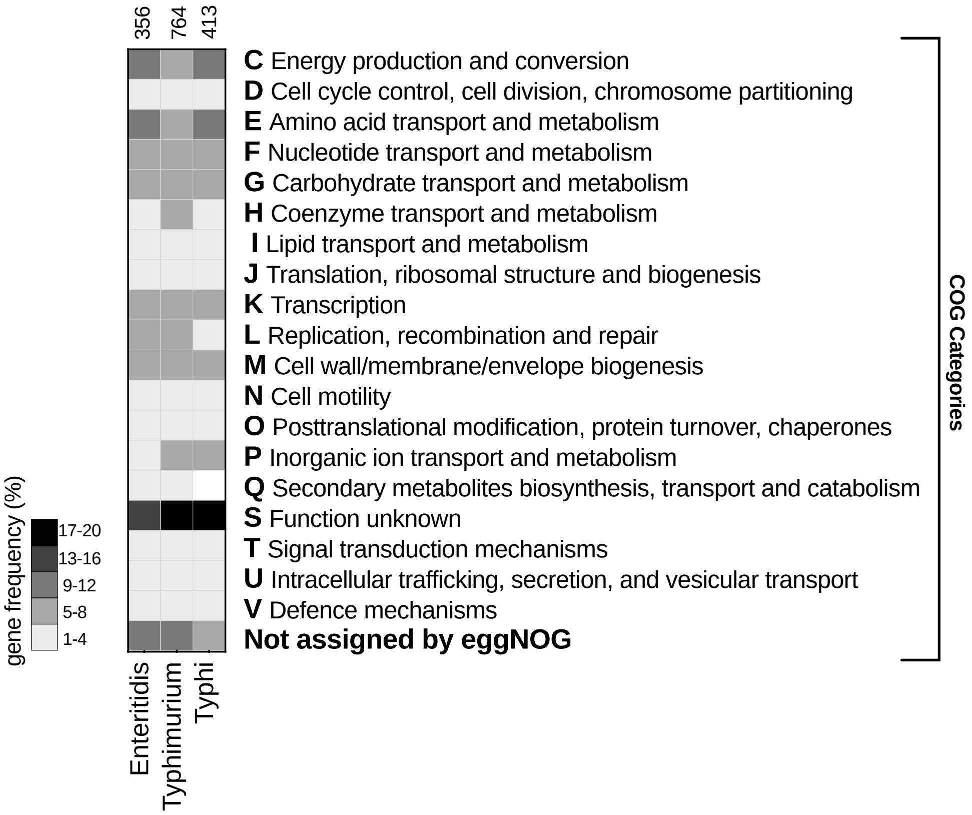


Figure 4.5 The Frequency distribution of the functional diversity among the genes targeted by potential regulatory spacers (PRS) in *S.* Enteritidis, Typhimurium and Typhi as per eggNOG classification.

Further, we analysed the targeting preferences of the CRISPR-Cas system within the coding/sense (RNA-) and non-coding/anti-sense (RNA+) strands of the gene. No discernible preference for either gene strand was observed across the three serovars (**Fig. 4.4A**). To further investigate the target site preferences within the gene, the gene was divided into quartiles of 25% of its length. Analysis of serovar Typhi indicated a slight preference for PRS targets within the second (26-50%) quartile of the gene. For serovars, Enteritidis and Typhimurium a lower preference was observed for PRS targets within the fourth (76-100%) quartile of the gene length (**Fig. 4.4B**). The PRS of serovar Typhi are situated in the 34-67% of the array and in serovar Enteritidis, they are majorly first 67% of the CRISPR array, whereas no such preference was observed in Typhimurium (**Fig. 4.4C**). Though we adopted a word size criterion of 5 bp in the BLAST analysis, we observe that the length of PRS complementary to the genomic regions is greater than 8 bp (**Fig. 4.4D**). For serovar Typhi, the size of the PRS targeting the genes ranged from 17-24 bp in more than 80% of the cases, while for serovars Enteritidis and Typhimurium it was variable with some showing 100% spacer match (**Fig. 4.4D**).

4.3.4 Comprehensive analysis of the PRS gene targets in *S.* Enteritidis, Typhimurium and Typhi

To gain insights into the functional relevance of the PRS, we analysed the biochemical processes regulated by the genes being targeted. The target gene classification was performed using the eggNOG database (**Fig. 4.5**). It was observed that a significant proportion of the gene targets lacked assigned categories or were placed in the "function unknown" category. Analysing those allocated to the defined eggNOG category revealed predominant associations with critical biological processes. Notably, a substantial portion of the gene targets appeared to play roles in energy production, amino acid transport and metabolism, carbohydrate transport and metabolism, nucleotide transport and metabolism, transcription, and cell wall/membrane biogenesis. Further stratification based on *Salmonella* serovars indicated distinct patterns. In the case of Enteritidis and Typhi, a significant proportion of gene targets was identified in categories related to energy production and conversion, as well as amino acid transport and metabolism. However, Typhimurium displayed a more diversified distribution, with gene targets present across all aforementioned functional categories (**Fig. 4.5**).

Detailed examination of the genes potentially targeted by PRS was conducted on a refined set encompassing genes targeted in at least 1% of instances/strains. The results for each serovar are discussed below.

Serovar Enteritidis: We identified 21 distinct genes involved in 22 pathways, potentially targeted by 23 unique spacers, as illustrated in **Fig. 4.6**. The following genes were detected as PRS targets in over 50% of the strains - *tag, ahpF, nrfA, pepB, igaA, recA, srlD, ydhP, galT, lptB, cysM, tcuB, cueO, entE, yhhT, fabH* and *mggB* (**Table 4.1**). It is noteworthy that all these genes exhibited an overall negative interaction energy of crRNA binding to the RNA of the gene (**Table 4.2**). Most of the genes are associated with metabolic pathways. These details can be visualised with the interactive networks, https://github.com/SimranKushwaha/Exploring-and-Exploiting-Prokaryotic-Immunity-in-Salmonella (refer to **Appendix II)**. It is interesting to note that the *recA* (DNA repair protein), *igaA* (intracellular growth attenuator protein) and *pepB* (peptidase B) have been identified as PRS targets in 83% of strains with experimentally verified PAM, for *recA* and *igaA* in ~78% instances. We also identified the *cysM* gene (Cysteine Synthase) as one of the PRS targets in 748 (74%) strains, with experimentally verified PAM in 64% of cases.

Serovar Typhimurium: We identified 98 distinct genes, involved in 103 pathways subjected to targeting by 121 unique spacers, as illustrated in **Fig. 4.7**. The analysis in over 50% of the strains showed 50 gene targets (**Table 4.1**) all exhibiting an overall negative interaction energy with the spacers targeting them (**Table 4.3**). These details can be visualised with the interactive networks, https://github.com/SimranKushwaha/Exploring-and-Exploiting-Prokaryotic-Immunity-in-Salmonella (refer to **Appendix II)**. The topmost PRS targets include *bcsC* (cellulase synthase operon protein C), *lon* (protease), *leuO* (HTH-type transcriptional regulator), *mrcB* (penicillin-binding protein B), *mdtB* (multi-drug resistant protein), *cadC* (transcriptional activator) targeted in 89%, 84%, 83%, 78%, 59% and 58% strains, respectively. A significant fraction of the targets had PAMs that are reported to be functional for type I-E CRISPR-Cas system.

Serovar Typhi: We identified 29 distinct genes, involved in 39 pathways subjected to targeting by 16 unique spacers, as illustrated in **Fig. 4.8**. The analysis in over 50% of the strains showed the gene targets *accA, ispF, ruvB, glcR, manX, ppc, serB, sipD*, ATP synthase, *birA, dacD, ratB*, Endonuclease, *murA, yciH* and *gtrB* (**Table 4.1**) all exhibiting an overall negative interaction energy with the spacers targeting them (**Table 4.4**). These details can

Table 4.1 Gene targeted by potential regulatory spacers (PRS) and their products

Gene	Product
accA	Acetyl-coenzyme A carboxylase carboxyl transferase
ahpF	Alkyl hydroperoxide reductase
alsT	Amino-acid carrier protein alst
appB	Cytochrome bd-II ubiquinol oxidase
arnE	Putative 4-amino-4-deoxy-L-arabinose-phosphoundecaprenol flippase
aroA	3-phosphoshikimate 1-carboxyvinyltransferase
bcsC	Cellulose synthase operon protein C
bepF	Efflux pump periplasmic linker
bioA	Adenosylmethionine-8-amino-7-oxononanoate aminotransferase
bioH	Pimeloyl-[acyl-carrier protein] methyl ester esterase
birA	Bifunctional ligase/repressor
cadC	Transcriptional activator
cbiE	Cobalt-precorrin-7 C(5)-methyltransferase
clcB	Voltage-gated clc-type chloride channel
cueO	Blue copper oxidase cueo
cysM	Cysteine synthase B
dacD	D-alanyl-D-alanine carboxypeptidase dacd
degQ	Periplasmic ph-dependent serine endoprotease degq
entE	Enterobactin synthase component E
fabH	3-oxoacyl-[acyl-carrier-protein] synthase 3
fmt	Methionyl-trna formyltransferase
frdA	Fumarate reductase flavoprotein subunit
fruA	PTS system fructose-specific EIIB'BC component
galT	Galactose-1-phosphate uridylyltransferase
gapA	Glyceraldehyde-3-phosphate dehydrogenase A
gdhA	Glutamate dehydrogenase
glcR	HTH-type transcriptional repressor
gtrB	Bactoprenol glucosyl transferase
igaA	Intracellular growth attenuator protein
ispF	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase
lepA	Elongation factor 4
leuC	3-isopropylmalate dehydratase
leuO	HTH-type transcriptional regulator
lon	Lon protease
lptB	Lipopolysaccharide export system ATP-binding protein
manX	PTS system mannose-specific EIIAB component
mdtB	Multidrug resistance protein
mggB	Mannosylglucosyl-3-phosphoglycerate phosphatase
mrcB	Penicillin-binding protein 1B
mrdA	Peptidoglycan D,D-transpeptidase
murA	UDP-N-acetylglucosamine 1-carboxyvinyltransferase
napA	Periplasmic nitrate reductase
narG	Respiratory nitrate reductase 1 alpha chain
nrfA	Cytochrome c-552
pepB	Peptidase B
pgi	Glucose-6-phosphate isomerase
ppc	Phosphoenolpyruvate carboxylase
purH	Bifunctional purine biosynthesis protein
ratB	Outer membrane protein
recA	DNA repair protein
rep	ATP-dependent DNA helicase
ruvB	Holliday junction ATP-dependent DNA helicase
SAM_*YgiQ*	Ygiq family radical SAM protein
selD	Selenide, water dikinase
serB	Phosphoserine phosphatase
sifB	T3SS effector protein
sipD	Cell invasion protein
srlD	Sorbitol-6-phosphate 2-dehydrogenase
srlR	Glucitol operon repressor
tag	DNA-3-methyladenine glycosylase 1
tcuB	Tricarballylate utilization protein B
troA	ABC transporter substrate-binding protein
trpE	Anthranilate synthase component 1
uspG	Universal stress protein UP12
xylB	Xylulose kinase
yagG	Putative glycoside/cation symporter
ycaD, yciH, yhhT, yifk	Putative protein
ydhP	Inner membrane transport protein
ydiB	Quinate/shikimate dehydrogenase
yfhM	Alpha-2-macroglobulin
YhdP	Asma2 domain-containing protein
yhhJ	Inner membrane transport permease
yojI	ABC transporter ATP-binding/permease protein
yopJ	Effector protein

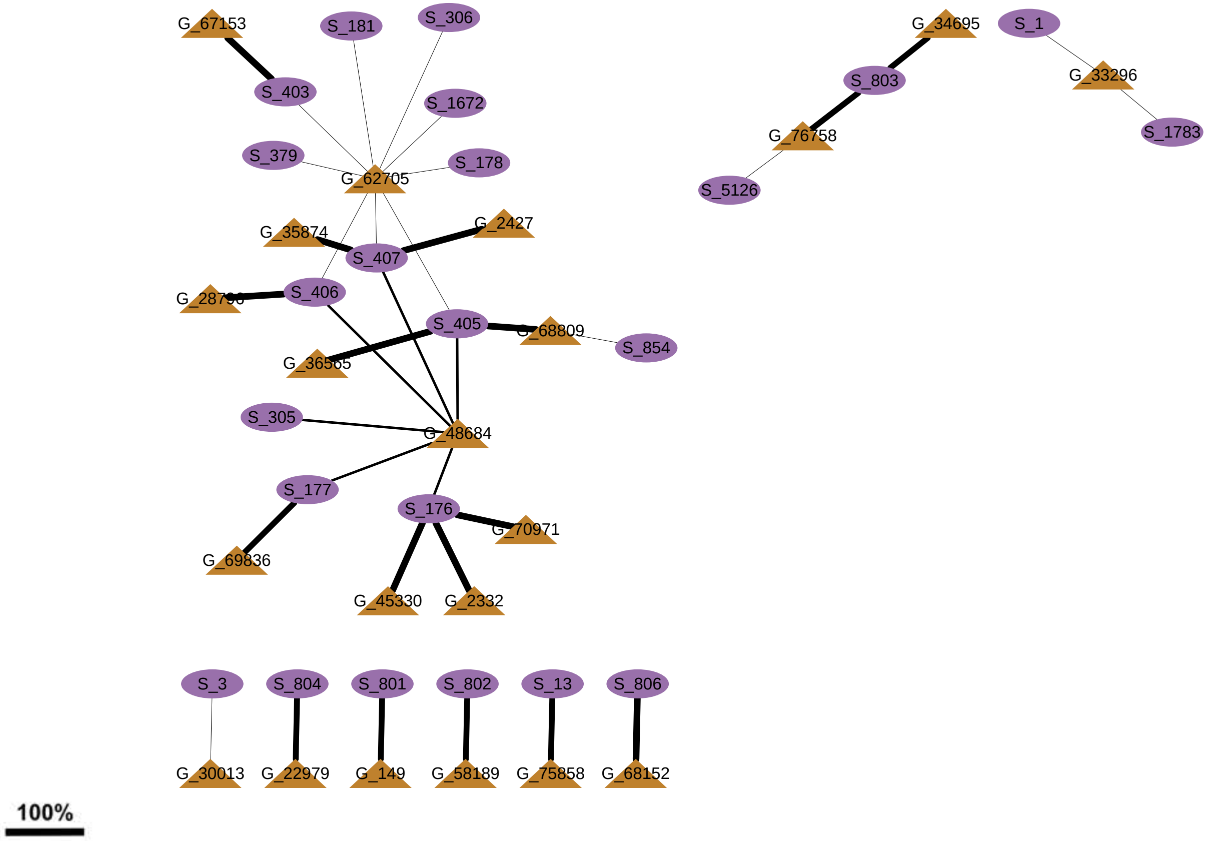


Figure 4.6 The network depicting potential regulatory spacers (PRS) spacers targeting genes in at least 1% of *S.* Enteritidis strains. The purple ellipses represent the PRS, while the brown triangles symbolise the genes (the names can be found in **Table 4.2**). The thickness of the connecting lines from the spacer to the gene is proportional to the number of strains in which the spacer is targeting the given gene.

Table 4.2 Detailed insights into significant genes targeted by potential regulatory spacers (PRS) in *S.* Enteritidis

Gene number	Gene name	Interacting energies (range) kcal/mol		Pathways associated											PAM																											
		Lowest	Highest	1	2	3	4	5	6	7	8	9	10	11	AAA	AAC	AAT	ACG	AGC	ATT	CAT	CCA	CCC	CCG	CGA	CGC	CGG	CTA	GAA	GCA	GCC	GCT	GTA	TAC	TAG	TAT	TCC	TCG	TGA	TGC	TGG	TTT
G_67153	*tag*	-22.21	-22.21						■								177														660											
G_36565	*ahpF*	-9.56	-9.56												656																	170										
G_2427	*nrfA*	-9.42	-9.42				■					■												250									552									
G_68809	*pepB*	-23.14	-13.12	■								■																									654				170	
G_35874	*igaA*	-13.67	-13.67																180										654													
G_28796	*recA*	-15.65	-15.65						■						662																							169				
G_2332	*srlD*	-21.87	-21.87		■							■					172											605														
G_70971	*ydhP*	-10.78	-10.78													172									604																	
G_45330	*galT*	-13.12	-13.12		■							■			597																					169						
G_68152	*lptB*	-5.64	-5.64					■																							759											
G_75858	*cysM*	-6.45	-6.45	■								■						482		260																						
G_76758	*tcuB*	-26.82	-15.65																								264												486			
G_58189	*cueO*	-7.44	-7.44																		479																			265		
G_34695	*entE*	-17.5	-17.5									■	■										263								478											
G_149	*yhhT*	-15.98	-15.98																			472																				265
G_69836	*fabH*	-25.36	-25.36								■	■														555				156												
G_22979	*mggB*	-8.49	-8.49																															251	477							

These genes are targeted in over 50% of the strains, excluding hypothetical proteins and those exhibiting overall positive interacting energy. The product details of these genes can be obtained in **Table 4.1**. The energies obtained are in the range between the lowest and highest values after matching PRS with the target gene. The pathways identified using 1 to 11 are 1-amino acid metabolism, 2-carbohydrate metabolism, 3-cellular processes, 4-energy metabolism, 5-environmental information processing, 6-genetic information processing, 7-human diseases, 8-lipid metabolism, 9-metabolic pathways, 10-metabolism-other and 11-nucleotide metabolism. The PAM column enumerates the count of PAM in each category for every gene, with PAM values <100 ignored due to their lack of significance. The grey-boxed PAM sequences correspond to those mentioned in the literature.

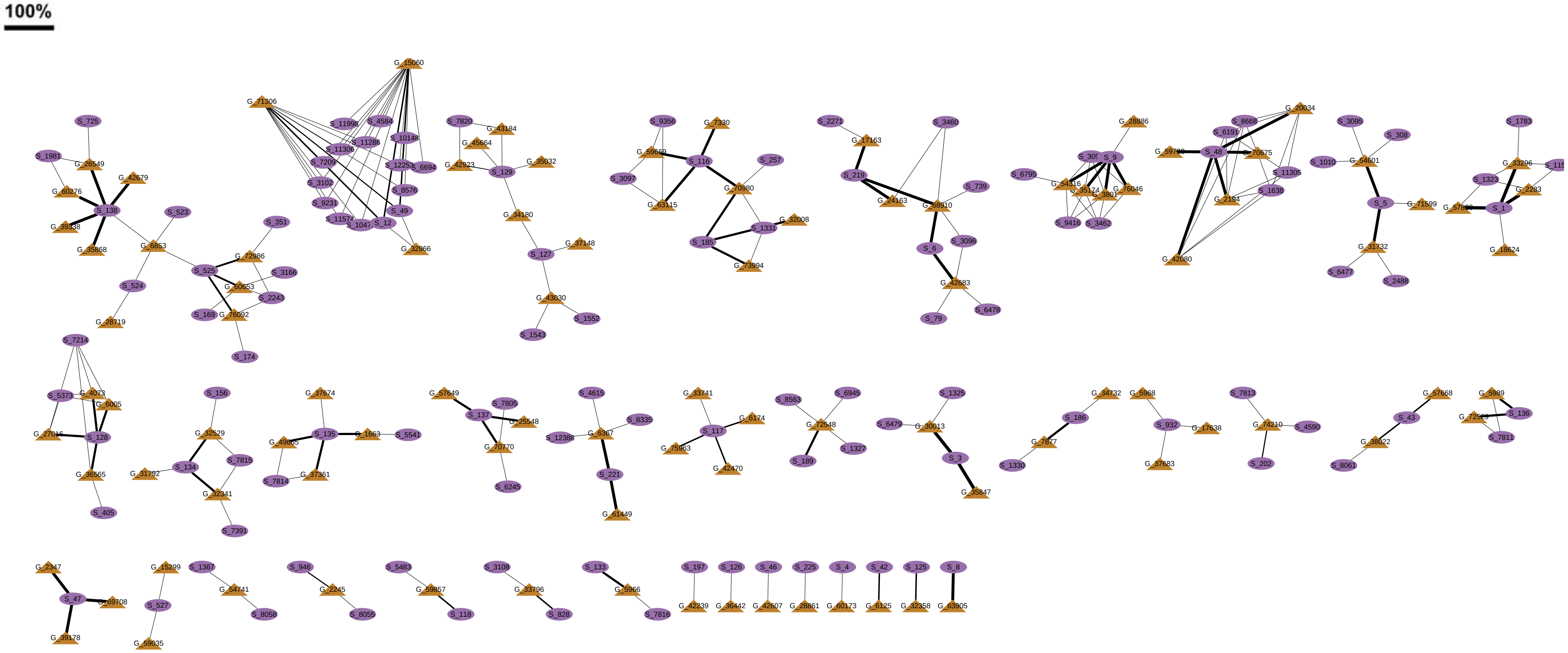


Figure 4.7 The network depicting potential regulatory spacers (PRS) spacers targeting genes in at least 1% of *S.* Typhimurium strains. The purple ellipses represent the PRS, while the brown triangles symbolise the genes (the names can be found in **Table 4.3**). The thickness of the connecting lines from the spacer to the gene is proportional to the number of strains exhibiting this interaction.

Table 4.3 Detailed insights into significant genes targeted by potential regulatory spacers (PRS) in *S.* Typhimurium

Gene number	Gene name	Interacting energies (range) kcal/mol		Pathways associated											PAM																																					
		Lowest	Highest	1	2	3	4	5	6	7	8	9	10	11	AAC	AAT	ACG	AGA	AGC	AGT	ATA	ATC	ATT	CAA	CAG	CAT	CCG	CGA	CGC	CGG	CTG	GAG	GAT	GCG	GCT	GGC	GGG	GGT	GTA	GTC	TAT	TCA	TCG	TCT	TGA	TGC	TGG	TGT	TTA	TTC	TTG	TTT
G_36565	*ahpF*	-14.71	-9.56																																							1268										
G_2347	*alsT*	-6.04	-6.04																				1734																													
G_49665	*appB*	-21.93	-13.23				▨	▨							703														547																							
G_63905	*arnE*	-21.58	-21.58							▨															1677																											
G_37361	*aroA*	-19.73	-12.12	▨								▨																			700												547									
G_2283	*bcsC*	-22.17	-12.68																								793											1052														
G_35174	*bepF*	-21.31	-9.04																											701															866							
G_27016	*bioA*	-18.25	-9.05									▨	▨			524																			700																	
G_73994	*bioH*	-21	-5.59									▨	▨												728			518																								
G_57649	*cadC*	-11.9	-11.9																													1255																				
G_59788	*cbiE*	-13.2	-13.2									▨	▨																																			1689				
G_26549	*clcB*	-14.45	-7.29														1806																																			
G_5989	*degQ*	-16.38	-9.17																										540												701											
G_25548	*fmt*	-7.26	-7.26						▨			▨	▨				1254																																			
G_69708	*frdA*	-7.53	-7.53		▨		▨	▨				▨																																	1734							
G_54316	*fruA*	-20.31	-9.8		▨			▨				▨																		865						703																
G_42683	*gapA*	-22.03	-20.01		▨		▨	▨		▨		▨																		889																						690
G_70675	*gdhA*	-15.68	-10.52	▨		▨	▨					▨																						707												887						
G_30013	*troA*	-7.55	-7																										801																							
G_17163	*YhdP*	-15.69	-15.44																																										1824							
G_71306	*sifB*	-19.69	-5.69																															698																914		
G_39178	Txn regulator	-5.46	-5.46																																																1681	
G_63115	SAM_*YgiQ*	-17.77	-15.15																851																667																	
G_7877	Peroxidase	-18.16	-5.02													544			742																																	
G_4073	HTH	-19.38	-6.96													700												520																								
G_32529	*lepA*	-26.72	-14.12							▨																		1271																								
G_39338	*leuC*	-5.98	-5.98	▨	▨							▨	▨																												1815											
G_42679	*leuO*	-6.78	-6.78																																						1806											
G_60276	*lon*	-9.73	-3.43			▨																		1005							796																					
G_1663	*mdtB*	-20.68	-16.62					▨																												1283																
G_54601	*mrcB*	-23.59	-8.35									▨	▨					1698																																		
G_2194	*mrdA*	-18.92	-13.49							▨		▨	▨								886																								707							
G_31732	*napA*	-18.2	-13.8				▨					▨																															1696									
G_33296	*narG*	-25.08	-15.08				▨	▨				▨																					793							1052												
G_32341	*pgi*	-14.69	-7.33		▨							▨					543																										690									
G_57828	*purH*	-11.4	-9.87							▨		▨	▨	▨																1901																						
G_6005	*rep*	-16.06	-9.15																												698										520											
G_76046	*selD*	-14.49	-5.74	▨								▨			866																			702																		
G_42080	*srlR*	-16.14	-12																																										708	887						
G_68910	*trpE*	-20.75	-11.19	▨		▨						▨	▨				800																																			976
G_5966	*uspG*	-16.34	-4.58																																				541						670							
G_70770	*xylB*	-12.11	-10.99		▨							▨																	535								695															
G_59669	*yagG*	-20.32	-17.54																	850		666																														
G_6367	*ycaD*	-19.58	-12.19																																									989					780			
G_35847	*ydiB*	-9.38	-9.38	▨								▨											1921																													
G_7330	*yfhM*	-17.54	-17.54														1598																																			
G_20034	*yhhJ*	-18.63	-16.18																								703								886																	
G_61449	*yifK*	-16.05	-16.05																																									1819								
G_70980	*yojI*	-23.04	-10.83					▨																		1597																										
G_72548	*yopJ*	-20.74	-8.89							▨																762																					561					

These genes are targeted in over 50% of the strains, excluding hypothetical proteins and those exhibiting overall positive interacting energy. The product details of these genes can be obtained in **Table 4.1**. The energies obtained are in the range between the lowest and highest values after matching PRS with the target gene. The pathways identified using 1 to 11 are 1-amino acid metabolism, 2-carbohydrate metabolism, 3-cellular processes, 4-energy metabolism, 5-environmental information processing, 6-genetic information processing, 7-human diseases, 8-lipid metabolism, 9-metabolic pathways, 10-metabolism-other and 11-nucleotide metabolism. The PAM column enumerates the count of PAM in each category for every gene, with PAM values <100 ignored due to their lack of significance. The grey-boxed PAM sequences correspond to those mentioned in the literature.

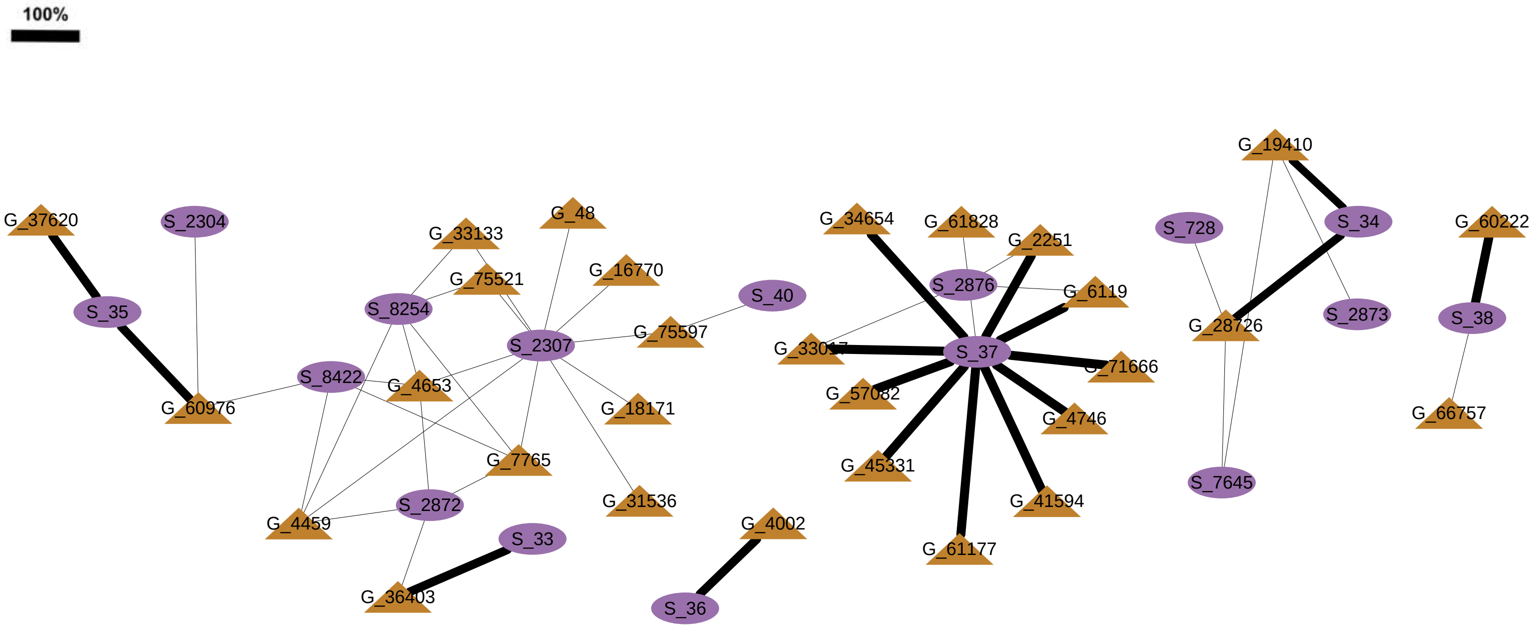


Figure 4.8 The network depicting potential regulatory spacers (PRS) spacers targeting genes in at least 1% of *S.* Typhi strains. The purple ellipses represent the PRS, while the brown triangles symbolise the genes (the names can be found in **Table 4.4**). The thickness of the connecting lines from the spacer to the gene is proportional to the number of strains exhibiting this interaction.

Table 4.4 Detailed insights into significant genes targeted by potential regulatory spacers (PRS) in *S.* Typhi

Gene number	Gene name	Interacting energies (range) kcal/mol		Pathways associated											PAM																
		Lowest	Highest	1	2	3	4	5	6	7	8	9	10	11	AAT	ACG	AGC	AGT	CCG	CGC	CTC	GCC	GCG	GGA	GGC	TAA	TAT	TCA	TTA	TTC	TTG
G_33017	*accA*	-8.63	-7.33		■		■				■	■								2397											
G_6119	*ispF*	-9.1	-8.19									■	■						1206												1188
G_2251	*ruvB*	-13.65	-10.91						■												1189	1206									
G_34654	*glcR*	-8.62	-8.62																												2396
G_45331	*manX*	-22.03	-22.03		■			■				■					1161						1188								
G_57082	*ppc*	-13.85	-13.85		■		■					■									2396										
G_61177	*serB*	-10.41	-10.41	■			■					■								2396											
G_4746	*sipD*	-8.23	-8.23							■										2396											
G_60222	ATP synthase	-15.6	-15.6															1001									1394				
G_41594	*birA*	-6.36	-6.36									■	■									1162				1186					
G_71666	*dacD*	-12.07	-12.07																			1181			1159						
G_4002	*ratB*	-14.45	-14.45													1158			1171												
G_36403	Endonuclease	-27.98	-25.08																	1137										1169	
G_60976	*murA*	-21.01	-10.3		■							■	■								1157			1123							
G_37620	*yciH*	-20.1	-20.1																									2272			
G_19410	*gtrB*	-4.31	-2.29												1022														1047		

These genes are targeted in over 50% of the strains, excluding hypothetical proteins and those exhibiting overall positive interacting energy. The product details of these genes can be obtained in **Table 4.1**. The energies obtained are in the range between the lowest and highest values after matching PRS with the target gene. The pathways identified using 1 to 11 are 1-amino acid metabolism, 2-carbohydrate metabolism, 3-cellular processes, 4-energy metabolism, 5-environmental information processing, 6-genetic information processing, 7-human diseases, 8-lipid metabolism, 9-metabolic pathways, 10-metabolism-other and 11-nucleotide metabolism. The PAM column enumerates the count of PAM in each category for every gene, with PAM values <100 ignored due to their lack of significance. The grey-boxed PAM sequences correspond to those mentioned in the literature.

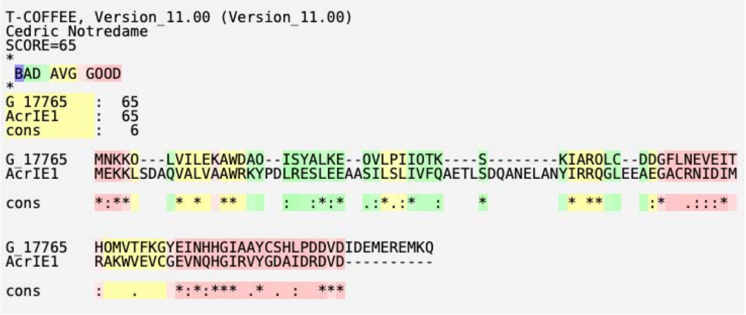
be visualised with the interactive networks, https://github.com/SimranKushwaha/Exploring-and-Exploiting-Prokaryotic-Immunity-in-Salmonella (refer to **Appendix II)**. Notably, *ruvB* (Holliday junction ATP-dependent DNA helicase), *ratB* (outer membrane protein) and *sipD* (cell invasion protein) were identified as PRS targets in 98% of strains. In all cases of *ratB* and 50% of *ruvB*, the PAMs are known to be functional. However, for *sipD* both the PAMs are not reported to be functional.

4.3.5 Analysis of Anti-CRISPR proteins in *S.* Enteritidis, Typhimurium and Typhi

The self-targeting CRISPR spacers exhibit detrimental effects, resulting in self-killing if the spacer shows a 100% protospacer match within its genome. One of the mechanisms preventing self-killing is mutating the protospacer region to reduce complementarity. Partial complementarity prevents self-targeting as the nuclease cannot act. Another way is to mutate or lose *cas* genes to prevent functional activity (Wimmer & Beisel, 2019).

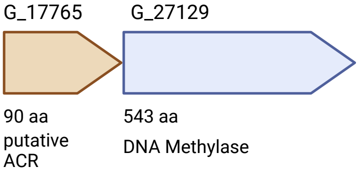
Our study identified PRS, with the majority showing partial complementarity (base-pair match in the range of 9-24 bp, **Fig. 4.4D**). However, serovar Enteritidis and Typhimurium have 12% and 1.5% spacers, respectively, showing a 100% protospacer match and can be self-targeting. Auto-immunity by such self-targeting spacers is generally prevented by using anti-CRISPR proteins (Acr) (Nobrega *et al*., 2020). To identify the Acr proteins in serovar Enteritidis, Typhimurium and Typhi, we mapped all the unique proteins against the already known anti-CRISPR proteins of the type I-E CRISPR-Cas system (Nobrega *et al*., 2020). We found 51 proteins showing homology to existing Acrs (AcrIE1-E7 and AcrIE4-IF7). The percentage protein identity was within the range of 23%-52%. One of the proteins G_17765, identified as a hypothetical protein of 89 amino acids, showed 49% identity with AcrIE1 (90 amino acids) (**Fig. 4.9A**). This protein is present in one strain of Enteritidis and 569 strains of Typhimurium. On a closer look, we found the protein is preceded by a DNA methylase protein in >91% of instances. Based on our analysis and criteria for Acrs identification (Nobrega *et al*., 2020), the gene coding for G_17765 may be a new type of I-E Acr (**Fig. 4.9B**). This gene is predominantly found in *Enterobacteriaceae.*

A)



```
T-COFFEE, Version_11.00 (Version_11.00)
Cedric Notredame
SCORE=65
*
 BAD AVG GOOD
*
G_17765   :  65
AcrIE1    :  65
cons      :   6

G_17765   MNKKQ---LVILEKAWDAQ--ISYALKE--QVLPIIQTK----S--------KIARQLC--DDGFLNEVEIT
AcrIE1    MEKKLSDAQVALVAAWRKYPDLRESLEEAASILSLIVFQAETLSDQANELANYIRRQGLEEAEGACRNIDIM

cons      *:**     * *  **     :  :*:*  .:*.:*  :    *         * **     :*  .:::*


G_17765   HQMVTFKGYEINHHGIAAYCSHLPDDVDIDEMEREMKQ
AcrIE1    RAKWVEVCGEVNQHGIRVYGDAIDRDVD----------

cons      :   .    *:*:*** .* . :  ***
```

B)



>G_17765
MNKKQLVILEKAWDAQISYALKEQVLPIIQTKSKIARQLCDDGFLNEVEITHQMVTFKGYEINHHGIAAYCSHLPDDVDIDEMEREMKQ

>G_27129
MKFKADQTSQKLRGGYYTPQNLADYVTKWVLSKNPKTILEPSCGDGVFIQAIANNGYDPNIELSCFELFDTEASKALDRCKLNNFSNAT
ITEGDFLVWANECLKKNKPIFDGVLGNPPFIRYQFLERNFQEQAQLVFEHLDLKFTKHTNAWVPFLLSSLALLKQGGRIGMVIPSEISHVM
HAQSLRSYLGHVCSKIVIIDPKEIWFEDTLQGAVILLAEKKQYPDEASQGVGIASVSGFEFLQEDPNVLFNDTVGINGETVEGKWTKATLDI
DELQLIKRVIAHPDVRKFKDIAKVDVGIVTGANNYFLVDNETVKSYKLERFAHPMFGRSQHCPGIIYDEKQHIENQEKGLPTNLLYIDEEFED
LSRSVKNYIELGEAEEYHKRYKCRIRKPWFKVPSVYSTEIGMLKRCHDAPRLIHNKVRAYTTDTAYRISSTVTSIENLVCSFLNPLTVITAEL
EGRFYGGGVLELVPSEIEKLYIPIVEGLEHNVEEINLLIKNGQIERVIRQQGLLILDKLGFTQEENEKLVEIWKKLRDRRLRK

Figure 4.9 Detection of Anti-CRISPR proteins. A) Alignment of the putative type I-E anti-CRISPR protein with the known anti-CRISPR AcrIE1. **B)** The arrangement and amino acid sequence of a putative type I-E anti-CRISPR protein and the DNA methylase gene.

4.4 Discussion

The type I-E CRISPR-Cas system in *Salmonella* is a highly conserved defence system present in ~94% of the sequenced strains. The average spacer count per strain varies between the serovars. Likewise, the results from **Chapter 3**, in general, the broad-host range serovars contain higher spacer counts. This observation could be probably because they can infect various hosts and encounter a more diverse set of phages and other foreign genetic elements, allowing them to defend against a broader range of potential threats. While host-restricted or host-adapted serovars have evolved to survive within a specific host environment, the lower spacer count may thus be explained.

A total of 7,624 unique spacer sequences were identified from these strains, but only 4.8% of the spacers had protospacers within plasmids and 0.6% within phages. This suggests that a substantial proportion of the spacers may have alternative functional roles beyond their participation in the adaptive immune system. However, we acknowledge that our analysis may be constrained by the limited availability of plasmid and phage datasets specific to *Salmonella*. Next, we anticipated that the CRISPR-Cas system, acting as an adaptive immune system, would display a distinct negative correlation between the number of spacers in the genome and the prevalence of prophages. We observe patterns similar to those found in other bacterial species like *Streptococcus* for type I-C CRISPR (Yamada *et al*., 2019) with no such correlations, further hinting about its functional roles beyond adaptive immunity.

We observe that in *S*. Enteritidis, Typhimurium and Typhi, all the spacers conserved in more than 80% of strains are PRS. In contrast, only ~40% of the spacers conserved in less than 20% of strains are PRS, except for serovar Typhi, with 80% being PRS. We believe that spacers with lower conservation might be subject to less selective pressure and could be remnants from past interactions with foreign genetic elements or less functionally relevant. However, the highly conserved spacers are generally expected to be under stronger evolutionary pressure, indicating some selective advantage conferred upon the bacteria possessing them. Out of the seven, ten and five spacers conserved in >80% strains of *S.* Enteritidis, Typhimurium and Typhi, five, six and four spacers target persistent genes. For example, the genes targeted in Enteritidis are *ahpF, igaA, nrfA, pepB, recA,* and *tag*; the genes targeted in Typhimurium are *bcsC, clcB, leuC*, *marG, purH, trpE, ycaD* and *yifK*; and the genes in Typhi include *accA, birA, ispF*, *manX, murA, ppc, ruvB, serB, sipD* and *yciH*.

Some of these examples are discussed below. Reports in the literature authenticate some of the identified gene targets of PRS. In serovar Enteritidis, *sipD*, one of the PRS targets, is shown to be significantly downregulated in the strain lacking the *cas3* gene (Cui *et al*., 2020). In serovar Typhimurium str. 14028s, the Cas5 ChIP-seq occupancy data revealed 236 crRNA spacer-Cascade-binding sites (Stringer *et al*., 2020). Some of our PRS target sites that match this data set include *leuO, entE, mrdA, ratB, rep*, and *pgi*. Not all the Cascade-binding sites were detected as the PRS targets, but some of the genes targeted were of the same operon, like *aroA, bcsC, bioA, bioH, cbiE, cysM, fabH, frdA, galT, ispF, leuC, narG, recA, xylB, ycaD, yfhM,* and *yhdP*. This suggests that the CRISPR-Cas system influences specific regulatory pathways. Therefore, we assume that the CRISPR-Cas system could regulate the gene targets highlighted in **Table 4.2, 4.3 & 4.4**.

Further confidence in our PRS target prediction is obtained through the study by Sharma *et al*., 2022. They revealed that the CRISPR-Cas system in *S.* Typhimurium regulates pellicle biofilm by affecting cellulose secretion. They show that the *bcsC* gene (cellulose exporter), one of the PRS targets, is regulated by the CRISPR-Cas system probably *via* complementary base-pairing of the crRNA to the gene. The CRISPR-Cas system is also known to impact biofilm formation in serovar Enteritidis and Typhi (Cui *et al*., 2020; Medina-Aparicio *et al*., 2021). In serovar Enteritidis, *pepB,* a protease, was identified as a PRS target. It is involved in the formation and modulation of biofilms, as well as the degradation of host cell matrices during the pathogenesis. These peptidases also partake in cell signalling, influencing the behaviour of microbial cells within the biofilm (Ramírez-Larrota & Eckhard, 2022).

Our study identified *recA* and *ruvB* genes as PRS targets in Enteritidis and Typhi, respectively. Evidence indicates that the Cas1 protein of *E. coli* genetically interacts with *recA* (DNA repair protein) and *ruvB* (Holliday junction ATP-dependent DNA helicase) (Babu *et al*., 2011). RecA on stimulation by RecBCD inhibits the spacer acquisition by the CRISPR-Cas system (Radovcic *et al*., 2018). Thus, the intricate interplay between Cas1, RuvB, RecBCD, and RecA unravels a complex regulatory network that shapes the dynamics of the CRISPR-Cas system.

Some of the PRS targets are the genes regulating the CRISPR-Cas expression. These targets include *leuO* in Typhimurium and *igaA* in Enteritidis. As shown elsewhere and in **Chapter 5**, LeuO, a pivotal global regulator, positively regulates the CRISPR-Cas expression in *Salmonella* (Medina-Aparicio *et al*., 2011). In *Serratia marcescens,* IgaA positively regulates the CRISPR-Cas expression, acting *via* the Rcs phosphorelay signalling cascade. Under stress conditions, IgaA inhibits the Rcs phosphorelay which is involved in the repression of type I-E, I-F and III CRISPR-Cas expression. The IgaA in *Serratia* and *Salmonella* have ~60% identity (93% query coverage). The gene region targeted by the CRISPR spacer is conserved in both *Salmonella* and *Serratia*. Thus, we theorise that there may be a plausible scenario wherein LeuO and IgaA function as regulators of the CRISPR-Cas system, with reciprocal regulatory interactions.

An inverse correlation has been reported between the CRISPR-Cas system and antibiotic resistance in most pathogens (van Belkum *et al.,* 2015; Li *et al.,* 2018). This is linked to the degradation of antibiotic-resistance genes on the mobile genetic elements by the CRISPR-Cas system. Our analysis detected *mdtB* and *mrcB*, the genes involved in antibiotic resistance, as the PRS targets probably hinting at other regulatory mechanism through which the system contributes to antibiotic resistance. We hypothesise that the bacteria might use the CRISPR-Cas system to selectively modulate the expression of the resistance gene in response to environmental cues. This could allow the bacterium to fine-tune its antibiotic resistance based on the presence or absence of specific selective pressures, such as the presence of varying concentrations of antibiotics in the environment. This conjecture needs to be validated with proper wet-lab experiments. Certain virulence genes like *lon* protease in Typhimurium and *sipD* in Typhi were also detected as PRS targets. Lon protease is essential for systemic infection of *S.* Typhimurium in mice and controls the expression of SPI-1 genes (Jiang, Li, Lv, & Feng, 2019). SipD is an SPI-1 protein essential for the invasion of the host cells. Results from our lab (unpublished data) and Cui *et al.,* suggest downregulation of *sipD* in serovars Typhimurium and Enteritidis respectively. The regulation of *sipD* by the CRISPR-Cas system in serovar Typhi awaits confirmation. Although preliminary results from our lab suggest decreased invasion of the CRISPR-Cas knockout Typhi strains in the intestinal epithelial cells.

Numerous investigations into the type I-E CRISPR system underscore the pivotal role of PAM in the adaptation and interference phase of the CRISPR-Cas complex (Xue & Sashital, 2019). The Cas1-Cas2 adaptation complex exhibits a robust affinity for canonical PAMs, ATG, AAG, AGG, and GAG. Yet, intriguingly, diverse studies propose that Cas1-Cas2 can also engage with non-canonical PAMs (like AGG, AWA, AWC, GAG, TTR, WWR, RWR, RRR, GRW, WWG, WWA, WWY, RWY, and WCA), albeit with diminished affinity. During interference, after (i) base pairing between the crRNA and complementary DNA target, and (ii) sequence recognition, the Cascade complex recruits Cas3 for target degradation (Hochstrasser *et al*., 2014). If the PAM is mutated at two or all three nucleotides, the interference phase is completely blocked, and the Cse1 cannot recruit the Cas3. However, the Cascade complex binds to the target DNA and acts in an interference-independent manner (Fineran *et al*., 2014; Xue & Sashital, 2019). Thus, the canonical and non-canonical PAMs could lead to the binding of the Cascade complex on the target with or without target cleavage. Hence, we hypothesise that irrespective of the presence of correct PAM, the Cascade complex containing the PRS may bind to the target genes and regulate their expression.

Nevertheless, if the CRISPR-Cas system gets activated under various conditions like biofilm and virulence, some STS may lead to self-killing. Hence, we think that there may be Acrs active to prevent self-killing. The analysis for the presence of Acrs in *Salmonella* genomes did not reveal a complete match with any known type I-E Acr. However, we identified one gene that matched with AcrIE1 with an alignment score of 65 and can be thought of as a new type I-E Acr for *Enterobacteriaceae*. It is already known that Acrs cluster with anti-RM and other anti-defence genes like the methyltransferase gene (Pinilla-Redondo *et al*., 2020). We too observed a DNA methylase protein juxtaposed with the putative Acr, hinting that this may be an anti-defence island present in a few strains of *Salmonella.*

In unveiling the intricacies of the CRISPR-Cas system in *Salmonella*, our study provides novel insights into its diverse functional roles beyond adaptive immunity. Our analysis not only confirms known gene hits associated with the CRISPR-Cas system but also expands the repertoire by identifying additional genes that we propose to be regulated by this system. Additionally, the identification of a putative Acr in *Salmonella* opens new avenues for research, underscoring the intricate and dynamic nature of the CRISPR-Cas system in bacterial defence and adaptation.

Chapter 5

Investigating the functional activation of the CRISPR-Cas system and repurposing it for *Salmonella*-specific killing

5.1 Introduction

Salmonellosis is a gastrointestinal illness caused by the bacterium *Salmonella*. It is the most significant among the 22 major food-borne pathogens in terms of impact on disability-adjusted life years (Bintsis 2017, Kurtz, Goggins and McLachlan 2017). Annually, about 14.3 million individuals suffer from typhoid fever, resulting in 136,000 global fatalities (Stanaway *et al.,* 2019). An epidemiological study in parts of Asia revealed that almost 5-7% of those affected by *Salmonella* Typhi were persistent carriers, raising concern as these carriers can be a primary source for *Salmonella* infections (Shu-Kee Eng *et al.,* 2015, Di Domenico *et al.,* 2017). Typhoidal *Salmonella* is more common in underdeveloped regions due to poor sanitation, while non-typhoidal *Salmonella* (NTS) is global (Feasey *et al.,* 2012). *Salmonella* outbreaks are recurrent, such as the 2012 *S.* Typhi outbreak in Northern India linked to water supply issues, incidents involving tainted peanut butter (2006-2007) and beef (2019) in the US, uncooked ham in the Netherlands (2016-2017), and soft cheese in Mexico (2018-2019) (Sheth *et al.,* 2011, Purighalla *et al.,* 2017, Brandwagt *et al.,* 2018, Plumb *et al.,* 2019).

The most prevalent issue is *Salmonella*'s growing antibiotic resistance. Historically, antibiotics like cephalosporins, chloramphenicol, and azithromycin have played a crucial role in mitigating the severity and spread of this bacterial pathogen (Antony *et al.,* 2018, Gut *et al.,* 2018). However, unchecked antibiotic use has led to numerous antibiotic-resistant strains, particularly for Typhimurium, Newport, and Heidelberg serovars (Gut *et al.,* 2018, Wang *et al.,* 2019). Global antibiotic consumption rose by 36% from 2000 to 2010, with India as a major consumer (Britto *et al.,* 2018). A 2018 report noted a 65% increase from 2000 to 2015, driven by China, India, and Pakistan (Klein *et al.,* 2018). Antibiotic overuse, especially cephalosporins in India, worsened the *Salmonella* resistance (Britto *et al.,* 2018). Furthermore, antibiotics' effect on infant gut microbiomes can also hinder immune development and aid *Salmonella* infection. Hence, using antibiotics for basic *Salmonella* gastroenteritis is discouraged (Vangay *et al.,* 2015, Bruzzese, Giannattasio and Guarino 2018).

While various alternatives to antibiotics offer benefits, they also confront regulatory and patent-related hurdles. Antimicrobial peptides, sourced from animals and plants, exhibit potential for targeted infection treatment, albeit with high production costs (Lei *et al.,* 2019). Additionally, predatory bacteria and engineered bacteria designed to target specific strains are under scrutiny. However, concerns linger about their precision, emergence of resistance, and long-term effects (Kadouri *et al.,* 2013). The use of prophylactics and probiotics to treat salmonellosis exists but has limitations (Antony *et al.,* 2018). Two vaccines against *S.* Typhi show just 50-70% efficacy and are not recommended for toddlers (below two years of age) (Gayet *et al.,* 2017). Additionally, no vaccines are available for non-typhoidal salmonellosis (MacLennan, Martin and Micoli 2014, Gayet *et al.,* 2017). In pursuit of precision and diminished resistance, researchers are investigating the CRISPR-Cas system for targeted pathogen killing (Gomaa *et al.,* 2014). Its programmability offers precision and adaptability, potentially providing a sustainable treatment approach. While alternatives to antibiotics have benefits, current challenges require solutions for successful medical use. Innovative, tailored strategies are crucial to address antibiotic resistance effectively.

Hamilton *et al.,* explored the exogenous type II CRISPR-Cas system for *Salmonella* elimination, employing interspecies conjugation to transfer Cas9 and guide RNA *via* plasmids for effective gene targeting and killing (Hamilton *et al.,* 2019). Challenges include toxicity of constitutive Cas9 expression, large-sized DNA to be transferred and escape mutations in protospacer or Cas9. Further, the crRNA vital for Cas9 activity is 20 bp long (Jiang and Doudna 2017) and the double-stranded DNA breaks by Cas9 are repairable (Wimmer and Beisel 2019). To overcome these challenges, we suggest exploiting the endogenous CRISPR-Cas3 system for small-sized DNA to be transferred, increased specificity due to larger (32 bp) crRNA (Kushwaha *et al.,* 2020), and a lesser chance of escape mutations as the system is crucial in regulating vital physiological functions (Medina-Aparicio *et al.,* 2011, Cui *et al.,* 2020, Medina-Aparicio *et al.,* 2021, Sharma *et al.,* 2022).

Salmonella's type I-E CRISPR-Cas system, regulated by LeuO, histone-like nucleoid structuring protein (H-NS), and leucine-responsive regulatory protein (LRP), has intricate functions (Medina-Aparicio *et al.,* 2018, Kushwaha *et al.,* 2022). Under lab conditions, H-NS represses *cas* genes by binding to its low GC-content promoter region (Medina-Aparicio *et al.,* 2011). Even though the system has roles in governing *Salmonella* physiology (Medina-Aparicio *et al.,* 2011, Cui *et al.,* 2020, Medina-Aparicio *et al.,* 2021, Sharma *et al.,* 2022) the conditions activating its functions are less understood.

Our primary objective is to comprehensively characterise the functional activation of *Salmonella*'s endogenous CRISPR-Cas system. By doing so, we can strategically employ the native system for self-targeting, ultimately eradicating the bacteria. This innovative approach holds the promise of effectively tackling the challenges posed by conventional strategies and ushering in a more precise and tailored means of combatting *Salmonella* infections.

5.2 Materials and Methods

5.2.1 Bacterial strains and culture conditions

The parent strain, *S.* Typhimurium str. 14028s (referred to as wildtype, WT 14028s) (**Table 5.1**), was cultivated in Luria-Bertani (LB) medium from HiMedia, supplemented with suitable antibiotics at 37 °C with continuous shaking at 120 rpm.

A. Growth conditions in nutrient-rich media

Overnight-grown WT 14028s bacterial cultures were sub-cultured in triplicates at a 1:100 ratio in LB medium and grown at 37 °C with continuous shaking at 120 rpm. The cells were collected at different time points, specifically at 0.3, 0.6, 1 and 1.5 optical density at 600 nm (OD_{600}), measured using Multiskan GO (Thermo Scientific, USA).

B. Growth in intracellular mimicking conditions F-media

Overnight-grown WT 14028s bacterial cultures were sub-cultured at a 1:50 ratio in triplicates in F-media (5 mM KCl, 7.5 mM NH_4SO_4, 0.5 mM K_2SO_4, 10 mM 2-(N-morpholino) ethane sulfonic acid buffer, 0.27% glycerol, 0.1% Casein Acid hydrolysate and 10 µM $MgCl_2$), pH-5.4 and grown at 37 °C with continuous shaking at 120 rpm. The cells were collected at different time points, specifically at OD_{600} 0.3, 0.6, and 1.

C. Growth conditions for inducing envelope stress with ethylenediamine tetraacetic acid (EDTA)

Overnight-grown WT 14028s bacterial cultures were subcultured in triplicates at a 1:100 ratio in LB medium. The secondary cultures were grown for 1.5 hours. EDTA (HiMedia) was added at concentrations of 0.5 mM, 1 mM, 2.5 mM and 5 mM, and the bacterial cultures were grown at 37 °C with continuous shaking at 120 rpm. The bacterial cells were collected at OD_{600} 1.

D. Biofilm formation

Overnight-grown WT 14028s bacterial culture was sub-cultured in triplicates at 1:100 ratio in biofilm media (LB without NaCl) and grown at 25 °C, static. The planktonic bacteria and biofilms (ring and pellicle) were collected at 24, 48 and 96 hours.

5.2.2 RNA isolation, cDNA synthesis and semi-quantitative RT-PCR

RNA isolation

The bacterial cells grown under different conditions, as mentioned in 5.2.1, were harvested and used for RNA isolation. To aid RNA isolation, the bacterial cells were first lysed using 4 mg/mL lysozyme (GeNei) to break down the cell walls. Total RNA was isolated using TRIzol reagent (HiMedia) and precipitated using isopropanol (HiMedia). The pellet was washed with 70% ethanol (HiMedia) to remove impurities and resuspended in nuclease-free water, yielding the purified RNA preparation.

Total RNA isolation from pellicle biofilm: Pellicle biofilms were resuspended in a solution containing 70% ammonium sulphate and 10% cetyltrimethylammonium bromide (CTAB). Subsequently, the pellicles were gently crushed using a toothpick and incubated at room temperature for 10 minutes. The suspensions were then centrifuged, and the pellets were resuspended in 500 mL of lysis solution (10 mM Tris, 10 mM EDTA, and 1 mg/mL lysozyme), followed by another 10 minutes of incubation at room temperature. Next, 10% SDS and 3M sodium acetate were added to the samples. The RNA was purified using phenol-chloroform-isoamyl alcohol extraction, and the RNA present in the aqueous phase was precipitated overnight at -80 °C using isopropanol. The precipitated RNA was washed with 70% ethanol to remove impurities and resuspended in nuclease-free water, resulting in the purified RNA preparation.

cDNA synthesis and semi-quantitative RT-PCR analysis

After RNA extraction, cDNA synthesis was performed using ProtoScript II reverse transcriptase from NEB. Semi-quantitative RT-PCR was then employed to assess the expression of eight *cas* genes (*cas1, cas2, cas5, cas7, cas6, cas3, cse1*, and *cse2*) using Taq DNA polymerase (GeNei), with 16S rRNA serving as a positive control. The intensity of PCR bands (calculated by the software Image Lab v6.1, Bio-Rad) was utilised to estimate the relative expression levels of the *cas* genes. The primers used in RT-PCR are listed in **Table 5.2**.

Table 5.1 List of bacterial strains and plasmids used in this study

Bacterial Strain	Genotype and Characteristics	Source/Ref
S. enterica serovar Typhimurium str. 14028s	Wildtype	A kind gift from Prof. Dipshikha Chakravorty, Indian Institute of Science, Bangalore, India
S. enterica serovar Typhi str. CT18	Wildtype	A kind gift from Prof. Dipshikha Chakravorty, Indian Institute of Science, Bangalore, India
S. enterica serovar Paratyphi A (MTCC 735)	Wildtype	Microbial Type Culture Collection and Gene Bank, India
S. enterica serovar Welterveden (MTCC 3227)	Wildtype	Microbial Type Culture Collection and Gene Bank, India
WT-pQE60-L-C	WT 14028s transformed with pQE60 containing *leuO* under constitutive T5 promoter	This study
WT-pQE60-L-I	WT 14028s transformed with pQE60 containing *leuO* under the inducible pBAD promoter	This study
WT-pQE60-L-I-CR	pQE60-L-I with a constitutively expressed CRISPR array with spacer against pTarget	This study
WT-pEmpty	WT 14028s transformed with empty pJUMP26-1A vector	This study
WT-pTarget	WT 14028s transformed with pJUMP26-1A vector containing the protospacer	This study
WT-pQE60-L-I-STS	WT-pQE60-L-I with a constitutively expressed self-targeting CRISPR array	This study

5.2.3 Plasmid construction for induction of *cas* genes

The *leuO* gene from the WT 14028s bacterial strain was amplified using PCR with specific cloning primers listed in **Table 5.3**. The resulting amplicon was then inserted into the plasmid pQE60, a kind gift from Prof. Dipshikha Chakravorty, at the *Nco*I and *Hind*III restriction sites, positioning the gene under the control of the constitutive T5 promoter of pQE60 (**Fig. 5.1A**). After successful construction, the positive clones containing the *leuO* gene in pQE60 were transformed into the WT strains. These transformed strains were termed WT pQE60-L, representing the bacterial cells constitutively expressing the *leuO* gene from the pQE60 vector.

The T5 promoter of the WT-LeuO construct was substituted with the pBAD promoter, which was obtained by PCR amplification from the pKD46 plasmid, a kind gift from Prof. Dipshikha Chakravorty, using the specified cloning primers listed in **Table 5.3**. The resultant amplicon was ligated into the plasmid pQE60 at the *Eco*RI and *Xho*I restriction sites. Upon successful construction, the positive clones were transformed into the WT bacterial strains. These transformed strains were denoted as WT-pQE60-L-I, signifying bacterial cells in which the *leuO* gene is expressed under the control of the pBAD promoter, which can be induced by arabinose.

5.2.4 Quantitative real-time PCR

Overnight-grown WT 14028s bacterial cultures were sub-cultured in triplicates at a 1:100 ratio in LB medium containing ampicillin (100 µg/mL) and arabinose (10 mM and 40 mM) and incubated at 37 °C, continuous shaking at 120 rpm for 4 hours. The total RNA was then isolated from the bacterial cultures, and cDNA synthesis was performed using the abovementioned method. For gene expression analysis, quantitative real-time PCR (RT-qPCR) was conducted using PowerUp SYBR Green master mix (Thermo Fisher Scientific). The relative expression of the target gene was determined by the threshold cycle method ($2^{(-\Delta\Delta CT)}$) with normalisation to the reference gene *rpoD*. The specific primer sequences used in the RT-qPCR assay are provided in **Table 5.2**.

5.2.5 Plasmid loss assay

A. Generation of plasmid constructs for the assay

A protospacer sequence (5' AAGATCACGCGCTCCCACTTGAAGCCCTCGGGGAA 3')

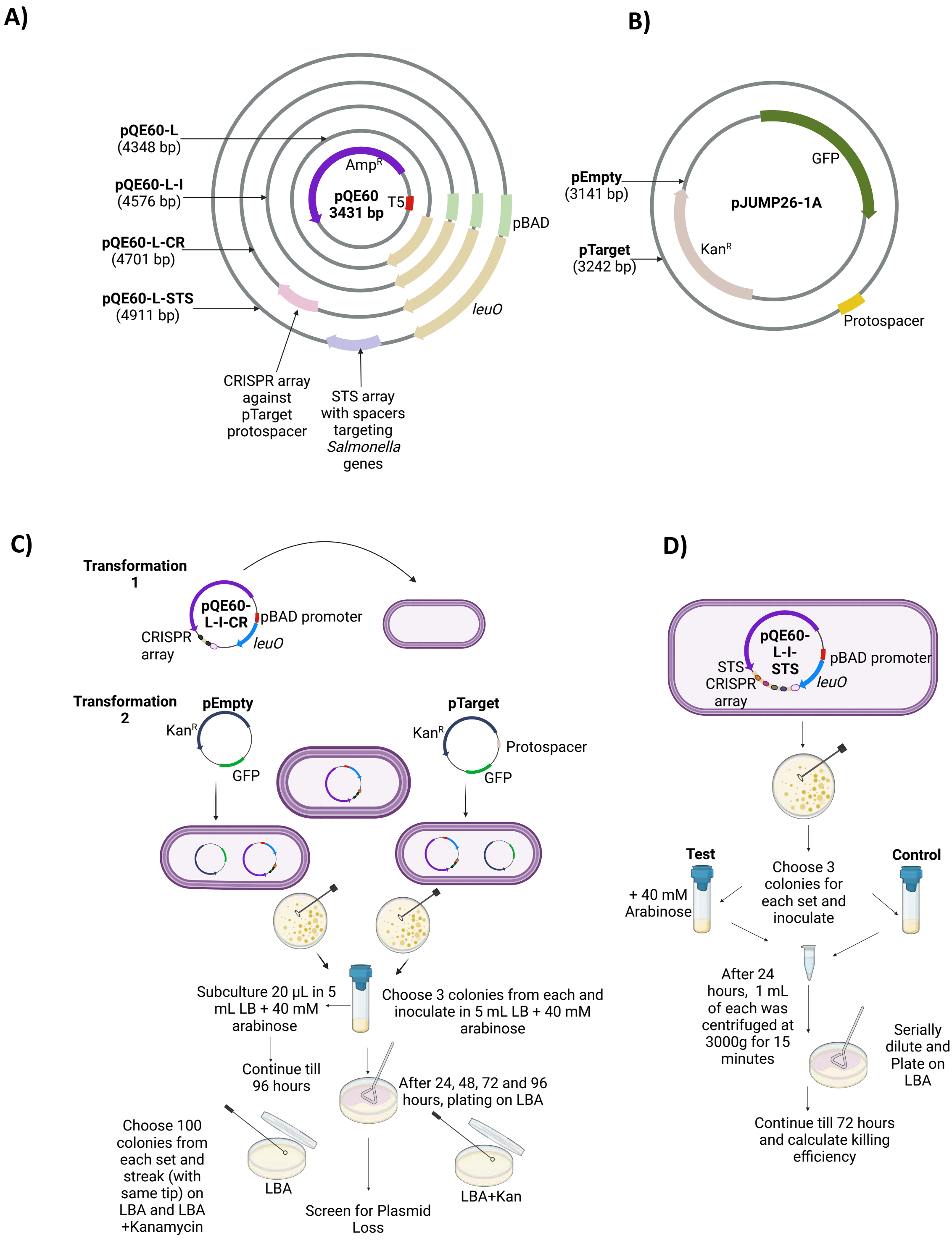

A)
pQE60-L
(4348 bp)
pQE60-L-I
(4576 bp)
pQE60-L-CR
(4701 bp)
pQE60-L-STS
(4911 bp)
AmpR
pQE60
3431 bp
T5
pBAD
leuO
CRISPR array against pTarget protospacer
STS array with spacers targeting Salmonella genes
B)
pEmpty
(3141 bp)
pTarget
(3242 bp)
pJUMP26-1A
GFP
KanR
Protospacer
C)
Transformation 1
pQE60-L-I-CR
pBAD promoter
CRISPR array
leuO
Transformation 2
pEmpty
KanR
GFP
pTarget
KanR
Protospacer
GFP
Subculture 20 µL in 5 mL LB + 40 mM arabinose
Choose 3 colonies from each and inoculate in 5 mL LB + 40 mM arabinose
Continue till 96 hours
After 24, 48, 72 and 96 hours, plating on LBA
Choose 100 colonies from each set and streak (with same tip) on LBA and LBA +Kanamycin
LBA
LBA+Kan
Screen for Plasmid Loss
D)
pQE60-L-I-STS
pBAD promoter
STS CRISPR array
leuO
Test
+ 40 mM Arabinose
Choose 3 colonies for each set and inoculate
Control
After 24 hours, 1 mL of each was centrifuged at 3000g for 15 minutes
Serially dilute and Plate on LBA
Continue till 72 hours and calculate killing efficiency

Figure 5.1 Graphical representation of the plasmid constructs and assays used in the study. A) pQE60-L is pQE60 containing *leuO,* pQE60-L-I is pQE60 containing *leuO* under the inducible pBAD promoter, pQE60-L-I-CR is pQE60-L-I with a constitutively expressed CRISPR array, and pQE60-L-I-STS is pQE60-L-I with a constitutively expressed self-targeting CRISPR array. The green, yellow, pink and light purle arrows represent the pBAD promoter, *leuO*, CRISPR array and STS array, respectively. **B)** Plasmid pTarget was obtained by cloning the protospacer into pJUMP26-1 A. **C)** The experimental procedure for the plasmid loss assay. **D)** The experimental procedure for the self-targeting assay in *Salmonella.*

from the pSW002-PpsbA-DsRed-Express2 (addgene) was amplified using PCR, with specific cloning primers listed in **Table 5.3**. The resulting amplicon was inserted into the plasmid pJUMP26-1A at the *Eco*RI and *Xba*I restriction sites. pJUMP26-1A contains a p15A ori, kanamycin resistance and constitutively expressed green fluorescent protein (GFP). The plasmid obtained was termed pTarget, while the empty pJUMP26-1A plasmid was named pEmpty (**Fig. 5.1B**).

A synthetic CRISPR array containing one spacer flanked by two direct repeats and a constitutive Anderson promoter was cloned in pQE60-L-I, using around-the-horn PCR cloning (primers listed in **Table 5.3**). The linear product was self-ligated using T4 DNA ligase (NEB) to obtain pQE60-L-I-CR (**Fig. 5.1A**) and transformed into the WT 14028s strain.

B. <u>Plasmid loss assay</u>

Plasmid loss assay was carried out in bacterial strains *S. enterica* subsp. *enterica* serovar Typhimurium str. 14028s, Typhi str. CT18, Paratyphi A and Welterveden (**Table 5.1**). Each bacterial strain was first transformed with pQE60-L-I-CR. Subsequently, these strains were transformed either with the pEmpty (control) or pTarget (test) plasmid (**Fig. 5.1C**). Three colonies from each set were selected and cultured in 5 mL of LB medium containing ampicillin (100 µg/mL) and arabinose (40 mM), which served as an inducer for LeuO that is expressed under the pBAD promoter. After 24 hours of incubation at 37 °C, 20 µL of the initial culture was sub-cultured in 5 mL of fresh LB medium with ampicillin and arabinose and incubated at 37 °C. The remaining culture was pelleted down, washed with 1 mL of MilliQ, and then serially diluted to a concentration of 10^{-4} to 10^{-6} cells per mL. Subsequently, 50 µL of this dilution was plated on Luria-Bertani agar (LBA). From these plates, 100 colonies were randomly picked and individually streaked on LBA and LBA supplemented with kanamycin (50 µg/mL). Based on the colony-forming units (CFU) obtained, plasmid loss was estimated using the formula [(CFU in Control - CFU in Test) / CFU in Control] * 100. This process was repeated over 96 hours to observe the plasmid loss in the bacterial population.

5.2.6 Targeted species-specific killing

A. <u>*In silico* selection of protospacer targets for testing self-targeting</u>

Based on the comprehensive literature survey and pangenome analysis outlined in

Chapter 1, we identified and selected four promising targets for self-killing. These targets were situated within the well-conserved genes responsible for pathogenicity and survival: *hilA, invA, ttrA,* and *sdiA.* The spacers were designed against these genes by selecting the regions within the genes that contain the PAM sequence AWG at its 5' end. These selected sequences (**Table 5.4**) would act as protospacers for the self-targeting spacers (STS).

B. Generation of plasmid constructs for the assay

A custom-designed CRISPR array with the spacers against the above-selected protospacers and a strong constitutive Anderson promoter was commercially synthesised (GeneArt Gene Synthesis, Thermo Fisher). This CRISPR array was cloned in the plasmid pQE60 containing the gene *leuO* under the pBAD promoter using Gibson Assembly by NEBuilder HiFi cloning kit (NEB). The primer sequences used for cloning are listed in **Table 5.3**. The construct was termed pQE60-L-I-STS and was transformed in the bacterial strain *S. enterica* subsp. *enterica* serovar Typhimurium str. 14028s, Typhi str. CT18, Paratyphi A and Welterveden.

C. Self-targeting assay

Three colonies of these pQE60-L-I-STS transformed *Salmonella* strains were selected and cultured in 5 mL of LB medium containing arabinose (40 mM), which served as an inducer for the LeuO expression. A control set, without arabinose induction, was used to provide a baseline. The bacterial cultures were incubated for 24 hours at 37 °C under shaking. 1 mL of each was pelleted down and washed with 1 mL of MilliQ water. Subsequently, the pellet was diluted to achieve cell concentrations ranging from 10^{-4} to 10^{-6} per mL. 50 µL of each dilution was plated onto LBA plates, and the LBA plates were incubated at 37 °C for colony growth. Following incubation, the colonies that developed on the LBA plates were counted to determine the CFU for each dilution. This process was repeated over 72 hours (**Fig. 5.1D**). Furthermore, 50 µL of each dilution was plated onto LBA supplemented with ampicillin (100 µg/mL) to check for plasmid curing. The percentage of surviving cells was calculated by the formula [(CFU in Control - CFU in Test) / CFU in Control] * 100. This quantification of surviving cells allowed the plotting of the killing efficiency.

5.3 Results

5.3.1 Generation of plasmid constructs

The plasmid constructs for activation of the CRISPR-Cas system, plasmid loss assay and self-targeting assay were generated using the protocols mentioned in section 5.2. We then confirmed the clones by restriction digestion of the potential clones and observed for respective insert release (**Fig. 5.2**). These recombinant plasmids were transformed into the *Salmonella* strains and used for further assays. The following terminologies will be used henceforth- (i) pQE60-L: pQE60 containing *leuO,* (ii) pQE60-L-I: pQE60 containing *leuO* under the inducible pBAD promoter, (iii) pQE60-L-I-CR: pQE60-L-I containing a constitutively expressed CRISPR array with spacer against pTarget, (iv) pQE60-L-I-STS: is pQE60-L-I containing a constitutively expressed self-targeting CRISPR array, (v) pEmpty: empty pJUMP26-1A and (vi) pTarget: pJUMP26-1A with cloned protospacer.

5.3.2 Inspecting the activation of the CRISPR-Cas system under various conditions

To inspect the activation of the CRISPR-Cas system in *S. enterica* subsp. *enterica* serovar Typhimurium str. 14028s, we checked the expression of 8 *cas* genes *cas2*, *cas1*, *cas6*, *cas5, cas7, cse2, cse1,* and *cas3* under various conditions: (A) Nutrient-rich media - Cultures at OD_{600} values of 0.3, 0.6, 1 and 1.5; (B) F-Media, mimicking the intravacuolar conditions during intracellular growth of *Salmonella* - Cultures at OD_{600} values of 0.3, 0.6, and 1; (C) Envelope stress with EDTA at concentrations 0.5 mM, 1 mM, 2.5 mM and 5 mM; and (D) Biofilm - ring and pellicle biofilms were collected at 24, 48 and 96 hours. The RNA was isolated from these cultures and processed for semi-quantitative RT-PCR. We did not see any visible amplification of the *cas* genes under the conditions tested, indicating the absence/undetectable *cas* gene expression. Representative gels for the amplification of *cas* genes under nutrient-rich media at OD_{600} value 0.6 are depicted in **Fig. 5.3**.

5.3.3 Inducing the activation of the CRISPR-Cas system using LeuO

As we did not detect any expression of *cas* genes under different conditions tested, we resorted to inducing its expression using its transcriptional regulators. Various transcription factors regulate the CRISPR-Cas system in *Salmonella* (Medina-Aparicio *et al.,* 2011, Kushwaha *et al.,* 2022). The major ones are the LeuO and H-NS. H-NS is the negative regulator of the system, silencing the CRISPR-Cas activity, while LeuO regulates the system

Table 5.2 List of expression primers used in this study

No.	Name	Sequence (5'- 3')
1	*cas1*- Forward	ATGATGACCTGCGGCTGA
2	*cas1*- Reverse	TTCACGCCATACTGCTTCG
3	*cas2*- Forward	TCTTGCCGTCTGGTTACTCG
4	*cas2*- Reverse	CGTCTGTTTTCACCCCAGGT
5	*cas3*- Forward	AACATGCCGGTTGGATTTGC
6	*cas3*- Reverse	CCACAGCGTGACAGACTCTT
7	*cas5*- Forward	GATTTCCCGACGCGACTACT
8	*cas5*- Reverse	ACTTTTGCGCCCCAGATACA
9	*cas6*- Forward	GCGTCACGATTTGCTGATGG
10	*cas6*- Reverse	TCATCTGCCGATCTTTCCC
11	*cas7*- Forward	GCCGGATGTTAGCGAAGAA
12	*cas7*- Reverse	CCTGCATCTTCTGCCGAT
13	*cse1*- Forward	TACCAGACCAGTGTGATGC
14	*cse1*- Reverse	CTGTAAGGTGGCAAAATCCA
15	*cse2*- Forward	TGATGCCTGTTTGGCTGAGG
16	*cse2*- Reverse	TGTCGCCACCTTTCTTCTGT
17	*hns*- Forward	ACATCCGTACTCTTCGTG
18	*hns*- Reverse	ACGAGTGCGTTCTTCCAC
19	*leuO*- Forward	AGCATCAGTTACGCTATCAGG
20	*leuO*-Reverse	AACATCGCCTTCCAGTAGC
21	*rpoD*- Forward	GATAAGACGAACGGTGAGG
22	*rpoD*- Reverse	AGCCTCTGTCAAATCAGC

Table 5.3 List of cloning primers used in this study

No.	Name	Sequence (5'- 3')	Role
1	*leuO* - Forward	GACTCCATGGATGCCAGAGGTCAAAACC	cloning *leuO* in pQE60
2	*leuO* - Reverse	GACTAAGCTTCGGTTTTATCGCTTACAAAC	
3	pBAD - Forward	ATGCCTCGAGACTCCCGCCATTCAGAG	cloning *leuO* under inducible promoter pBAD in pQE60
4	pBAD - Reverse	ATGCGAATTCAACGGGTATGGAGAAACAGT	
5	Array - Forward	ATCACGCGCTCCCACTTGAAGCCCTCGGGGAAATGTTCCCCGCGCCAGCGGGGATAAACACGGTTATCCACAGAATCAGG	cloning artificial CRISPR array in pQE60
6	Array - Reverse	CGGTTTATCCCCGCTGGCGCGGGGAACACGCTAGCACTGTACCTAGGACTGAGCTAGCCGTCAAGTATTACCGCCTTTGAGTG	
7	Protospacer - Forward	GATCGAATTCAAGATCACGCGCTCCCACTTG	cloning protospacer in pJUMP26-1A
8	Protospacer - Reverse	GATCTCTAGATCCAAGGTGTACGTGAAGCA	
9	STS - Forward	TACCTAGGACTGAGCTAGCCGTCAACGTCATCACCGAAACG	cloning self-targeting CRISPR array in pQE60
10	STS - Reverse	GTAGGACTGCTCAGTTCAAACATGATCGTGAAAACCTCTGACACAT	

Table 5.4 Individual breakdown of the sequence of the STS CRISPR array

Gene	Sequence (5'- 3')
Anderson Promoter	TTGACGGCTAGCTCAGTCCTAGGTACAGTGCTAGC
hilA	GCGCAAATGGGGATTTTTGATAAACAAAACGC
invA	CGAAATTTCCTGATTTACTTAAAGAAGTGCTC
ttrA	TCTGGGATATGACGTAAAATGCTGGACGCAGG
sdiA	AAGCGCAGGCGATGTGGGATGCCGCCCAGCGT
Direct repeats 1-4	GTGTTCCCCGCGCCAGCGGGGATAAACCG
Direct repeat 5	ATGTTCCCCGCGCCAGCGGGGATAAACAC

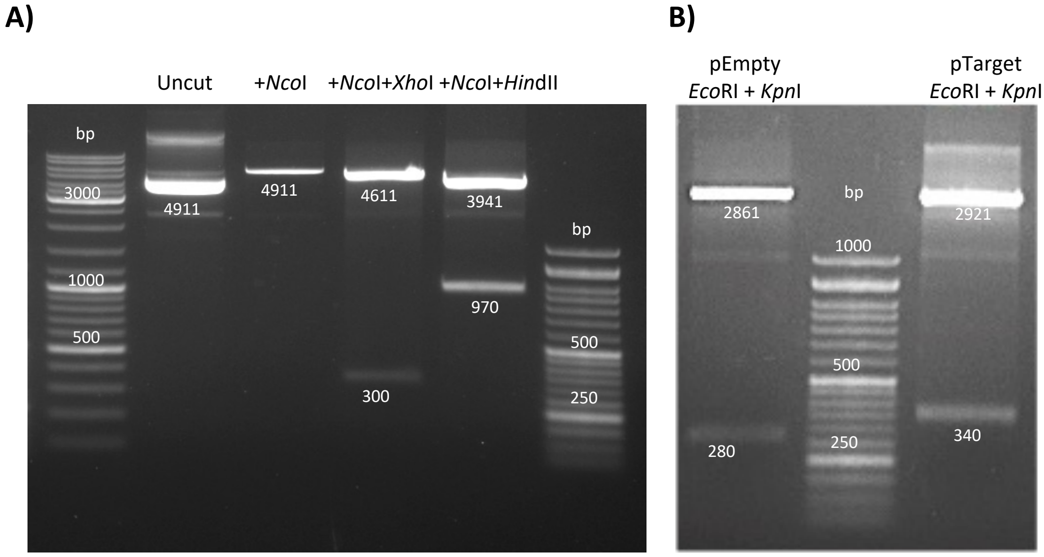


Figure 5.2 Clone verification by restriction digestion and PCR. A) pBAD promoter (lane 4, 300 bp) and *leuO* (lane 5, 970 bp) cloned in pQE60 verified by restriction digestion of the potential clones. **B)** Restriction digestion of pEmpty (lane 1, 280 bp plasmid fragment) and the potential clone (lane 3, 340: 60 + 280 bp) confirms the protospacer (60 bp) cloning.

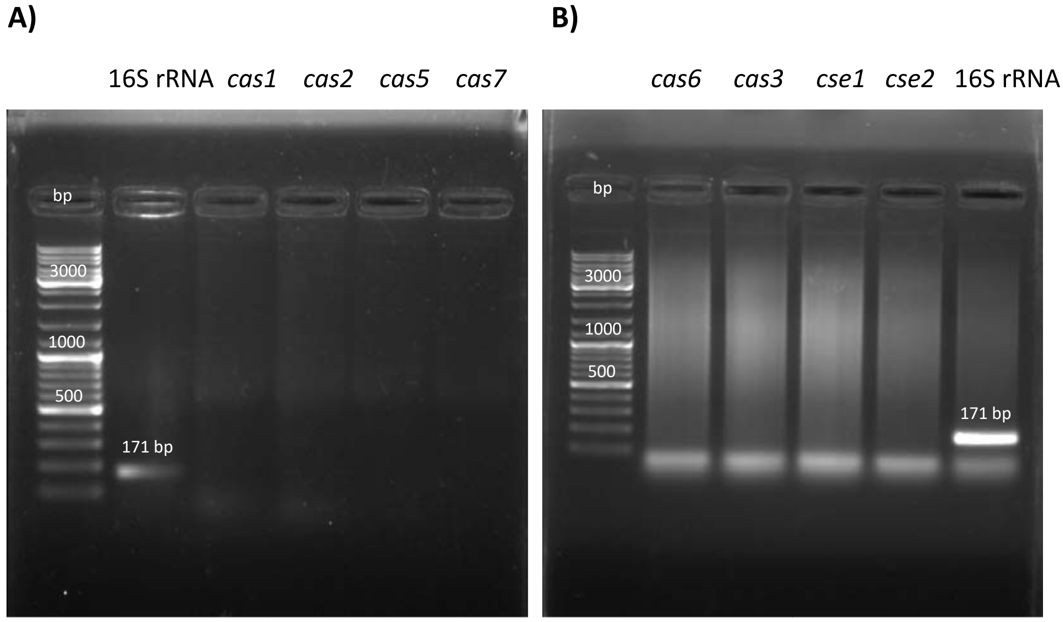


Figure 5.3 Expression of *cas* genes in log phase. A) Expression of the genes *cas1, cas2, cas5, cas7* and 16S rRNA. **B)** Expression of the genes *cas6, cas3, cse1, cse2* and 16S rRNA from the *S.* Typhimurium str. 14028s. Detectable bands were obtained for the 16S rRNA gene and not for *cas* genes.

positively. To achieve a robust active CRISPR-Cas system, we overexpressed LeuO in trans using pQE60- L-I plasmid.

The WT 14028s transformed with pQE60-L-I was cultivated in a nutrient-rich medium supplemented with 10 mM and 40 mM arabinose. Arabinose was used at two different concentrations to titre the expression level of LeuO. Given that the CRISPR-Cas system exhibited negligible expression in nutrient-rich media, the 10 mM arabinose served as a reference point, resulting in minimal *cas* gene expression. Utilising 40 mM arabinose would ensure a strong and robust expression of the *cas* genes. The RNA was isolated from these cultures. The expression of the *cas* genes and its regulators *leuO* and *hns* was evaluated using RT-qPCR. When induced with 40 mM arabinose, we see a >30-fold increase in the expression of *leuO*, with no discernible alteration detected in *hns* expression. In accordance, the *cas* genes show an increase in expression as follows- *cas7* by 10-fold, *cas3* by 7.5-fold, *cas6* by 6.5-fold, *cas1* by 5.7-fold, *cas2* by 3.6-fold, *cse2* by 3.1-fold, *cas5* and *cse1* by 3.1-fold (**Fig. 5.4A**).

5.3.4 Validating the functional activation of the CRISPR-Cas system

We executed the plasmid loss assay to evaluate the functional efficacy of the CRISPR-Cas system as an adaptive immune mechanism. The test condition included the *S. enterica* strains with pTarget and pQE60-L-I-CR, while the control included strains with pEmpty and pQE60-L-I-CR. The pTarget contains the protospacer for the spacer in pQE60-L-I-CR. Thus, the functionally active CRISPR-Cas system would degrade the pTarget, making the bacterial cell kanamycin susceptible without any green fluorescence. However, in the control condition, such a mechanism would not be operational (**Fig. 5.4B**).

We observed a remarkable 93% loss of pTarget plasmid after just 24 hours of incubation, which further increased to ~95% reduction by 96 hours. As expected, no significant plasmid loss was observed for the control set containing pEmpty without any protospacer (**Fig. 5.4B**). To enhance the robustness of our findings, we extended the plasmid loss assay to *Salmonella* serovars belonging to typhoidal (Typhi str. CT18, Paratyphi A) and non-typhoidal (Welterveden) groups. A consistent and noteworthy loss of over 95% in pTarget plasmids was observed across all these serovars (**Fig. 5.4B**), validating the functional activation of the CRISPR-Cas system in different *S. enterica* serovars.

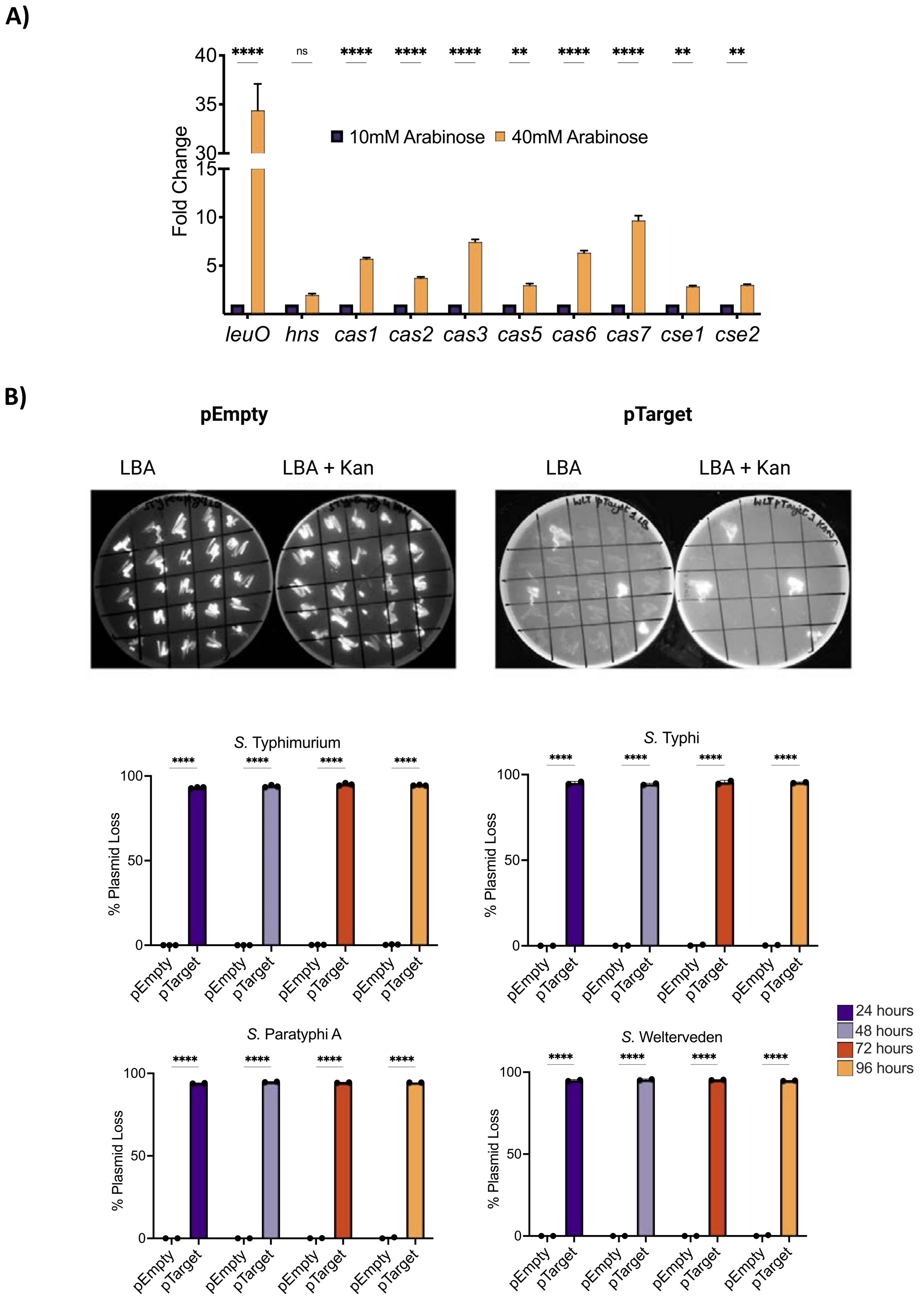


Figure 5.4 Activation of the CRISPR-Cas system. A) Induction of *cas* genes in log phase with overexpressed *leuO*. The relative expression of the target gene was determined by qRT-PCR using the threshold cycle method ($2^{(-\Delta\Delta CT)}$) with normalisation to the reference gene *rpoD*. Statistical significance: **P≤ 0.01, ****P≤ 0.0001, ns = not significant. **B)** Plasmid loss assay. Representative images of colonies obtained using the plasmid loss assay from pEmpty and pTarget. The percentage loss of plasmids was determined post-induction of the CRISPR-Cas system using the formula: [(CFU in Control - CFU in Test) / CFU in Control] * 100. Statistical significance: ****P≤ 0.0001.

5.3.5 Pangenome analysis to choose protospacer targets for self-targeting

Existing literature (Guo *et al.,* 2000; Malorny *et al.,* 2004; Halatsi *et al.,* 2006; Siala *et al.,* 2017) has highlighted key genes like *hilA*, *invA*, *ttrA*, and *sdiA* useful in identifying *Salmonella* strains. Through pangenome analysis from **Chapter 2**, we checked the conservation of the genes and selected four *Salmonella*-specific and persistent genes — *hilA*, *invA*, *ttrA*, and *sdiA*—as target protospacers for self-killing.

Employing a manual approach, we meticulously scanned the sequences of these chosen genes to locate the PAM sequence, AWG. Subsequently, 32 bp segments beyond these identified PAM regions were chosen as the protospacers and used to obtain spacer sequences for the STS CRISPR array. To ascertain the prevalence of the spacer sequences, a nucleotide BLAST analysis was executed with a criterion of 100% sequence identity over a 35 bp stretch encompassing the spacer region and PAM across the 12.2K strain dataset. This analysis revealed a remarkable conservation rate of ~99% among the strains (**Fig. 5.5A**), affirming the suitability of these spacer sequences for further experimentation.

5.3.6 Validating the self-targeting for species-specific killing

We executed the self-targeting assay to check the use of the CRISPR-Cas system in *Salmonella*-specific killing. This procedure involves introducing pQE60-L-I-STS that contains STS spacers. Assuming the self-targeting mechanism works, the CRISPR-Cas system would target the selected protospacer in the genome, cleaving it, thereby leading to cell death. The assay results revealed 35±2.0% killing after 24 hours, which dropped to 17±6.5% by 72 hours (**Fig. 5.5B**).

We then checked for self-targeting in other serovars. Serovar Typhi str. CT18 showed a 19±3.8% death, Paratyphi A 22±2.0%, and Welterveden 10±1.5% 72 hours post-induction of the CRISPR-Cas system (**Fig. 5.5B**). Consequently, 50 µL of each dilution was plated onto LBA supplemented with ampicillin to assess plasmid curing but no significant difference in CFU was observed between the induced and uninduced conditions.

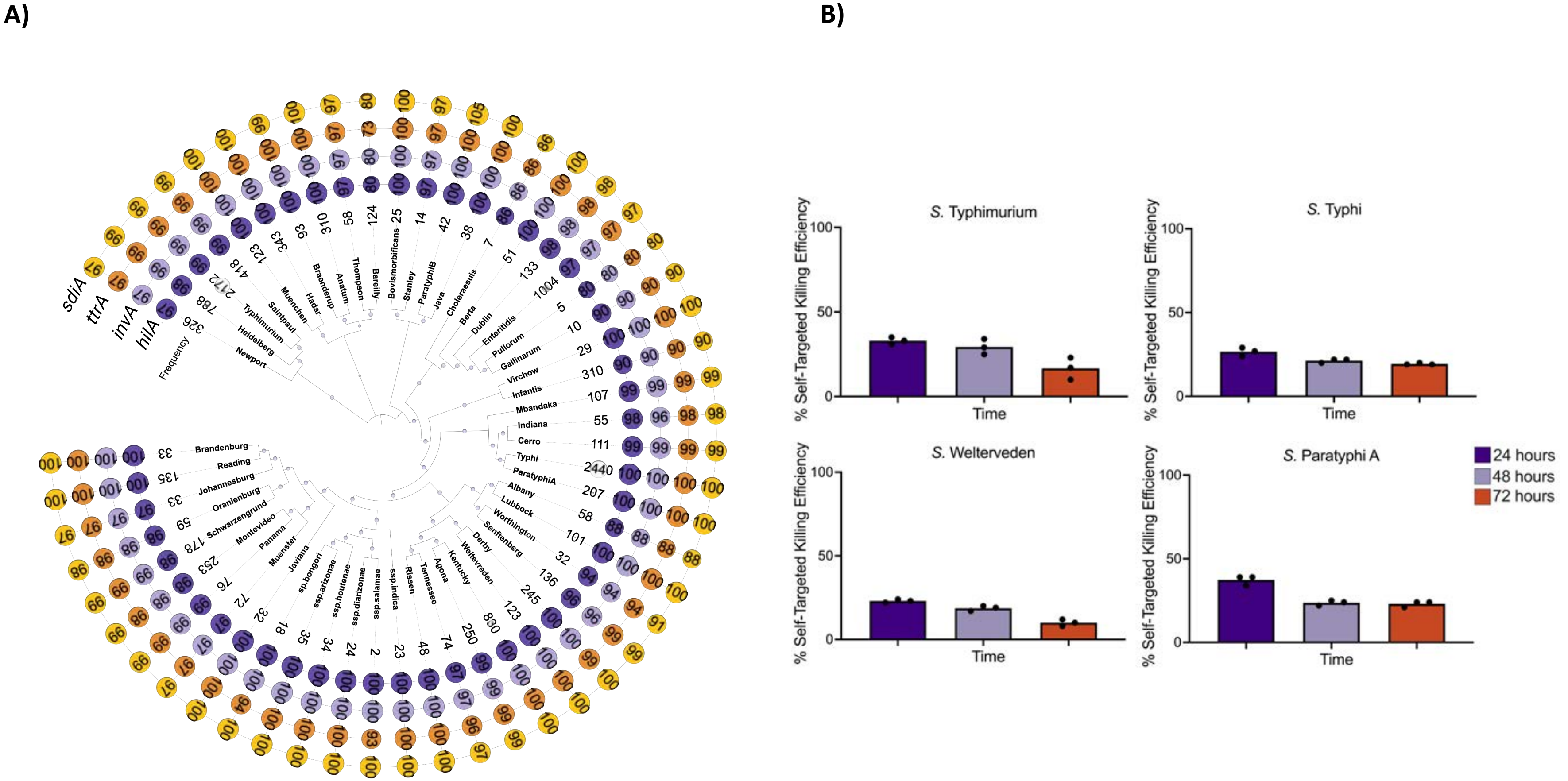


Figure 5.5 Self-targeting for species-specific killing. A) Pangenome conservation and frequency calculation of *hilA*, *invA*, *ttrA*, and *sdiA* genes of *Salmonella*. The frequency represents the strain count within a given *Salmonella* serovar. The size of the circles on the phylogenetic tree indicates the bootstrap values. The circle's values denote the percentage strain count containing the gene. **B)** Self-targeted killing assay. Percentage of self-targeted killing efficiency in *S.* Typhimurium, *S.* Typhi, *S.* Paratyphi A and *S.* Welterveden. The percentage killing was calculated using the formula: [(CFU in Control - CFU in Test) / CFU in Control] * 100.

5.4 Discussion

In this study, our primary objective was to elucidate the conditions under which the CRISPR-Cas system is active and explore its potential applications for species-specific killing of *Salmonella*. Many factors, including temperature, nutrient availability, and stress conditions, influence the expression of the CRISPR-Cas components, as seen in various species of *Enterobacteriaceae* (Fang *et al.,* 2000, Eswarappa *et al.,* 2009, Pul *et al.,* 2010, Louwen *et al.,* 2014, Sun *et al.,* 2020, Wu *et al.,* 2022, Zakrzewska and Burmistrz 2023). In our study, we subjected *Salmonella* to various growth conditions like nutrient-rich and nutrient-depleted media, envelope stress, and biofilm growth to explore the CRISPR-Cas activation. Intriguingly, we consistently observed an undetectable expression of *cas* genes. This resonates with the previous studies in *E. coli,* showing no detectable *cas* expression under lab conditions (Paul *et al.,* 2010). Nevertheless, recent studies on the *Salmonella* CRISPR-Cas system (by knocking out various components of the system) highlight its roles in regulating bacterial physiology, virulence, and biofilm (Cui *et al.,* 2020; Stringer *et al*., 2020; Medina-Aparicio *et al.,* 2021, Sharma *et al.,* 2022). This warrants further investigation into the underlying factors influencing the CRISPR-Cas functional dynamics.

A study by Westra *et al.,* 2010, reveals the significance of H-NS and LeuO in CRISPR-Cas regulation in *E. coli*. Our investigation expanded upon this research by examining the roles of LeuO and H-NS regulators in activating the CRISPR-Cas system in *Salmonella*. We carried out the experiment by overexpressing LeuO in trans to activate the system. The plasmid loss assay in various *Salmonella* serovars demonstrated its functional activation, exhibiting ~95% efficacy. This highlights the potential of LeuO to positively influence the expression of the CRISPR-Cas system in *Salmonella*. Westra *et al.,* also commented on the mechanism of action in *E. coli*. They showed that H-NS molecules can spread along the DNA, binding to adjacent regions and forming a larger complex. This spreading can lead to the repression of gene transcription over a wider region of DNA than just the initial binding site (Westra *et al*., 2010). As the H-NS expression remains unaltered upon LeuO overexpression, we hypothesised that in *Salmonella*, the activation of the CRISPR-Cas transcription by LeuO is either by (i) pushing aside H-NS, which is already bound to the promoter or (ii) overexpressed LeuO curbs the H-NS from effectively constructing the expanded complexes along the DNA. This leads to the disruption of the H-NS complex, facilitating the accessibility of RNA polymerase to the *cas* promoter.

We intended to harness this potential for species-specific killing by optimising the utilisation of our plasmid construct, overexpressing LeuO and functionally activating the endogenous CRISPR-Cas system. Through pangenome analysis, we strategically selected protospacer targets well-conserved throughout *Salmonella* for implementing a self-targeting mechanism. Four genes—*hilA*, *invA*, *ttrA*, and *sdiA* were identified as protospacer candidates for the STS CRISPR array. Subsequently, by employing a self-targeting assay, we explored the system's potential for species-specific killing. However, the self-targeting results were not as anticipated, leading to <35% self-killing, indicating a non-significant effect. The experiment resulted in bacterial colonies that could have escaped genome targeting. However, no curing of pQE60-L-I-STS was observed. The generation of escaped colonies (Gomaa *et al.,* 2014, Hamilton *et al.,* 2019) can be by multiple factors. The bacteria utilise DNA repair pathways, developing mutations that resist the intended modification, resulting in escape mutants. When bacterial cells undergo genome editing, they often activate stress responses, triggering protective mechanisms and preventing modifications. The use of the CRISPR-Cas editing system can create a form of selective pressure, favouring resistant cells over edited ones. Also, unsuccessful crRNA targeting can allow bacteria to evade unintended modification, yielding off-target escape mutants.

In conclusion, our study sheds light on the potential role of LeuO in positively modulating the CRISPR-Cas system in *Salmonella* while hinting at the complex factors influencing its expression. However, the challenges in achieving species-specific killing through self-targeting using the endogenous CRISPR-Cas system in *Salmonella* are significant and call for further research and improvement. One approach is to improve and design more specific and effective protospacer sequences less prone to mutations. Increasing the number of spacers to target a broad spectrum of essential genomic locations might potentially enhance self-killing efficiency. Improving CRISPR array design and delivery methods (e.g., using bacterial conjugation) could increase the precision and efficiency of the CRISPR-Cas system. Lastly, combining the endogenous CRISPR-Cas system with other genome editing techniques, such as phage therapy, could also provide a more effective species-specific killing in *Salmonella*.

Chapter 6

Conclusion and Future Prospects

6.1 Conclusion

Salmonella exhibits complex evolutionary patterns encompassing over 2600 serovars with diverse pathogenic profiles (Tanner & Kingsley, 2018). This diversity of *Salmonella* lineages is influenced by horizontal gene transfer events of mobile genetic elements (MGEs) and pathogenicity islands, making their genome flexible (Ferreira, Buckner, & Finlay, 2012). Hence, understanding the intricacies of *Salmonella*'s genomic plasticity is crucial for elucidating its pathogenesis and devising effective strategies for its control.

Within a species, the core genome, comprising genes universally present, handles essential cellular functions. In contrast, the flexible genome consists of genes that vary between strains, allowing bacteria to adapt to specific environments and acquire pathogenic traits. Comprehensive pangenome analysis of 12K *Salmonella* strains in this thesis has revealed the roles played by dynamic genome segments known as regions of genome plasticity (RGPs) in shaping its evolution. We observed a purposeful and non-random integration pattern of pathogenicity-related gene clusters into specific RGPs. Most RGPs were preferably located at strategic locations (spots) to gain potential benefits of co-regulation, leading to functional synergy among genes. These benefits are provided by the persistent border genes. For instance, inserting RGP with metal resistance genes near stress resistance genes. This arrangement allows them to share a regulatory network, making it more efficient to respond to stressors. Furthermore, the type I-E CRISPR-Cas system, an adaptive immune mechanism, is highly conserved and prevalent in spot #22.

Reports from our laboratory and other relevant studies underscore the significance of *Salmonella*'s CRISPR-Cas system in non-canonical functions, particularly in biofilm formation and virulence (Cui *et al.*, 2020; Medina-Aparicio *et al.*, 2021; Sharma, Das, Raja, & Marathe, 2022). To understand the system's role in such processes, we investigated the nuances of this system, studying its evolution and roles in endogenous gene regulation. Our analysis of 22 *Salmonella* serovars validated a preliminary study with four *Salmonella* serovars suggesting two varieties of the CRISPR-Cas system within this genus. Phylogenomic analysis categorised the strains of these serovars into two predominant clades, CRISPR-STM/*cas*-STM and CRISPR-STY/*cas*-STY, possessing mainly the non-typhoidal serovars and typhoidal serovars, respectively. We also observed the conservation of CRISPR spacers within the serovars. The CRISPR arrays of the broad-host-range serovars (e.g., Typhimurium) were longer than the host-specific serovars (e.g., Typhi). Further, we could not map protospacers onto MGEs for a significant fraction of spacers. Thus, to gain a profound understanding, we thoroughly analysed the CRISPR-Cas systems and their spacer targets across the 12K *Salmonella* strains, representing 52 distinct serovars.

We identified 7,624 unique spacers, with only 4.8% (365 spacers) displaying protospacers within reported plasmids and 0.6% (43 spacers) within phages. We explored the potential CRISPR spacer targets in the genomes of *S.* Enteritidis, *S.* Typhimurium, and *S.* Typhi belonging to CRISPR-STM (Enteritidis and Typhimurium) and CRISPR-STY (Typhi) categories. All highly conserved spacers potentially exhibit regulatory functions. Such spacers were named potential regulatory spacers (PRS). Some of the targets of PRS, like *recA* and *ruvB* in Enteritidis and Typhi, respectively, reportedly exhibit genetic interactions with *cas1* in *E. coli*. We also identified *bcsC, entE, leuO, mrdA, pgi, ratB,* and *rep* as PRS targets. The congruence of this result with available reports demonstrating CRISPR-Cas mediated regulation of binding of the Cas5 to these targets (Stringer, Baniulyte, Lasek-Nesselquist, Seed, & Wade, 2020; Sharma, Das, Raja, & Marathe, 2022) adds credence to our data. Consequently, we posit that the CRISPR-Cas system potentially regulates the gene targets identified in our research but awaits experimental validation.

The exogenous type II CRISPR-Cas system containing Cas9 has been used for targeted *Salmonella* elimination (Hamilton *et al.*, 2019). However, it posed some challenges, including the toxicity of overexpressed Cas9 and the transfer of bigger-sized plasmid into *Salmonella*. Our strategy of exploiting an endogenous system bypassed these challenges. Subsequently, for clinical feasibility of the strategy, the delivery mechanism of the CRISPR-Cas system can be optimised by exploring phage-mediated delivery systems that could act synergistically (phage therapy and CRISPR-Cas mediated self-killing) to improve the killing efficacy. However, the observed efficacy of self-targeting was below anticipated levels, with less than 35% species-specific killing. This outcome prompts a necessity for further refinement of our strategy.

6.2 Future scope

The present study on *Salmonella*'s genomic plasticity and the intricate dynamics of its CRISPR-Cas system opens up avenues for significant future research. One can experimentally validate (i) the role of conserved flanking genes to support the purposeful integration of RGPs at a given spot and (ii) if the regulation and function of RGPs and flanking genes are coordinated. This integrated analysis is vital for deciphering the intricate regulatory networks governing RGPs and providing valuable insights into *Salmonella*'s adaptability and pathogenicity.

The phylogenomic study of the CRISPR-Cas system suggests that the CRISPR array of *Salmonella* is uniquely tailored to each serovar. Further, the variations in the CRISPR spacers and *cas* genes may manifest a competitive advantage to the bacteria under plightful situations like antimicrobial stress. Exploring such cases could illuminate the influence of environmental factors like antibiotics and host defences on the evolution of the CRISPR-Cas system. Furthermore, the variability in CRISPR spacers holds promise for revolutionising serotyping methodologies.

Further, the functional roles of PRS and PAMs identified within the CRISPR-Cas system in *Salmonella* can be experimentally verified. Understanding the functional roles of PRS may provide insights into CRISPR's regulatory mechanisms. One can discern whether the CRISPR-Cas system predominantly targets DNA or RNA using DNA footprinting/chromatin immunoprecipitation assays and RNA degradation/ electrophoretic mobility shift assays, respectively.

A multi-step approach can be implemented to optimise the efficiency of the CRISPR-Cas system to kill *Salmonella*. Firstly, the selection and quantity of the self-targeting spacers may be refined to ensure specificity and efficiency, resulting in a lethal outcome. Subsequently, the delivery mechanism of the CRISPR-Cas system can be optimised by exploring phage-mediated delivery systems that could act synergistically to improve the killing efficacy. This presents a transformative solution for clinical and environmental applications combating *Salmonella* infections.

In conclusion, our study on *Salmonella*'s genomic plasticity and the CRISPR-Cas system provides key insights into its adaptability and pathogenicity. However, our attempt at repurposing the CRISPR-Cas system for targeted killing showed lower efficacy, prompting the need for refinement. This work holds promise for reducing the global burden of *Salmonella* infections, offering a valuable contribution to the ongoing efforts to fight against bacterial pathogens.

REFERENCES

Abby, S., & Daubin, V. (2007). Comparative genomics and the evolution of prokaryotes. *Trends in Microbiology*, 15(3), 135-141.

Aklujkar, M., & Lovley, D. R. (2010). Interference with histidyl-tRNA synthetase by a CRISPR spacer sequence as a factor in the evolution of *Pelobacter carbinolicus. BMC Evol Biol*, 10, 230.

Althouse, C., Patterson, S., Fedorka-Cray, P., & Isaacson, R. E. (2003). Type 1 fimbriae *of Salmonella enterica* serovar Typhimurium bind to enterocytes and contribute to colonization of swine in vivo. *Infect Immun*, 71(11), 6446-6452.

An, S. Q., Lu, G. T., Su, H. Z., Li, R. F., He, Y. Q., Jiang, B. L., *et al*. (2011). Systematic mutagenesis of all predicted *gntR* genes in *Xanthomonas campestris* pv. *campestris* reveals a GntR family transcriptional regulator controlling hypersensitive response and virulence. *Mol Plant Microbe Interact*, 24(9), 1027-1039.

Anderson, C. J., & Kendall, M. M. (2017). *Salmonella enterica* serovar Typhimurium Strategies for Host Adaptation*. Front Microbiol*, 8, 1983.

Andino, A., & Hanning, I. (2015). *Salmonella enterica*: survival, colonization, and virulence differences among serovars. *ScientificWorldJournal*, 2015, 520179.

Antony, L., M. Behr, D. Sockett, D. Miskimins, N. Aulik, J. Christopher-Hennings, E. Nelson, M. W. Allard & J. Scaria (2018). Genome divergence and increased virulence of outbreak associated *Salmonella enterica* subspecies *enterica* serovar Heidelberg*. Gut Pathog*, 10, 53.

Argimón, S., Abudahab, K., Goater, R. J. E., Fedosejev, A., Bhai, J., Glasner, C., *et al*. (2016). Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb Genom*, 2(11), e000093.

Ashkenazi, S., Cleary, T. G., Murray, B. E., Wanger, A., & Pickering, L. K. (1988). Quantitative analysis and partial characterization of cytotoxin production by *Salmonella* strains. *Infect Immun*, 56(12), 3089-3094.

Babu, M., N. Beloglazova, R. Flick, C. Graham, T. Skarina, B. Nocek, A. Gagarinova, O. Pogoutse, G. Brown, A. Binkowski, S. Phanse, A. Joachimiak, E. V. Koonin, A. Savchenko, A. Emili, J. Greenblatt, A. M. Edwards & A. F. Yakunin (2011). A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. *Mol Microbiol*, 79, 484-502.

Bang, I. S., Frye, J. G., McClelland, M., Velayudhan, J., & Fang, F. C. (2005). Alternative sigma factor interactions in *Salmonella*: σ^E and σ^H promote antioxidant defences by enhancing σ^S levels. *Mol Microbiol*, 56(3), 811-823.

Barrangou, R., & van der Oost, J. (2015). Bacteriophage exclusion, a new defense system. *EMBO J*, 34(2), 134-135.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., *et al*. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, 315(5819), 1709-1712.

Bäumler, A. J. (1997). The record of horizontal gene transfer in *Salmonella*. *Trends Microbiol*, 5(8), 318-322.

Bäumler, A. J., Tsolis, R. M., Bowe, F. A., Kusters, J. G., Hoffmann, S., & Heffron, F. (1996). The *pef* fimbrial operon of *Salmonella* Typhimurium mediates adhesion to murine small intestine and is necessary for fluid accumulation in the infant mouse. *Infection and Immunity*, 64(1), 61-68.

Bazin, A., Gautreau, G., Médigue, C., Vallenet, D., & Calteau, A. (2020). panRGP: a pangenome-based method to predict genomic islands and explore their diversity. *Bioinformatics*, 36(Supplement_2), i651-i658.

Beceiro, A., Tomás, M., & Bou, G. (2013). Antimicrobial resistance and virulence: a successful or deleterious association in the bacterial world? *Clin Microbiol Rev*, 26(2), 185-230.

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

Bernick, D. L., Cox, C. L., Dennis, P. P., & Lowe, T. M. (2012). Comparative genomic and transcriptional analyses of CRISPR systems across the genus *Pyrobaculum. Front Microbiol*, 3, 251.

Bintsis, T. (2017). Foodborne pathogens. *AIMS Microbiol*, 3, 529-563.

Biswas, A., Gagnon, J. N., Brouns, S. J., Fineran, P. C., & Brown, C. M. (2013). CRISPRTarget: bioinformatic prediction and analysis of crRNA targets. *RNA Biol*, 10(5), 817-827.

Bozic, B., J. Repac & M. Djordjevic (2019). Endogenous Gene Regulation as a Predicted Main Function of Type I-E CRISPR/Cas System in *E. coli*. *Molecules*, 24.

Brandwagt, D., C. van den Wijngaard, A. D. Tulen, A. C. Mulder, A. Hofhuis, R. Jacobs, M. Heck, A. Verbruggen, H. van den Kerkhof, I. Slegers-Fitz-James, L. Mughini-Gras & E. Franz (2018). Outbreak of *Salmonella* Bovismorbificans associated with the consumption of uncooked ham products, the Netherlands, 2016 to 2017. *Euro Surveill*, 23.

Britto, C. D., Wong, V. K., Dougan, G., & Pollard, A. J. (2018). A systematic review of antimicrobial resistance in *Salmonella enterica* serovar Typhi, the etiological agent of typhoid. *PLoS Negl Trop Dis*, 12(10), e0006779.

Brooks, A. N., Turkarslan, S., Beer, K. D., Lo, F. Y., & Baliga, N. S. (2011). Adaptation of cells to new environments. *Wiley Interdiscip Rev Syst Biol Med*, 3(5), 544-561.

Brouns, S. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J., Snijders, A. P., *et al*. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, 321(5891), 960-964.

Bruzzese, E., A. Giannattasio & A. Guarino (2018). Antibiotic treatment of acute gastroenteritis in children. *F1000Res*, 7, 193.

Buckner, M. M., Croxen, M. A., Arena, E. T., & Finlay, B. B. (2011). A comprehensive study of the contribution of *Salmonella enterica* serovar Typhimurium SPI2 effectors to bacterial colonization, survival, and replication in typhoid fever, macrophage, and epithelial cell infection models. *Virulence*, 2(3), 208-216.

Buyse, J., Swennen, Q., Niewold, T. A., Klasing, K. C., Janssens, G. P., Baumgartner, M., *et al*. (2007). Dietary L-carnitine supplementation enhances the lipopolysaccharide-induced acute phase protein response in broiler chickens*. Vet Immunol Immunopathol*, 118(1-2), 154-159.

Caliando, B. J. & C. A. Voigt (2015). Targeted DNA degradation using a CRISPR device stably carried in the host genome. *Nat Commun*, 6, 6989.

Campos, J., Cristino, L., Peixe, L., & Antunes, P. (2016). MCR-1 in multidrug-resistant and copper-tolerant clinically relevant *Salmonella* 1,4,[5],12:i:- and *S.* Rissen clones in Portugal, 2011 to 2015. *Eurosurveillance*, 21(26), 30270.

Chan, K., Baker, S., Kim, C. C., Detweiler, C. S., Dougan, G., & Falkow, S. (2003). Genomic comparison of *Salmonella enterica* serovars and *Salmonella bongori* by use of an *S. enterica* serovar Typhimurium DNA microarray. *J Bacteriol*, 185(2), 553-563.

Chary, P., Prasad, R., Chopra, A. K., & Peterson, J. W. (1993). Location of the enterotoxin gene from *Salmonella* Typhimurium and characterization of the gene products. *FEMS Microbiol Lett*, 111(1), 87-92.

Chen, C. C., Chou, M. Y., Huang, C. H., Majumder, A., & Wu, H. Y. (2005). A cis-spreading nucleoprotein filament is responsible for the gene silencing activity found in the promoter relay mechanism. *J Biol Chem*, 280(6), 5101-5112.

Chen, L., Zheng, D., Liu, B., Yang, J., & Jin, Q. (2016). VFDB 2016: hierarchical and refined dataset for big data analysis--10 years on. *Nucleic Acids Res*, 44(D1), D694-697.

Cheng, L., Wang, J., Zhao, X., Yin, H., Fang, H., Lin, C., *et al*. (2020). An antiphage *Escherichia coli* mutant for higher production of L-threonine obtained by atmospheric and room temperature plasma mutagenesis. *Biotechnol Prog*, 36(6), e3058.

Cheng, R. A., & Wiedmann, M. (2020). Recent Advances in Our Understanding of the Diversity and Roles of Chaperone-Usher Fimbriae in Facilitating *Salmonella* Host and Tissue Tropism. *Front Cell Infect Microbiol*, 10, 628043.

Cheng, R. A., Eade, C. R., & Wiedmann, M. (2019). Embracing Diversity: Differences in Virulence Mechanisms, Disease Severity, and Host Adaptations Contribute to the Success of Nontyphoidal *Salmonella* as a Foodborne Pathogen*. Front Microbiol*, 10, 1368.

Colavecchio, A., Cadieux, B., Lo, A., & Goodridge, L. D. (2017). Bacteriophages Contribute to the Spread of Antibiotic Resistance Genes among Foodborne Pathogens of the *Enterobacteriaceae* Family - A Review*. Front Microbiol*, 8, 1108.

Collaborators, G. T. a. P. (2019). The global burden of typhoid and paratyphoid fevers: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Infect Dis*, 19, 369-381.

Collinson, S. K., Liu, S. L., Clouthier, S. C., Banser, P. A., Doran, J. L., Sanderson, K. E., *et al*. (1996). The location of four fimbrin-encoding genes, *agfA, fimA, sefA* and *sefD*, on the *Salmonella* Enteritidis and/or *S.* Typhimurium XbaI-BlnI genomic restriction maps. *Gene*, 169(1), 75-80.

Cook, R. B., Nathan; Redgwell, Tamsin; Rihtman, Branko; Barnes, Megan; Clokie, Martha; Stekel, Dov J.; Hobman, Jon ; Jones, Michael A.; Millard, Andrew. (2021). INfrastructure for a PHAge REference Database: Identification of Large-Scale Biases in the Current Collection of Cultured Phage Genomes. *PHAGE*, 2(4), 214-223.

Coombes, B. K., Wickham, M. E., Brown, N. F., Lemire, S., Bossi, L., Hsiao, W. W. L., *et al*. (2005). Genetic and Molecular Analysis of GogB, a Phage-encoded Type III-secreted Substrate in *Salmonella enterica* serovar Typhimurium with Autonomous Expression from its Associated Phage. *Journal of Molecular Biology*, 348(4), 817-830.

Cooper, L. A., Stringer, A. M., & Wade, J. T. (2018). Determining the Specificity of Cascade Binding, Interference, and Primed Adaptation. *mBio*, 9(2).

Costa, A. R., Berg, D. F. v. d., Esser, J. Q., Muralidharan, A., Bossche, H. v. d., Bonilla, B. E., *et al*. (2023). Accumulation of defense systems in phage resistant strains of *Pseudomonas aeruginosa*. *bioRxiv*, 2022.2008.2012.503731.

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., *et al*. (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.

Crooks, G. E., Hon, G., Chandonia, J. M., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res*, 14(6), 1188-1190.

Cui, L., Wang, X., Huang, D., Zhao, Y., Feng, J., Lu, Q., *et al*. (2020). CRISPR-*cas3* of *Salmonella* Upregulates Bacterial Biofilm Formation and Virulence to Host Cells by Targeting Quorum-Sensing Systems. *Pathogens*, 9(1).

Cunrath, O., & Palmer, J. D. (2021). An overview of *Salmonella enterica* metal homeostasis pathways during infection. *Microlife*, 2, uqab001.

Dai, C., Qu, Y., Wu, W., Li, S., Chen, Z., Lian, S., *et al*. (2023). QSP: An open sequence database for quorum sensing related gene analysis with an automatic annotation pipeline. *Water Research*, 235, 119814.

Daigle, F. (2008). Typhi genes expressed during infection or involved in pathogenesis. J Infect Dev Ctries, 2(6), 431-437.

Dalbey, R. E., Kaushik, S., & Kuhn, A. (2023). YidC as a potential antibiotic target. *Biochim Biophys Acta Mol Cell Res*, 1870(2), 119403.

Das, S., Bombaywala, S., Srivastava, S., Kapley, A., Dhodapkar, R., & Dafale, N. A. (2022). Genome plasticity as a paradigm of antibiotic resistance spread in ESKAPE pathogens. *Environ Sci Pollut Res Int*, 29(27), 40507-40519.

Daubin, V., Lerat, E., & Perrière, G. (2003). The source of laterally transferred genes in bacterial genomes. *Genome Biol*, 4(9), R57.

Dauga, C., Zabrovskaia, A., & Grimont, P. A. (1998). Restriction fragment length polymorphism analysis of some flagellin genes of *Salmonella enterica*. *J Clin Microbiol*, 36(10), 2835-2843.

De Groote, M. A., Ochsner, U. A., Shiloh, M. U., Nathan, C., McCord, J. M., Dinauer, M. C., *et al*. (1997). Periplasmic superoxide dismutase protects *Salmonella* from products of phagocyte NADPH-oxidase and nitric oxide synthase. *Proceedings of the National Academy of Sciences*, 94(25), 13997-14001.

Debarbieux, L., Bohin, A., & Bohin, J. P. (1997). Topological analysis of the membrane-bound glucosyltransferase, MdoH, required for osmoregulated periplasmic glucan synthesis in *Escherichia coli*. *J Bacteriol*, 179(21), 6692-6698.

Dekker, J. P., & Frank, K. M. (2015). *Salmonella*, *Shigella*, and *Yersinia*. *Clin Lab Med*, 35(2), 225-246.

Deng, Y., Xu, H., Su, Y., Liu, S., Xu, L., Guo, Z., *et al*. (2019). Horizontal gene transfer contributes to virulence and antibiotic resistance of *Vibrio harveyi* 345 based on complete genome sequence analysis. *BMC Genomics*, 20(1), 761.

Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., *et al*. (2008). Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res*, 36(Web Server issue), W465-469.

Di Domenico, E. G., I. Cavallo, M. Pontone, L. Toma & F. Ensoli (2017). Biofilm Producing *Salmonella* Typhi: Chronic Colonization and Development of Gallbladder Cancer. *Int J Mol Sci*, 18.

Díez-Villaseñor, C., Almendros, C., García-Martínez, J., & Mojica, F. J. (2010). Diversity of CRISPR loci in *Escherichia coli*. *Microbiology* (Reading), 156(Pt 5), 1351-1361.

Dillon, S. C., Espinosa, E., Hokamp, K., Ussery, D. W., Casadesús, J., & Dorman, C. J. (2012). LeuO is a global regulator of gene expression in *Salmonella enterica* serovar Typhimurium. *Mol Microbiol*, 85(6), 1072-1089.

Dobrindt, U., Agerer, F., Michaelis, K., Janka, A., Buchrieser, C., Samuelson, M., *et al*. (2003). Analysis of genome plasticity in pathogenic and commensal *Escherichia coli* isolates by use of DNA arrays. *J Bacteriol*, 185(6), 1831-1840.

Dobrindt, U., Hochhut, B., Hentschel, U., & Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews Microbiology*, 2(5), 414-424.

Dobrindt, U., Zdziarski, J., Salvador, E., & Hacker, J. (2010). Bacterial genome plasticity and its impact on adaptation during persistent infection. *Int J Med Microbiol*, 300(6), 363-366.

Donnenberg, C. R. a. M. S. (2014). Human Pathogenic Enterobacteriaceae *Reference Module in Biomedical Sciences*: Elsevier.

Doolittle, R. F., Feng, D. F., Tsang, S., Cho, G., & Little, E. (1996). Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science*, 271(5248), 470-477.

Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., *et al*. (2018). Systematic discovery of antiphage defense systems in the microbial pangenome. *Science*, 359(6379).

Doughty, S., Sloan, J., Bennett-Wood, V., Robertson, M., Robins-Browne, R. M., & Hartland, E. L. (2002). Identification of a novel fimbrial gene cluster related to long polar fimbriae in locus of enterocyte effacement-negative strains of enterohemorrhagic *Escherichia coli*. *Infect Immun*, 70(12), 6761-6769.

Dy, R. L., Pitman, A. R., & Fineran, P. C. (2013). Chromosomal targeting by CRISPR-Cas systems can contribute to genome plasticity in bacteria. *Mob Genet Elements*, 3(5), e26831.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32(5), 1792-1797.

Egido, J. E., Costa, A. R., Aparicio-Maldonado, C., Haas, P. J., & Brouns, S. J. J. (2022). Mechanisms and clinical importance of bacteriophage resistance. *FEMS Microbiol Rev*, 46(1).

Elhadad, D., Desai, P., Rahav, G., McClelland, M., & Gal-Mor, O. (2015). Flagellin Is Required for Host Cell Invasion and Normal *Salmonella* Pathogenicity Island 1 Expression by *Salmonella enterica* serovar Paratyphi A. *Infect Immun*, 83(9), 3355-3368.

Enault, F., Briet, A., Bouteille, L., Roux, S., Sullivan, M. B., & Petit, M.-A. (2017). Phages rarely encode antibiotic resistance genes: a cautionary tale for virome analyses. *The ISME Journal*, 11(1), 237-247.

Endo, A., Watanabe, T., Ogata, N., Nozawa, T., Aikawa, C., Arakawa, S., *et al*. (2015). Comparative genome analysis and identification of competitive and cooperative interactions in a polymicrobial disease. *ISME J*, 9(3), 629-642.

Eng, S.-K., Pusparajah, P., Ab Mutalib, N.-S., Ser, H.-L., Chan, K.-G., & Lee, L.-H. (2015). *Salmonella*: A review on pathogenesis, epidemiology and antibiotic resistance. *Frontiers in Life Science*, 8(3), 284-293.

Erickson, J. W., & Gross, C. A. (1989). Identification of the sigma E subunit of *Escherichia coli* RNA polymerase: a second alternate sigma factor involved in high-temperature gene expression. *Genes Dev*, 3(9), 1462-1471.

Eswarappa, S. M., G. Karnam, A. G. Nagarajan, S. Chakraborty & D. Chakravortty (2009). *lac* repressor is an antivirulence factor of *Salmonella enterica*: its role in the evolution of virulence in *Salmonella*. *PLoS One*, 4, e5789.

Fabre, L., Njamkepo, E., & Weill, F. X. (2021). Comment on Tanmoy *et al*. CRISPR-Cas Diversity in Clinical *Salmonella enterica* serovar Typhi Isolates from South Asian Countries. *Genes* (Basel), 12(8).

Fabre, L., Zhang, J., Guigon, G., Le Hello, S., Guibert, V., Accou-Demartin, M., *et al*. (2012). CRISPR typing and subtyping for improved laboratory surveillance of *Salmonella* infections. *PLoS One*, 7(5), e36995.

Fang, M., A. Majumder, K. J. Tsai & H. Y. Wu (2000). ppGpp-dependent *leuO* expression in bacteria under stress. *Biochem Biophys Res Commun*, 276, 64-70.

Faucher, S. P., Curtiss, R., & Daigle, F. (2005). Selective capture of *Salmonella enterica* serovar Typhi genes expressed in macrophages that are absent from the *Salmonella enterica* serovar Typhimurium genome. *Infect Immun*, 73(8), 5217-5221.

Feasey, N. A., G. Dougan, R. A. Kingsley, R. S. Heyderman & M. A. Gordon (2012). Invasive non-typhoidal *Salmonella* disease: an emerging and neglected tropical disease in Africa. *Lancet*, 379, 2489-2499.

Feldgarden, M., Brover, V., Haft, D. H., Prasad, A. B., Slotta, D. J., Tolstoy, I., *et al*. (2019). Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrob Agents Chemother*, 63(11).

Feng, Y., Liu, J., Li, Y.-G., Cao, F.-L., Johnston, R. N., Zhou, J., *et al*. (2012). Inheritance of the *Salmonella* virulence plasmids: Mostly vertical and rarely horizontal. *Infection, Genetics and Evolution*, 12(5), 1058-1063.

Ferrari, R. G., Rosario, D. K. A., Cunha-Neto, A., Mano, S. B., Figueiredo, E. E. S., & Conte-Junior, C. A. (2019). Worldwide Epidemiology of *Salmonella* serovars in Animal-Based Foods: a Meta-analysis. *Appl Environ Microbiol*, 85(14).

Ferreira, R. B., Buckner, M. M., & Finlay, B. B. (2012). Genome Plasticity in *Salmonella enterica* and Its Relevance to Host-Pathogen Interactions. *Genome Plasticity and Infectious Diseases* (pp. 84-102).

Fierer, J., & Guiney, D. G. (2001). Diverse virulence traits underlying different clinical outcomes of *Salmonella* infection. *The Journal of Clinical Investigation*, 107(7), 775-780.

Fineran, P. C., Gerritzen, M. J., Suárez-Diez, M., Künne, T., Boekhorst, J., van Hijum, S. A., *et al*. (2014). Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc Natl Acad Sci U S A*, 111(16), E1629-1638.

Fischbach, M., & Voigt, C. A. (2010). Prokaryotic gene clusters: a rich toolbox for synthetic biology. *Biotechnol J*, 5(12), 1277-1296.

Foley, S. L., Johnson, T. J., Ricke, S. C., Nayak, R., & Danzeisen, J. (2013). *Salmonella* pathogenicity and host adaptation in chicken-associated serovars. *Microbiol Mol Biol Rev*, 77(4), 582-607.

Fookes, M., Schroeder, G. N., Langridge, G. C., Blondel, C. J., Mammina, C., Connor, T. R., *et al*. (2011). *Salmonella bongori* provides insights into the evolution of the Salmonellae. *PLoS Pathog*, 7(8), e1002191.

Francino, M. P. (2012). The ecology of bacterial genes and the survival of the new. *Int J Evol Biol*, 2012, 394026.

Fricke, W. F., Mammel, M. K., McDermott, P. F., Tartera, C., White, D. G., Leclerc, J. E., *et al*. (2011). Comparative genomics of 28 *Salmonella enterica* isolates: evidence for CRISPR-mediated adaptive sublineage evolution. *J Bacteriol*, 193(14), 3556-3568.

Fronzes, R., Remaut, H., & Waksman, G. (2008). Architectures and biogenesis of non-flagellar protein appendages in Gram-negative bacteria. *EMBO J*, 27(17), 2271-2280.

Fukushima, M., Kakinuma, K., & Kawaguchi, R. (2002). Phylogenetic analysis of *Salmonella*, *Shigella*, and *Escherichia coli* strains on the basis of the *gyrB* gene sequence. *J Clin Microbiol*, 40(8), 2779-2785.

Gal-Mor, O. (2019). Persistent Infection and Long-Term Carriage of Typhoidal and Nontyphoidal Salmonellae. *Clin Microbiol Rev*, 32(1).

Gal-Mor, O., Boyle, E. C., & Grassl, G. A. (2014). Same species, different diseases: how and why typhoidal and non-typhoidal *Salmonella enterica* serovars differ. *Front Microbiol*, 5, 391.

Galperin, M. Y., & Koonin, E. V. (2000). Who's your neighbor? New computational approaches for functional genomics. *Nat Biotechnol*, 18(6), 609-613.

Gao, L., Altae-Tran, H., Böhning, F., Makarova, K. S., Segel, M., Schmid-Burgk, J. L., *et al*. (2020). Diverse enzymatic activities mediate antiviral immunity in prokaryotes. *Science*, 369(6507), 1077-1084.

Gao, X., Deng, L., Stack, G., Yu, H., Chen, X., Naito-Matsui, Y., *et al*. (2017). Evolution of host adaptation in the *Salmonella* typhoid toxin. *Nat Microbiol*, 2(12), 1592-1599.

García-Gutiérrez, E., Almendros, C., Mojica, F. J., Guzmán, N. M., & García-Martínez, J. (2015). CRISPR Content Correlates with the Pathogenic Potential of *Escherichia coli*. *PLoS One*, 10(7), e0131935.

Garcia-Russell, N., Elrod, B., & Dominguez, K. (2009). Stress-induced prophage DNA replication in *Salmonella enterica* serovar Typhimurium*. Infect Genet Evol*, 9(5), 889-895.

Gautreau, G., Bazin, A., Gachet, M., Planel, R., Burlot, L., Dubois, M., *et al*. (2020). PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph. *PLoS Computational Biology*, 16(3), e1007732.

Gayet, R., G. Bioley, N. Rochereau, S. Paul & B. Corthésy (2017). Vaccination against *Salmonella* Infection: the Mucosal Way. *Microbiol Mol Biol Rev*, 81.

Gillespie, J. J., Wattam, A. R., Cammer, S. A., Gabbard, J. L., Shukla, M. P., Dalay, O., *et al*. (2011). PATRIC: the comprehensive bacterial bioinformatics resource with a focus on human pathogenic species*. Infect Immun*, 79(11), 4286-4298.

Goldfarb, T., Sberro, H., Weinstock, E., Cohen, O., Doron, S., Charpak-Amikam, Y., *et al*. (2015). BREX is a novel phage resistance system widespread in microbial genomes. *EMBO J*, 34(2), 169-183.

Gomaa, A. A., H. E. Klumpe, M. L. Luo, K. Selle, R. Barrangou & C. L. Beisel (2014). Programmable removal of bacterial strains by use of genome-targeting CRISPR-Cas systems. *mBio*, 5, e00928-13.

Gong, J., Zeng, X., Zhang, P., Zhang, D., Wang, C., & Lin, J. (2019). Characterization of the emerging multidrug-resistant *Salmonella enterica* serovar Indiana strains in China. *Emerging Microbes & Infections*, 8(1), 29-39.

Gorlova, O., Fedorov, A., Logothetis, C., Amos, C., & Gorlov, I. (2014). Genes with a large intronic burden show greater evolutionary conservation on the protein level. *BMC Evol Biol*, 14(1), 50.

Grishkevich, V., & Yanai, I. (2014). Gene length and expression level shape genomic novelties. *Genome Res*, 24(9), 1497-1503.

Groisman, E. A., & Ochman, H. (1996). Pathogenicity islands: bacterial evolution in quantum leaps. *Cell*, 87(5), 791-794.

Guadarrama, C., Medrano-López, A., Oropeza, R., Hernández-Lucas, I., & Calva, E. (2014). The *Salmonella enterica* serovar Typhi LeuO global regulator forms tetramers: residues involved in oligomerization, DNA binding, and transcriptional regulation. *J Bacteriol*, 196(12), 2143-2154.

Guindon, S., & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*, 52(5), 696-704.

Guiney, D. G., Fang, F. C., Krause, M., & Libby, S. (1994). Plasmid-mediated virulence genes in non-typhoid *Salmonella* serovars. *FEMS Microbiology Letters*, 124(1), 1-9.

Guiney, D. G., Fang, F. C., Krause, M., Libby, S., Buchmeier, N. A., Fierer, J., *et al*. (1995). Biology and Clinical Significance of Virulence Plasmids in *Salmonella* serovars. *Clinical Infectious Diseases*, 21(Supplement_2), S146-S151.

Gulig, P. A., Danbara, H., Guiney, D. G., Lax, A. J., Norel, F., & Rhen, M (1993). Molecular analysis of *spv* virulence genes of the *Salmonella* virulence plasmids. *Molecular Microbiology*, 7(6), 825-830.

Guo, X., J. Chen, L. R. Beuchat & R. E. Brackett (2000). PCR detection of *Salmonella enterica* serotype Montevideo in and on raw tomatoes using primers derived from *hilA*. *Appl Environ Microbiol*, 66, 5248-52.

Gupta, S. K., Sharma, P., McMillan, E. A., Jackson, C. R., Hiott, L. M., Woodley, T., *et al*. (2019). Genomic comparison of diverse *Salmonella* serovars isolated from swine. *PLoS One*, 14(11), e0224518.

Gut, A. M., T. Vasiljevic, T. Yeager & O. N. Donkor (2018). *Salmonella* infection - prevention and treatment by antibiotics and probiotic yeasts: a review. *Microbiology* (Reading), 164, 1327-1344.

Hacker, J., & Carniel, E. (2001). Ecological fitness, genomic islands and bacterial pathogenicity. *EMBO reports*, 2(5), 376-381.

Halatsi, K., I. Oikonomou, M. Lambiri, G. Mandilara, A. Vatopoulos & A. Kyriacou (2006). PCR detection of *Salmonella* spp. using primers targeting the quorum sensing gene *sdiA*. *FEMS Microbiol Lett*, 259, 201-7.

Hale, C. R., Majumdar, S., Elmore, J., Pfister, N., Compton, M., Olson, S., *et al*. (2012). Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. *Mol Cell*, 45(3), 292-302.

Hamilton, T. A., Pellegrino, G. M., Therrien, J. A., Ham, D. T., Bartlett, P. C., Karas, B. J., *et al*. (2019). Efficient inter-species conjugative transfer of a CRISPR nuclease for targeted bacterial killing. *Nat Commun*, 10(1), 4544.

Hansen-Wester, I., Chakravortty, D., & Hensel, M. (2004). Functional Transfer of *Salmonella* Pathogenicity Island 2 to *Salmonella bongori and Escherichia coli*. *Infection and Immunity*, 72(5), 2879-2888.

Hasman, H., & Aarestrup, F. M. (2002). *tcrB*, a Gene Conferring Transferable Copper Resistance in *Enterococcus faecium*: Occurrence, Transferability, and Linkage to Macrolide and Glycopeptide Resistance. *Antimicrobial Agents and Chemotherapy*, 46(5), 1410-1416.

Hayes, R. P., Xiao, Y., Ding, F., van Erp, P. B., Rajashankar, K., Bailey, S., *et al*. (2016). Structural basis for promiscuous PAM recognition in type I-E Cascade from *E. coli*. *Nature*, 530(7591), 499-503.

Hayward, M. R., AbuOun, M., La Ragione, R. M., Tchórzewska, M. A., Cooley, W. A., Everest, D. J., *et al*. (2014). SPI-23 of *S.* Derby: role in adherence and invasion of porcine tissues. *PLoS One*, 9(9), e107857.

Hille, F., Richter, H., Wong, S. P., Bratovič, M., Ressel, S., & Charpentier, E. (2018). The Biology of CRISPR-Cas: Backward and Forward. *Cell*, 172(6), 1239-1259.

Hitchcock, P. J., Leive, L., Mäkelä, P. H., Rietschel, E. T., Strittmatter, W., & Morrison, D. C. (1986). Lipopolysaccharide nomenclature--past, present, and future. *J Bacteriol*, 166(3), 699-705.

Ho, T. D., & Slauch, J. M. (2001). Characterization of *grvA*, an Antivirulence Gene on the Gifsy-2 Phage in *Salmonella enterica* serovar Typhimurium. *Journal of Bacteriology*, 183(2), 611-620.

Hochhauser, D., Millman, A., & Sorek, R. (2023). The defense island repertoire of the *Escherichia coli* pan-genome. *PLoS Genet*, 19(4), e1010694.

Hochstrasser, M. L., Taylor, D. W., Bhat, P., Guegler, C. K., Sternberg, S. H., Nogales, E., *et al*. (2014). CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. *Proc Natl Acad Sci U S A*, 111(18), 6618-6623.

Holt, K. E., Thomson, N. R., Wain, J., Phan, M. D., Nair, S., Hasan, R., *et al*. (2007). Multidrug-resistant *Salmonella enterica* serovar Paratyphi A harbors IncHI1 plasmids similar to those found in serovar Typhi. *J Bacteriol*, 189(11), 4257-4264.

Horváth, P., Kato, T., Miyata, T., & Namba, K. (2019). Structure of *Salmonella* Flagellar Hook Reveals Intermolecular Domain Interactions for the Universal Joint Function. *Biomolecules*, 9(9).

Horvath, P., Romero, D. A., Coûté-Monvoisin, A. C., Richards, M., Deveau, H., Moineau, S., *et al*. (2008). Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J Bacteriol*, 190(4), 1401-1412.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., *et al*. (2019). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res*, 47(D1), D309-D314.

Humphries, A. D., Raffatellu, M., Winter, S., Weening, E. H., Kingsley, R. A., Droleskey, R., *et al*. (2003). The use of flow cytometry to detect expression of subunits encoded by 11 *Salmonella enterica* serotype Typhimurium fimbrial operons. *Mol Microbiol*, 48(5), 1357-1376.

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.

Ilyas, B., Tsai, C. N., & Coombes, B. K. (2017). Evolution of *Salmonella*-Host Cell Interactions through a Dynamic Bacterial Genome. *Front Cell Infect Microbiol*, 7, 428.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., & Nakata, A. (1987). Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *J Bacteriol*, 169(12), 5429-5433.

Jaglic, Z., & Cervinkova, D. (2012). Genetic basis of resistance to quaternary ammonium compounds- the *qac* genes and their role: a review. *Veterinární medicína*, 57(6), 275-281.

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun*, 9(1), 5114.

Janda, J. M., & Abbott, S. L. (2021). The Changing Face of the Family *Enterobacteriaceae* (Order: "*Enterobacterales*"): New Members, Taxonomic Issues, Geographic Expansion, and New Diseases and Disease Syndromes. *Clin Microbiol Rev*, 34(2).

Jennings, E., Thurston, T. L. M., & Holden, D. W. (2017). *Salmonella* SPI-2 Type III Secretion System Effectors: Molecular Mechanisms And Physiological Consequences. *Cell Host Microbe*, 22(2), 217-231.

Jia, B., Raphenya, A. R., Alcock, B., Waglechner, N., Guo, P., Tsang, K. K., *et al*. (2017). CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res*, 45(D1), D566-D573.

Jiang, F., & Doudna, J. A. (2017). CRISPR-Cas9 Structures and Mechanisms. *Annu Rev Biophys*, 46, 505-529.

Jiang, L., Li, X., Lv, R., & Feng, L. (2019). LoiA directly represses *lon* gene expression to activate the expression of *Salmonella* pathogenicity island-1 genes. *Res Microbiol*, 170(3), 131-137.

Jin, G., Nakhleh, L., Snir, S., & Tuller, T. (2007). Inferring phylogenetic networks by the maximum parsimony criterion: a case study. *Mol Biol Evol*, 24(1), 324-337.

Johnson, M. C., Laderman, E., Huiting, E., Zhang, C., Davidson, A., & Bondy-Denomy, J. (2023). Core defense hotspots within *Pseudomonas aeruginosa* are a consistent and rich source of anti-phage defense systems. *Nucleic Acids Res*, 51(10), 4995-5005.

Jolley, K. A., Bray, J. E., & Maiden, M. C. J. (2018). Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res*, 3, 124.

Jones, T. F., Ingram, L. A., Cieslak, P. R., Vugia, D. J., Tobin-D'Angelo, M., Hurd, S., *et al*. (2008). Salmonellosis outcomes differ substantially by serotype. *J Infect Dis*, 198(1), 109-114.

Kadouri, D. E., K. To, R. M. Shanks & Y. Doi (2013). Predatory bacteria: a potential ally against multidrug-resistant Gram-negative pathogens. *PLoS One*, 8, e63397.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., & Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res*, 32(Database issue), D277-280.

Kaniuk, N. A., Monteiro, M. A., Parker, C. T., & Whitfield, C. (2002). Molecular diversity of the genetic loci responsible for lipopolysaccharide core oligosaccharide assembly within the genus *Salmonella*. *Mol Microbiol*, 46(5), 1305-1318.

Karimi, Z., Ahmadi, A., Najafi, A., & Ranjbar, R. (2018). Bacterial CRISPR Regions: General Features and their Potential for Epidemiological Molecular Typing Studies. *Open Microbiol J*, 12, 59-70.

Khan, M. A. S., & Rahman, S. R. (2022). Use of Phages to Treat Antimicrobial-Resistant. *Vet Sci,* 9(8).

Kim, M., & Ryu, S. (2012). Spontaneous and transient defence against bacteriophage by phase-variable glucosylation of O-antigen in *Salmonella enterica* serovar Typhimurium. *Mol Microbiol*, 86(2), 411-425.

Kirkconnell, K. S., Magnuson, B., Paulsen, M. T., Lu, B., Bedi, K., & Ljungman, M. (2017). Gene length as a biological timer to establish temporal transcriptional regulation. *Cell Cycle*, 16(3), 259-270.

Klein, E. Y., T. P. Van Boeckel, E. M. Martinez, S. Pant, S. Gandra, S. A. Levin, H. Goossens & R. Laxminarayan (2018). Global increase and geographic convergence in antibiotic consumption between 2000 and 2015. *Proc Natl Acad Sci U S A*, 115, E3463-E3470.

Koonin, E. V., & Makarova, K. S. (2019). Origins and evolution of CRISPR-Cas systems. *Philos Trans R Soc Lond B Biol Sci*, 374(1772), 20180087.

Koonin, E. V., Makarova, K. S., & Zhang, F. (2017). Diversity, classification and evolution of CRISPR-Cas systems. *Curr Opin Microbiol*, 37, 67-78.

Krivoy, A., Rutkauskas, M., Kuznedelov, K., Musharova, O., Rouillon, C., Severinov, K., *et al*. (2018). Primed CRISPR adaptation in *Escherichia coli* cells does not depend on conformational changes in the Cascade effector complex detected in Vitro. *Nucleic Acids Res*, 46(8), 4087-4098.

Kropinski, A. M., Sulakvelidze, A., Konczy, P., & Poppe, C. (2007). *Salmonella* Phages and Prophages—Genomics and Practical Aspects. In H. Schatten & A. Eisenstark (Eds.), *Salmonella*: *Methods and Protocols* (pp. 133-175). Totowa, NJ: Humana Press.

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol*, 35(6), 1547-1549.

Kurtz, J. R., J. A. Goggins & J. B. McLachlan (2017). *Salmonella* infection: Interplay between the bacteria and host immune system. *Immunol Lett*, 190, 42-50.

Kushwaha, S. K., Bhavesh, N. L. S., Abdella, B., Lahiri, C., & Marathe, S. A. (2020). The phylogenomics of CRISPR-Cas system and revelation of its features in *Salmonella*. *Sci Rep*, 10(1), 21156.

Kushwaha, S. K., L. P. Narasimhan, C. Chithananthan & S. A. Marathe (2022). Clustered regularly interspaced short palindromic repeats-Cas system: diversity and regulation in *Enterobacteriaceae*. *Future Microbiol*, 17, 1249-1267.

Kushwaha, S. K., Anand, A., Wu, Y., Avila, H. L., Sicheritz-Ponten, T., Millard, A., *et al*. (2023). Genomic plasticity is a blueprint of diversity in *Salmonella* lineages. *bioRxiv*, 2023.2012.2002.569618.

Kushwaha, S. K., Kumar, A. A., Gupta, H., & Marathe, S. A. (2023). The Phylogenetic Study of the CRISPR-Cas System in *Enterobacteriaceae*. *Curr Microbiol*, 80(6), 196.

Laing, C. R., Whiteside, M. D., & Gannon, V. P. J. (2017). Pan-genome Analyses of the Species *Salmonella enterica*, and Identification of Genomic Markers Predictive for Species, Subspecies, and Serovar. *Front Microbiol*, 8, 1345.

Lamas, A., Miranda, J. M., Regal, P., Vázquez, B., Franco, C. M., & Cepeda, A. (2018). A comprehensive review of non-*enterica* subspecies of *Salmonella enterica*. *Microbiol Res*, 206, 60-73.

Lawrence, J. G. (2002). Shared Strategies in Gene Organization among Prokaryotes and Eukaryotes. *Cell*, 110(4), 407-413.

Ledeboer, N. A., Frye, J. G., McClelland, M., & Jones, B. D. (2006). *Salmonella enterica* serovar Typhimurium Requires the Lpf, Pef, and Tafi Fimbriae for Biofilm Formation on HEp-2 Tissue Culture Cells and Chicken Intestinal Epithelium. *Infection and Immunity*, 74(6), 3156-3169.

Lee, Y. H., Kim, B. H., Kim, J. H., Yoon, W. S., Bang, S. H., & Park, Y. K. (2007). CadC has a global translational effect during acid adaptation in *Salmonella enterica* serovar Typhimurium. *J Bacteriol*, 189(6), 2417-2425.

Lees, J. A., Harris, S. R., Tonkin-Hill, G., Gladstone, R. A., Lo, S. W., Weiser, J. N., *et al*. (2019). Fast and flexible bacterial genomic epidemiology with PopPUNK. *Genome Res*, 29(2), 304-316.

Lei, J., L. Sun, S. Huang, C. Zhu, P. Li, J. He, V. Mackey, D. H. Coy & Q. He (2019). The antimicrobial peptides and their potential clinical applications. *Am J Transl Res*, 11, 3919-3931.

Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL).v4: recent updates and new developments. *Nucleic Acids Res*, 47(W1), W256-W259.

Lewis, M., Chang, G., Horton, N. C., Kercher, M. A., Pace, H. C., Schumacher, M. A., *et al*. (1996). Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science*, 271(5253), 1247-1254.

Li, H. Y., Kao, C. Y., Lin, W. H., Zheng, P. X., Yan, J. J., Wang, M. C., *et al*. (2018). Characterization of CRISPR-Cas Systems in Clinical *Klebsiella pneumoniae* Isolates Uncovers Its Potential Association With Antibiotic Susceptibility. *Front Microbiol*, 9, 1595.

Li, R., Fang, L., Tan, S., Yu, M., Li, X., He, S., *et al*. (2016). Type I CRISPR-Cas targets endogenous genes and regulates virulence to evade mammalian host immunity. *Cell Res*, 26(12), 1273-1287.

Li, S., Zhang, S., Baert, L., Jagadeesan, B., Ngom-Bru, C., Griswold, T., *et al*. (2019). Implications of Mobile Genetic Elements for *Salmonella enterica* Single-Nucleotide Polymorphism Subtyping and Source Tracking Investigations. *Applied and Environmental Microbiology*, 85(24), e01985-01919.

Libby, S. J., Lesnick, M., Hasegawa, P., Kurth, M., Belcher, C., Fierer, J., *et al*. (2002). Characterization of the *spv* locus in *Salmonella enterica* serovar Arizona. *Infect Immun*, 70(6), 3290-3294.

Lima, T., Domingues, S., & Da Silva, G. J. (2019). Plasmid-Mediated Colistin Resistance in *Salmonella enterica*: A Review*. Microorganisms*, 7(2).

Lin, I. H., Liu, T. T., Teng, Y. T., Wu, H. L., Liu, Y. M., Wu, K. M., *et al*. (2011). Sequencing and comparative genome analysis of two pathogenic *Streptococcus gallolyticus* subspecies: genome plasticity, adaptation and virulence. *PLoS One*, 6(5), e20519.

Lin, J., Du, F., Long, M., & Li, P. (2022). Limitations of Phage Therapy and Corresponding Optimization Strategies: A Review. *Molecules*, 27(6).

Liu, G. R., Liu, W. Q., Johnston, R. N., Sanderson, K. E., Li, S. X., & Liu, S. L. (2006). Genome plasticity and *ori-ter* rebalancing in *Salmonella* Typhi. *Mol Biol Evol*, 23(2), 365-371.

Liu, W. Q., Liu, G. R., Li, J. Q., Xu, G. M., Qi, D., He, X. Y., *et al*. (2007). Diverse genome structures of *Salmonella* Paratyphi C. *BMC Genomics*, 8, 290.

Liu, W. T., Karavolos, M. H., Bulmer, D. M., Allaoui, A., Hormaeche, R. D., Lee, J. J., *et al*. (2007). Role of the universal stress protein UspA of *Salmonella* in growth arrest, stress and virulence. *Microb Pathog*, 42(1), 2-10.

Liu, Y., Chen, H., Kenney, L. J., & Yan, J. (2010). A divalent switch drives H-NS/DNA-binding conformations between stiffening and bridging modes. *Genes Dev*, 24(4), 339-344.

Lou, L., Zhang, P., Piao, R., & Wang, Y. (2019). Pathogenicity Island 1 (SPI-1) and Its Complex Regulatory Network. *Front Cell Infect Microbiol*, 9, 270.

Louwen, R., R. H. Staals, H. P. Endtz, P. van Baarlen & J. van der Oost (2014). The role of CRISPR-Cas systems in virulence of pathogenic bacteria. *Microbiol Mol Biol Rev*, 78, 74-88.

MacLennan, C. A., L. B. Martin & F. Micoli (2014). Vaccines against invasive *Salmonella* disease: current status and future directions. *Hum Vaccin Immunother*, 10, 1478-93.

Makarova, K. S., Anantharaman, V., Grishin, N. V., Koonin, E. V., & Aravind, L. (2014). CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems. *Front Genet*, 5, 102.

Makarova, K. S., Wolf, Y. I., Iranzo, J., Shmakov, S. A., Alkhnbashi, O. S., Brouns, S. J. J., *et al*. (2020). Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol*, 18(2), 67-83.

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.

Malo, M. S., & Loughlin, R. E. (1990). Promoter elements and regulation of expression of the *cysD* gene of *Escherichia coli* K-12. *Gene*, 87(1), 127-131.

Malorny, B., E. Paccassoni, P. Fach, C. Bunge, A. Martin & R. Helmuth (2004). Diagnostic real-time PCR for detection of *Salmonella* in food. *Appl Environ Microbiol*, 70, 7046-52.

Mambu, J., Virlogeux-Payant, I., Holbert, S., Grépinet, O., Velge, P., & Wiedemann, A. (2017). An Updated View on the Rck Invasin of *Salmonella*: Still Much to Discover. *Frontiers in Cellular and Infection Microbiology*, 7.

Mann, M., Wright, P. R., & Backofen, R. (2017). IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. *Nucleic Acids Res*, 45(W1), W435-W439.

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol*, 38(10), 4647-4654.

Marcus, S. L., Brumell, J. H., Pfeifer, C. G., & Finlay, B. B. (2000). *Salmonella* pathogenicity islands: big virulence in small packages. *Microbes Infect*, 2(2), 145-156.

Masi, M., & Pagès, J. M. (2013). Structure, Function and Regulation of Outer Membrane Proteins Involved in Drug Transport in *Enterobacteriaceae*: the OmpF/C - TolC Case. *Open Microbiol J*, 7, 22-33.

Mastrorilli, E., Petrin, S., Orsini, M., Longo, A., Cozza, D., Luzzi, I., *et al*. (2020). Comparative genomic analysis reveals high intra-serovar plasticity within *Salmonella* Napoli isolated in 2005-2017. *BMC Genomics*, 21(1), 202.

Mathee, K., Narasimhan, G., Valdes, C., Qiu, X., Matewish, J. M., Koehrsen, M., *et al*. (2008). Dynamics of *Pseudomonas aeruginosa* genome evolution. *Proceedings of the National Academy of Sciences*, 105(8), 3100-3105.

McClelland, M., Florea, L., Sanderson, K., Clifton, S. W., Parkhill, J., Churcher, C., *et al*. (2000). Comparison of the *Escherichia coli* K-12 genome with sampled genomes of a *Klebsiella pneumoniae* and three *Salmonella enterica* serovars, Typhimurium, Typhi and Paratyphi. *Nucleic Acids Res*, 28(24), 4974-4986.

McDonald, N. D., Regmi, A., Morreale, D. P., Borowski, J. D., & Boyd, E. F. (2019). CRISPR-Cas systems are present predominantly on mobile genetic elements in *Vibrio* species. *BMC Genomics*, 20(1), 105.

McMillan, E. A., Jackson, C. R., & Frye, J. G. (2020). Transferable Plasmids of *Salmonella enterica* Associated With Antibiotic Resistance Genes. *Frontiers in Microbiology*, 11.

McQuiston, J. R., Fields, P. I., Tauxe, R. V., & Logsdon, J. M. (2008). Do *Salmonella* carry spare tyres? *Trends Microbiol*, 16(4), 142-148.

Medina-Aparicio, L., Rebollar-Flores, J. E., Beltrán-Luviano, A. A., Vázquez, A., Gutiérrez-Ríos, R. M., Olvera, L., *et al*. (2017). CRISPR-Cas system presents multiple transcriptional units including antisense RNAs that are expressed in minimal medium and upregulated by pH in *Salmonella enterica* serovar Typhi. *Microbiology* (Reading), 163(2), 253-265.

Medina-Aparicio, L., Rebollar-Flores, J. E., Gallego-Hernández, A. L., Vázquez, A., Olvera, L., Gutiérrez-Ríos, R. M., *et al*. (2011). The CRISPR/Cas immune system is an operon regulated by LeuO, H-NS, and leucine-responsive regulatory protein in *Salmonella enterica* serovar Typhi. *J Bacteriol*, 193(10), 2396-2407.

Medina-Aparicio, L., Rodriguez-Gutierrez, S., Rebollar-Flores, J. E., Martínez-Batallar, Á., Mendoza-Mejía, B. D., Aguirre-Partida, E. D., *et al*. (2021). The CRISPR-Cas System Is Involved in OmpR Genetic Regulation for Outer Membrane Protein Synthesis in *Salmonella* Typhi. *Front Microbiol*, 12, 657404.

Medina-Aparicio, L., S. Dávila, J. E. Rebollar-Flores, E. Calva & I. Hernández-Lucas (2018). The CRISPR-Cas system in *Enterobacteriaceae*. *Pathog Dis*, 76.

Mey, A. R., Gómez-Garzón, C., & Payne, S. M. (2021). Iron Transport and Metabolism in *Escherichia*, *Shigella*, and *Salmonella*. *EcoSal Plus*, 9(2), eESP00342020.

Michael, P. S. a. W. J. K. (2012). Resistance and survival strategies of *Salmonella enterica* to environmental stresses. *Food Research International*, 45(2), 455-481.

Millman, A., Melamed, S., Leavitt, A., Doron, S., Bernheim, A., Hör, J., *et al*. (2022). An expanded arsenal of immune systems that protect bacteria from phages. *Cell Host Microbe*, 30(11), 1556-1569.e1555.

Nakano, M., Yamasaki, E., Ichinose, A., Shimohata, T., Takahashi, A., Akada, J. K., *et al*. (2012). *Salmonella* enterotoxin (Stn) regulates membrane composition and integrity. *Dis Model Mech*, 5(4), 515-521.

Navarre, W. W., Porwollik, S., Wang, Y., McClelland, M., Rosen, H., Libby, S. J., *et al*. (2006). Selective silencing of foreign DNA with low GC content by the H-NS protein in *Salmonella*. *Science*, 313(5784), 236-238.

Netea, M. G., Schlitzer, A., Placek, K., Joosten, L. A. B., & Schultze, J. L. (2019). Innate and Adaptive Immune Memory: an Evolutionary Continuum in the Host's Response to Pathogens. *Cell Host Microbe*, 25(1), 13-26.

Nguyen, S. V., Harhay, D. M., Bono, J. L., Smith, T. P. L., Fields, P. I., Dinsmore, B. A., *et al*. (2018). Comparative genomics of *Salmonella enterica* serovar Montevideo reveals lineage-specific gene differences that may influence ecological niche association. *Microb Genom*, 4(8).

Nitschké, P., Guerdoux-Jamet, P., Chiapello, H., Faroux, G., Hénaut, C., Hénaut, A., *et al*. (1998). Indigo: a World-Wide-Web review of genomes and gene functions. *FEMS Microbiol Rev*, 22(4), 207-227.

Nobrega, F. L., Walinga, H., Dutilh, B. E., & Brouns, S. J. J. (2020). Prophages are associated with extensive CRISPR-Cas auto-immunity. *Nucleic Acids Res*, 48(21), 12074-12084.

Nuccio, S. P., & Bäumler, A. J. (2007). Evolution of the chaperone/usher assembly pathway: fimbrial classification goes Greek. *Microbiol Mol Biol Rev*, 71(4), 551-575.

Nuñez, J. K., Kranzusch, P. J., Noeske, J., Wright, A. V., Davies, C. W., & Doudna, J. A. (2014). Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat Struct Mol Biol*, 21(6), 528-534.

Ochman, H., & Groisman, E. A. (1996). Distribution of pathogenicity islands in *Salmonella* spp. *Infect Immun*, 64(12), 5410-5412.

Oelschlaeger, T. A., Zhang, D., Schubert, S., Carniel, E., Rabsch, W., Karch, H., *et al*. (2003). The High-Pathogenicity Island Is Absent in Human Pathogens of *Salmonella enterica* Subspecies I but Present in Isolates of Subspecies III and VI. *Journal of Bacteriology*, 185(3), 1107-1111.

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.

Ondov, B. D., Starrett, G. J., Sappington, A., Kostic, A., Koren, S., Buck, C. B., *et al*. (2019). Mash Screen: high-throughput sequence containment estimation for genome discovery. *Genome Biol*, 20(1), 232.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., *et al*. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol*, 17(1), 132.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., & Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci U S A*, 96(6), 2896-2901.

Payne, L. J., Todeschini, T. C., Wu, Y., Perry, B. J., Ronson, Clive W., Fineran, Peter C., *et al*. (2021). Identification and classification of antiviral defence systems in bacteria and archaea with PADLOC reveals new system types. *Nucleic Acids Research*, 49(19), 10868-10878.

Pearce, M. E., Langridge, G. C., Lauer, A. C., Grant, K., Maiden, M. C. J., & Chattaway, M. A. (2021). An evaluation of the species and subspecies of the genus *Salmonella* with whole genome sequence data: Proposal of type strains and epithets for novel *S. enterica* subspecies VII, VIII, IX, X and XI. *Genomics*, 113(5), 3152-3162.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., & Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, 96(8), 4285-4288.

Pérez-Losada, M., Arenas, M., & Castro-Nallar, E. (2018). Microbial sequence typing in the genomic era. *Infect Genet Evol*, 63, 346-359.

Perez-Rodriguez, R., C. Haitjema, Q. Huang, K. H. Nam, S. Bernardis, A. Ke & M. P. DeLisa (2011). Envelope stress is a trigger of CRISPR RNA-mediated DNA silencing in *Escherichia coli*. *Mol Microbiol*, 79, 584-99.

Perez, J. C., & Groisman, E. A. (2007). Acid pH activation of the PmrA/PmrB two-component regulatory system of *Salmonella enterica*. *Mol Microbiol*, 63(1), 283-293.

Pettengill, J. B., Timme, R. E., Barrangou, R., Toro, M., Allard, M. W., Strain, E., *et al*. (2014). The evolutionary history and diagnostic utility of the CRISPR-Cas system within *Salmonella enterica* ssp. *enterica*. *PeerJ*, 2, e340.

Pilar, A. V. C., Reid-Yu, S. A., Cooper, C. A., Mulder, D. T., & Coombes, B. K. (2012). GogB Is an Anti-Inflammatory Effector that Limits Tissue Damage during *Salmonella* Infection through Interaction with Human FBXO22 and Skp1. *PLoS Pathogens*, 8(6), e1002773.

Pilla, G., Arcari, G., Tang, C. M., & Carattoli, A. (2022). Virulence plasmid pINV as a genetic signature for *Shigella flexneri* phylogeny. *Microbial Genomics*, 8(6).

Pinilla-Redondo, R., Shehreen, S., Marino, N. D., Fagerlund, R. D., Brown, C. M., Sørensen, S. J., *et al*. (2020). Discovery of multiple anti-CRISPRs highlights anti-defense gene clustering in mobile genetic elements. *Nat Commun*, 11(1), 5652.

Plumb, I. D., C. A. Schwensohn, L. Gieraltowski, S. Tecle, Z. D. Schneider, J. Freiman, A. Cote, D. Noveroske, J. Kolsin, J. Brandenburg, J. C. Chen, K. A. Tagg, P. B. White, H. J. Shah, L. K. Francois Watkins, M. E. Wise & C. R. Friedman (2019). Outbreak of *Salmonella* Newport Infections with Decreased Susceptibility to Azithromycin Linked to Beef Obtained in the United States and Soft Cheese Obtained in Mexico- United States, 2018-2019. *MMWR Morb Mortal Wkly Rep*, 68, 713-717.

Pontel, L. B., Audero, M. E., Espariz, M., Checa, S. K., & Soncini, F. C. (2007). GolS controls the response to gold by the hierarchical induction of *Salmonella*-specific genes that include a CBA efflux-coding operon. *Mol Microbiol*, 66(3), 814-825.

Porwollik, S., Boyd, E. F., Choy, C., Cheng, P., Florea, L., Proctor, E., *et al*. (2004). Characterization of *Salmonella enterica* subspecies I genovars by use of microarrays. *J Bacteriol*, 186(17), 5883-5898.

Porwollik, S., Wong, R. M., & McClelland, M. (2002). Evolutionary genomics of *Salmonella*: gene acquisitions revealed by microarray analysis. *Proc Natl Acad Sci U S A*, 99(13), 8956-8961.

Pul, U., R. Wurm, Z. Arslan, R. Geissen, N. Hofmann & R. Wagner (2010). Identification and characterization of *E. coli* CRISPR-*cas* promoters and their silencing by H-NS. *Mol Microbiol*, 75, 1495-512.

Purighalla, S., S. Esakimuthu, M. Reddy, T. Seth, S. D. Patil, G. K. Varghese, R. Dasarathy, V. S. Richard & V. K. Sambandamurthy (2017). Investigation into a community outbreak of *Salmonella* Typhi in Bengaluru, India. *Indian J Med Res*, 146, S15-S22.

Pursey, E., D. Sünderhauf, W. H. Gaze, E. R. Westra & S. van Houte (2018). CRISPR-Cas antimicrobials: Challenges and future prospects. *PLoS Pathog*, 14, e1006990.

Radovcic, M., Killelea, T., Savitskaya, E., Wettstein, L., Bolt, E. L., & Ivancic-Bace, I. (2018). CRISPR-Cas adaptation in *Escherichia coli* requires RecBCD helicase but not nuclease activity, is independent of homologous recombination, and is antagonized by 5' ssDNA exonucleases. *Nucleic Acids Res*, 46(19), 10173-10183.

Ramírez-Larrota, J. S., & Eckhard, U. (2022). An Introduction to Bacterial Biofilms and Their Proteases, and Their Roles in Host Infection and Immune Evasion. *Biomolecules*, 12(2).

Ravenhall, M., Škunca, N., Lassalle, F., & Dessimoz, C. (2015). Inferring horizontal gene transfer. *PLoS Comput Biol*, 11(5), e1004095.

Redding, S., Sternberg, S. H., Marshall, M., Gibb, B., Bhat, P., Guegler, C. K., *et al*. (2015). Surveillance and Processing of Foreign DNA by the *Escherichia coli* CRISPR-Cas System. *Cell*, 163(4), 854-865.

Richter, C., Chang, J. T., & Fineran, P. C. (2012). Function and regulation of clustered regularly interspaced short palindromic repeats (CRISPR) / CRISPR associated (Cas) systems. *Viruses*, 4(10), 2291-2311.

Robertson, J., Schonfeld, J., Bessonov, K., Bastedo, P., & Nash, J. H. E. (2023). A global survey of *Salmonella* plasmids and their associations with antimicrobial resistance. *Microb Genom*, 9(5).

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.

Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S. A., & Sørensen, S. J. (2020). CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. *CRISPR J*, 3(6), 462-469.

Rychlik, I., Gregorova, D., & Hradecka, H. (2006). Distribution and function of plasmids in *Salmonella enterica*. *Vet Microbiol*, 112(1), 1-10.

Sabbagh, S. C., Forest, C. G., Lepage, C., Leclerc, J. M., & Daigle, F. (2010). So similar, yet so different: uncovering distinctive features in the genomes of *Salmonella enterica* serovars Typhimurium and Typhi. *FEMS Microbiol Lett*, 305(1), 1-13.

Sampson, T. R., & Weiss, D. S. (2013). Degeneration of a CRISPR/Cas system and its regulatory target during the evolution of a pathogen. *RNA Biol*, 10(10), 1618-1622.

Sampson, T. R., & Weiss, D. S. (2014). CRISPR-Cas systems: new players in gene regulation and bacterial physiology. *Front Cell Infect Microbiol*, 4, 37.

Sanders, M. E., Akkermans, L. M., Haller, D., Hammerman, C., Heimbach, J., Hörmannsperger, G., *et al*. (2010). Safety assessment of probiotics for human use. *Gut Microbes*, 1(3), 164-185.

Schmartz, G. P., Hartung, A., Hirsch, P., Kern, F., Fehlmann, T., Müller, R., *et al*. (2022). PLSDB: advancing a comprehensive database of bacterial plasmids. *Nucleic Acids Res*, 50(D1), D273-D278.

Schmidt, H., & Hensel, M. (2004). Pathogenicity islands in bacterial pathogenesis. *Clin Microbiol Rev*, 17(1), 14-56.

Schwengers, O., Barth, P., Falgenhauer, L., Hain, T., Chakraborty, T., & Goesmann, A. (2020). Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microbial Genomics*, 6(10).

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, 30(14), 2068-2069.

Shabbir, M. A. B., Tang, Y., Xu, Z., Lin, M., Cheng, G., Dai, M., *et al*. (2018). The Involvement of the of the Cas9 Gene in Virulence of *Campylobacter jejuni*. *Front Cell Infect Microbiol*, 8, 285.

Shabbir, M. A., Hao, H., Shabbir, M. Z., Hussain, H. I., Iqbal, Z., Ahmed, S., *et al*. (2016). Survival and Evolution of CRISPR-Cas System in Prokaryotes and Its Applications. *Front Immunol*, 7, 375.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., *et al*. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11), 2498-2504.

Shariat, N., Timme, R. E., Pettengill, J. B., Barrangou, R., & Dudley, E. G. (2015). Characterization and evolution of *Salmonella* CRISPR-Cas systems. *Microbiology*, 161(Pt 2), 374-386.

Sharma, N., A. Das, P. Raja & S. A. Marathe (2022). The CRISPR-Cas System Differentially Regulates Surface-Attached and Pellicle Biofilm in *Salmonella enterica* serovar Typhimurium. *Microbiol Spectr*, e0020222.

Sheikh, A., Charles, R. C., Sharmeen, N., Rollins, S. M., Harris, J. B., Bhuiyan, M. S., *et al*. (2011). In vivo expression of *Salmonella enterica* serotype Typhi genes in the blood of patients with typhoid fever in Bangladesh. *PLoS Negl Trop Dis*, 5(12), e1419.

Sheppard, S. K., Guttman, D. S., & Fitzgerald, J. R. (2018). Population genomics of bacterial host adaptation. *Nat Rev Genet*, 19(9), 549-565.

Sheth, A. N., M. Hoekstra, N. Patel, G. Ewald, C. Lord, C. Clarke, E. Villamil, K. Niksich, C. Bopp, T. A. Nguyen, D. Zink & M. Lynch (2011). A national outbreak of *Salmonella* serotype Tennessee infections from contaminated peanut butter: a new food vehicle for salmonellosis in the United States. *Clin Infect Dis*, 53, 356-62.

Shi, Y., Li, J., Shen, Y., & Sun, Z. (2020). Using Probiotics to Mute *Salmonella enteric* serovar Typhimurium: An Opinion. *Front Bioeng Biotechnol*, 8, 558.

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.

Siala, M., A. Barbana, S. Smaoui, S. Hachicha, C. Marouane, S. Kammoun, R. Gdoura & F. Messadi-Akrout (2017). Screening and Detecting *Salmonella* in Different Food Matrices in Southern Tunisia Using a Combined Enrichment/Real-Time PCR Method: Correlation with Conventional Culture Method. *Front Microbiol*, 8, 2416.

Silva, C., Puente, J. L., & Calva, E. (2017). *Salmonella* virulence plasmid: pathogenesis and ecology. *Pathog Dis*.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), 3210-3212.

Sirén, K., Millard, A., Petersen, B., Gilbert, M Thomas P., Clokie, M. R. J., & Sicheritz-Pontén, T. (2021). Rapid discovery of novel prophages using biological feature engineering and machine learning. *NAR Genomics and Bioinformatics*, 3(1).

Solbiati, J., Duran-Pinedo, A., Godoy Rocha, F., Gibson, F. C., & Frias-Lopez, J. (2020). Virulence of the Pathogen *Porphyromonas gingivalis* Is Controlled by the CRISPR-Cas Protein Cas3. *mSystems*, 5(5).

Stanaway, J. D., Reiner, R. C., Blacker, B. F., Goldberg, E. M., Khalil, I. A., Troeger, C. E., *et al*. (2019). The global burden of typhoid and paratyphoid fevers: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet Infectious Diseases*, 19(4), 369-381.

Starikova, E. V., Tikhonova, P. O., Prianichnikov, N. A., Rands, C. M., Zdobnov, E. M., Ilina, E. N., *et al*. (2020). Phigaro: high-throughput prophage sequence annotation. *Bioinformatics*, 36(12), 3882-3884.

Stringer, A. M., Baniulyte, G., Lasek-Nesselquist, E., Seed, K. D., & Wade, J. T. (2020). Transcription termination and antitermination of bacterial CRISPR arrays. *Elife*, 9.

Suar, M., Jantsch, J., Hapfelmeier, S., Kremer, M., Stallmach, T., Barrow, P. A., *et al*. (2006). Virulence of broad- and narrow-host-range *Salmonella enterica* serovars in the streptomycin-pretreated mouse model. *Infect Immun*, 74(1), 632-644.

Sun, D., X. Mao, M. Fei, Z. Chen, T. Zhu & J. Qiu (2020). Histone-like Nucleoid-Structuring Protein (H-NS) Paralogue StpA Activates the Type I-E CRISPR-Cas System against Natural Transformation in *Escherichia coli*. *Appl Environ Microbiol*, 86.

Szklarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., *et al*. (2023). The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res*, 51(D1), D638-D646.

Tamura, K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Mol Biol Evol*, 9(4), 678-687.

Tang, B., Gong, T., Zhou, X., Lu, M., Zeng, J., Peng, X., *et al*. (2019). Deletion of *cas3* gene in *Streptococcus* mutants affects biofilm formation and increases fluoride sensitivity. *Arch Oral Biol*, 99, 190-197.

Tanmoy, A. M., Saha, C., Sajib, M. S. I., Saha, S., Komurian-Pradel, F., van Belkum, A., *et al*. (2020). CRISPR-Cas Diversity in Clinical *Salmonella enterica* serovar Typhi Isolates from South Asian Countries. *Genes* (Basel), 11(11).

Tanner, J. R., & Kingsley, R. A. (2018). Evolution of *Salmonella* within Hosts. *Trends Microbiol*, 26(12), 986-998.

Tao, S., Chen, H., Li, N., & Liang, W. (2022). The Application of the CRISPR-Cas System in Antibiotic Resistance. *Infect Drug Resist*, 15, 4155-4168.

Tesson, F., Hervé, A., Mordret, E., Touchon, M., d'Humières, C., Cury, J., *et al*. (2022). Systematic and quantitative view of the antiviral arsenal of prokaryotes. *Nature Communications*, 13(1), 2561.

Thompson, C. P., Doak, A. N., Amirani, N., Schroeder, E. A., Wright, J., Kariyawasam, S., *et al*. (2018). High-Resolution Identification of Multiple *Salmonella* serovars in a Single Sample by Using CRISPR-SeroSeq. *Appl Environ Microbiol*, 84(21).

Timme, R. E., Pettengill, J. B., Allard, M. W., Strain, E., Barrangou, R., Wehnes, C., *et al*. (2013). Phylogenetic diversity of the enteric pathogen *Salmonella enterica* subsp. *enterica* inferred from genome-wide reference-free SNP characters. *Genome Biol Evol*, 5(11), 2109-2123.

Touchon, M. & E. P. Rocha (2010). The small, slow and specialized CRISPR and anti-CRISPR of *Escherichia* and *Salmonella. PLoS One*, 5, e11126.

Townsend, S. M., Kramer, N. E., Edwards, R., Baker, S., Hamlin, N., Simmonds, M., *et al*. (2001). *Salmonella enterica* serovar Typhi possesses a unique repertoire of fimbrial gene sequences. *Infect Immun*, 69(5), 2894-2901.

Turcotte, M. R., Smith, J. T., Li, J., Zhang, X., Wolfe, K. L., Gao, F., *et al*. (2022). Genome characteristics of clinical *Salmonella enterica* population from a state public health laboratory, New Hampshire, USA, 2017-2020. *BMC Genomics*, 23(1), 537.

Tzeng, S. R., Huang, Y. W., Zhang, Y. Q., Yang, C. Y., Chien, H. S., Chen, Y. R., *et al*. (2020). A Celecoxib Derivative Eradicates Antibiotic-Resistant *Staphylococcus aureus* and Biofilms by Targeting YidC2 Translocase. *Int J Mol Sci*, 21(23).

Uribe, R. V., C. Rathmer, L. J. Jahn, M. M. H. Ellabaan, S. S. Li & M. O. A. Sommer (2021). Bacterial resistance to CRISPR-Cas antimicrobials. *Sci Rep*, 11, 17267.

Urrutia, A. O., & Hurst, L. D. (2003). The signature of selection mediated by expression on human genes. *Genome Res*, 13(10), 2260-2264.

Urrutia, I. M., Fuentes, J. A., Valenzuela, L. M., Ortega, A. P., Hidalgo, A. A., & Mora, G. C. (2014). *Salmonella* Typhi *shdA*: pseudogene or allelic variant? *Infect Genet Evol*, 26, 146-152.

Uzzau, S., Leori, G. S., Petruzzi, V., Watson, P. R., Schianchi, G., Bacciu, D., *et al*. (2001). *Salmonella enterica* serovar-host specificity does not correlate with the magnitude of intestinal invasion in sheep. *Infect Immun*, 69(5), 3092-3099.

V. T. Nair, D., Venkitanarayanan, K., & Kollanoor Johny, A. (2018). Antibiotic-Resistant *Salmonella* in the Food Supply and the Potential Role of Antibiotic Alternatives for Control. *Foods*, 7(10), 167.

Vaid, R. K., Thakur, Z., Anand, T., Kumar, S., & Tripathi, B. N. (2021). Comparative genome analysis of *Salmonella enterica* serovar Gallinarum biovars Pullorum and Gallinarum decodes strain specific genes. *PLoS One*, 16(8), e0255612.

van Belkum, A., Soriaga, L. B., LaFave, M. C., Akella, S., Veyrieras, J. B., Barbu, E. M., *et al*. (2015). Phylogenetic Distribution of CRISPR-Cas Systems in Antibiotic-Resistant *Pseudomonas aeruginosa*. *mBio*, 6(6), e01796-01715.

Vangay, P., Ward, T., Gerber, J. S., & Knights, D. (2015). Antibiotics, pediatric dysbiosis, and disease. *Cell Host Microbe*, 17(5), 553-564.

Vázquez, X., García, V., Fernández, J., Bances, M., de Toro, M., Ladero, V., *et al*. (2021). Colistin Resistance in Monophasic Isolates of *Salmonella enterica* ST34 Collected From Meat-Derived Products in Spain, With or Without CMY-2 Co-production. *Front Microbiol*, 12, 735364.

Vercoe, R. B., Chang, J. T., Dy, R. L., Taylor, C., Gristwood, T., Clulow, J. S., *et al*. (2013). Cytotoxic chromosomal targeting by CRISPR/Cas systems can reshape bacterial genomes and expel or remodel pathogenicity islands. *PLoS Genet*, 9(4), e1003454.

Vernikos, G. S., & Parkhill, J. (2006). Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics*, 22(18), 2196-2203.

Vila Nova, M., Durimel, K., La, K., Felten, A., Bessières, P., Mistou, M. Y., *et al*. (2019). Genetic and metabolic signatures of *Salmonella enterica* subsp. *enterica* associated with animal sources at the pangenomic scale. *BMC Genomics*, 20(1), 814.

Vishnoi, A., Kryazhimskiy, S., Bazykin, G. A., Hannenhalli, S., & Plotkin, J. B. (2010). Young proteins experience more variable selection pressures than old proteins. *Genome Res*, 20(11), 1574-1581.

Wahl, A., Battesti, A., & Ansaldi, M. (2019). Prophages in *Salmonella enterica*: a driving force in reshaping the genome and physiology of their bacterial host? *Mol Microbiol*, 111(2), 303-316.

Wang, J., Li, J., Zhao, H., Sheng, G., Wang, M., Yin, M., *et al*. (2015). Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. *Cell*, 163(4), 840-853.

Wang, M., Qazi, I. H., Wang, L., Zhou, G., & Han, H. (2020). Virulence and Immune Escape. *Microorganisms*, 8(3).

Wang, X., Quinn, P. J., & Yan, A. (2015). Kdo2-lipid A: structural diversity and impact on immunopharmacology. *Biol Rev Camb Philos Soc*, 90(2), 408-427.

Wang, X., S. Biswas, N. Paudyal, H. Pan, X. Li, W. Fang & M. Yue (2019). Antibiotic Resistance in *Salmonella* Typhimurium Isolates Recovered From the Food Chain Through National Antimicrobial Resistance Monitoring System Between 1996 and 2016. *Front Microbiol*, 10, 985.

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., *et al*. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic acids research*, 42 (Database issue), D581-D591.

Wattiau, P., Boland, C., & Bertrand, S. (2011). Methodologies for *Salmonella enterica* subsp. *enterica* subtyping: gold standards and alternatives. *Appl Environ Microbiol*, 77(22), 7877-7885.

Westra, E. R., U. Pul, N. Heidrich, M. M. Jore, M. Lundgren, T. Stratmann, R. Wurm, A. Raine, M. Mescher, L. Van Heereveld, M. Mastop, E. G. Wagner, K. Schnetz, J. Van Der Oost, R. Wagner & S. J. Brouns (2010). H-NS-mediated repression of CRISPR-based immunity in *Escherichia coli* K12 can be relieved by the transcription activator LeuO. *Mol Microbiol*, 77, 1380-93.

Westra, E. R., van Erp, P. B., Künne, T., Wong, S. P., Staals, R. H., Seegers, C. L., *et al*. (2012). CRISPR immunity relies on the consecutive binding and degradation of negatively supercoiled invader DNA by Cascade and Cas3. *Mol Cell*, 46(5), 595-605.

White, A. P., Gibson, D. L., Collinson, S. K., Banser, P. A., & Kay, W. W. (2003). Extracellular polysaccharides associated with thin aggregative fimbriae of *Salmonella enterica* serovar Enteritidis. *J Bacteriol*, 185(18), 5398-5407.

Wiedemann, A., Mijouin, L., Ayoub, M. A., Barilleau, E., Canepa, S., Teixeira-Gomes, A. P., *et al*. (2016). Identification of the epidermal growth factor receptor as the receptor for *Salmonella* Rck–dependent invasion. *The FASEB Journal*, 30(12), 4180-4191.

Wimmer, F., & Beisel, C. L. (2019). CRISPR-Cas Systems and the Paradox of Self-Targeting Spacers. *Front Microbiol*, 10, 3078.

Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V., & Lipman, D. J. (2009). The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci U S A*, 106(18), 7273-7280.

Woudstra, C., & Granier, S. A. (2023). A Glimpse at the Anti-Phage Defenses Landscape in the Foodborne Pathogen. *Viruses*, 15(2).

Wu, Q., Cui, L., Liu, Y., Li, R., Dai, M., Xia, Z., *et al*. (2022). CRISPR-Cas systems target endogenous genes to impact bacterial physiology and alter mammalian immune responses. *Mol Biomed*, 3(1), 22.

Wu, Y., Battalapalli, D., Hakeem, M. J., Selamneni, V., Zhang, P., Draz, M. S., *et al*. (2021). Engineered CRISPR-Cas systems for the detection and control of antibiotic-resistant infections. *J Nanobiotechnology,* 19(1), 401.

Wu, Y., Garushyants, S. K., Hurk, A. v. d., Aparicio-Maldonado, C., Kushwaha, S. K., King, C. M., *et al*. (2023). Synergistic anti-phage activity of bacterial defence systems. *bioRxiv*, 2022.2008.2021.504612.

Xiang, Y., Li, F., Dong, N., Tian, S., Zhang, H., Du, X., *et al*. (2020). Investigation of a Salmonellosis Outbreak Caused by Multidrug Resistant *Salmonella* Typhimurium in China. *Front Microbiol*, 11, 801.

Xie, X., Hu, Y., Xu, Y., Yin, K., Li, Y., Chen, Y., *et al*. (2017). Genetic analysis of *Salmonella enterica* serovar Gallinarum biovar Pullorum based on characterization and evolution of CRISPR sequence. *Vet Microbiol*, 203, 81-87.

Xu, X., Biswas, S., Gu, G., Elbediwi, M., Li, Y., & Yue, M. (2020). Characterization of Multidrug Resistance Patterns of Emerging *Salmonella enterica* serovar Rissen along the Food Chain in China. *Antibiotics*, 9(10), 660.

Xue, C., & Sashital, D. G. (2019). Mechanisms of Type I-E and I-F CRISPR-Cas Systems in *Enterobacteriaceae*. *EcoSal Plus*, 8(2).

Yamada, S., Shibasaki, M., Murase, K., Watanabe, T., Aikawa, C., Nozawa, T., *et al*. (2019). Phylogenetic relationship of prophages is affected by CRISPR selection in Group A *Streptococcus*. *BMC Microbiol*, 19(1), 24.

Yang, C., Li, P., Su, W., Li, H., Liu, H., Yang, G., *et al*. (2015). Polymorphism of CRISPR shows separated natural groupings of *Shigella* subtypes and evidence of horizontal transfer of CRISPR. *RNA Biol*, 12(10), 1109-1120.

Yang, Q. E., Agouri, S. R., Tyrrell, J. M., & Walsh, T. R. (2018). Heavy Metal Resistance Genes Are Associated with bla$_{\text{NDM-1-}}$ and bla$_{\text{CTX-M-15-}}$ Carrying *Enterobacteriaceae*. *Antimicrobial Agents and Chemotherapy*, 62(5), 10.1128/aac.02642-02617.

Yu, G. (2020). Using ggtree to Visualize Data on Tree-Like Structures. *Curr Protoc Bioinformatics*, 69(1), e96.

Zablewska, B., & Kur, J. (1995). Mutations in HU and IHF affect bacteriophage T4 growth: HimD subunits of IHF appear to function as homodimers. *Gene*, 160(1), 131-132.

Zakrzewska, M. & M. Burmistrz (2023). Mechanisms regulating the CRISPR-Cas systems. *Front Microbiol*, 14, 1060337.

Zegans, M. E., Wagner, J. C., Cady, K. C., Murphy, D. M., Hammond, J. H., & O'Toole, G. A. (2009). Interaction between bacteriophage DMS3 and host CRISPR region inhibits group behaviors of *Pseudomonas aeruginosa*. *J Bacteriol*, 191(1), 210-219.

Zhang, K., Zhang, Y., Wang, Z., Li, Y., Xu, H., Jiao, X., *et al*. (2021). Characterization of CRISPR array in *Salmonella enterica* from asymptomatic people and patients. *Int J Food Microbiol*, 355, 109338.

Zhang, Y., Yang, J., & Bai, G. (2018). Regulation of the CRISPR-Associated Genes by Rv2837c (CnpB) via an Orn-Like Activity in Tuberculosis Complex *Mycobacteria*. *Journal of Bacteriology*, 200(8), 10.1128/jb.00743-00717.

Zheng, J., Luo, Y., Reed, E., Bell, R., Brown, E. W., & Hoffmann, M. (2017). Whole-Genome Comparative Analysis of *Salmonella enterica* serovar Newport Strains Reveals Lineage-Specific Divergence. *Genome Biol Evol*, 9(4), 1047-1050.

APPENDIX

[APPENDIX I]

Isolation of *Salmonella*-specific bacteriophages from the Ganga water

Ganga is known as the holy river of India. Its sanctity is not merely symbolic; it is rooted in the extraordinary biodiversity it hosts, including billions of bacteriophages. Since 1998, phages have been systematically isolated from the Ganga River, specifically targeting pathogenic enteric bacterial species, including *Salmonella*. The inherent bacterial specificity of phages makes them an optimal choice for targeted bacterial eradication, emphasising their significant role in advancing medical interventions. Recent studies underscore the therapeutic promise of phages, employing phage therapy as a potent tool for eliminating bacteria in humans. Additionally, the utilisation of phages for transduction—where genetic material is transferred from one bacterium to another *via* phages—presents a compelling avenue for achieving highly efficient bacterial eradication.

In this pursuit, we aimed to enhance the specificity and efficiency of *Salmonella* elimination using our self-targeting CRISPR array (pQE60-L-I-STS) (**Chapter 5**) by integrating it into a *Salmonella*-specific phage. The idea is to generate a phage vector to specifically eliminate *Salmonella* while introducing the self-targeting CRISPR array. This genetic payload activates the endogenous CRISPR-Cas system, initiating a self-targeting mechanism that effectively eradicates the bacteria by targeting its genome. The combination of these elements aims to improve the precision and efficacy of the bacterial elimination process. In pursuit of this goal, we successfully isolated *Salmonella*-specific bacteriophages from the Ganga River.

Methodology

Water sampling: Water samples were collected from four locations (25.307002, 83.012100; 25.304512, 83.010356; 25.300380, 83.008052; 25.304046, 83.010016) in the Ganga River, Varanasi. Filtered using 0.45 µm PES syringe filters and stored at 4 °C.

Phage isolation: 500 µL of overnight grown *S.* Typhi str. CT18, *S.* Typhimurium str. 14028s, *S.* Paratyphi A, and *S.* Welterveden strains were inoculated in 50 mL LB and incubated for 2 hours. 1 mL of filtered water was added to these cultures, and incubated for 3-4 hours.

Then the samples were filtered with a 0.45 µm PES filter. 1 mL of overnight bacterial culture was inoculated in 50 mL of LB top agar (0.7%), and 4 mL of it was poured over the LB agar plate (2%). 50 µL of filtered culture was spotted in the centre and incubated at 37 °C overnight.

<u>*Phage Purification:*</u> 100 µL of an overnight bacterial culture was mixed with 4 mL of LB top agar. This was evenly poured onto an LB bottom agar plate. A phage plaque from the prior plate was picked using a toothpick, followed by gently piercing the double-layer agar and disseminating the phage using paper strips. The plate was incubated at 37 °C overnight, and the process was repeated until a consistent phage morphology was observed.

<u>*Phage Host Range Testing:*</u> The isolated phages were tested against all four *Salmonella* strains to assess their ability to lyse the bacteria.

Results and Conclusion

Three distinct phages, each exhibiting unique morphologies, were isolated and purified to target *Salmonella* strains (**Fig. 1A**). The phages were subjected to testing against four different *Salmonella* strains. The results revealed that all three phages demonstrated infectivity towards *S.* Typhi str. CT18 and *S.* Paratyphi A. Notably, these phages exhibited an inability to infect *S.* Typhimurium str. 14028s and *S.* Welterveden (**Fig. 1B**). As the isolated phages are specific to serovars Typhi and Paratyphi A they would be ineffective for utilisation in CRISPR-Cas mediated self-targeting.

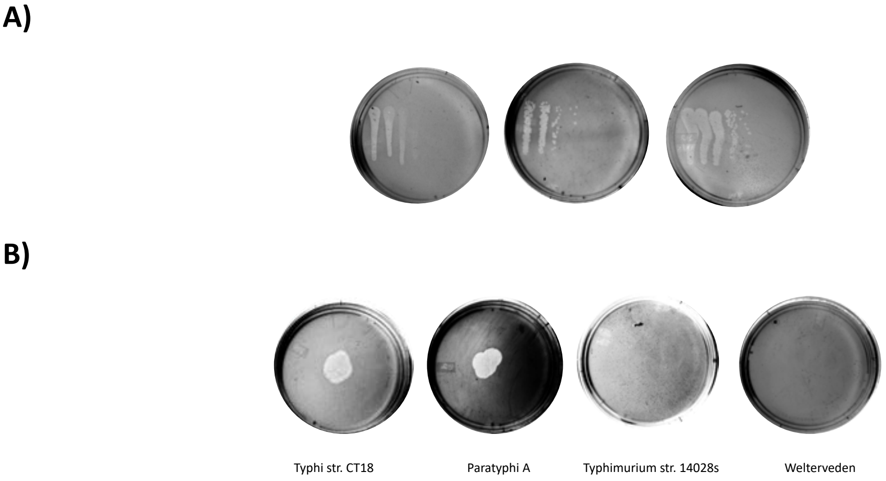


Figure 1 A) Three different morphologies of the obtained phages. B) Phage host range testing. The host range testing involved subjecting the phages to various *Salmonella* strains and visualising bacterial lysis.

[APPENDIX II]

Online Supplementary Data

The additional resources supporting this thesis are accessible at

https://github.com/SimranKushwaha/Exploring-and-Exploiting-Prokaryotic-Immunity-in-Salmonella

Chapter 2

- **An interactive visualisation highlighting 26 identified spots in *Salmonella* RefSeq.**

  Spot 1 – A hotspot for defence systems, particularly Septu type I

  Spot 9 – A hotspot for virulence factors, particularly *sod*

  Spot 11 – A hotspot for defence systems, particularly AbiE and Shango

  Spot 15 – A hotspot for virulence factors, particularly *cdt*

  Spot 17 – A hotspot for stress-resistance genes against gold

  Spot 21 – A hotspot for virulence factors *ste* and *sse*

  Spot 22 – A hotspot for defence systems CRISPR-Cas type I-E

  Spot 30 – A hotspot for virulence factors *lpf*

  Spot 31 – A hotspot for defence systems SEFIR

  Spot 32 – A hotspot for virulence factors *sse*

  Spot 36 – A hotspot for virulence factors, particularly *fae*

  Spot 39 – A hotspot for multiple defence systems

  Spot 43 – A hotspot for multiple defence systems

  Spot 44 – A hotspot for multiple antibiotic resistance genes

  Spot 47 – A hotspot for virulence factor *tcp* and defence system Thoeris

  Spot 51 – A hotspot for multiple antibiotic and stress-resistance genes

  Spot 53 – A hotspot for stress-resistance genes against copper and silver

  Spot 54 – A hotspot for multiple virulence factors

  Spot 63 – A hotspot for multiple defence systems

  Spot 66 – A hotspot for defence system CBASS type I

Spot 68 – A hotspot for defence system RM type I and IV

Spot 79 – A hotspot for virulence factor *ssp*

Spot 89 – A hotspot for defence system RM type III

Spot 92 – A hotspot for virulence factor *sop*

Spot 94 – A hotspot for virulence factor *rat*

Spot 103 – A hotspot for stress-resistance genes against arsenic

- **Microreact project 1**, which presents metadata for isolates, phylogenetic analysis, and the country of isolation of *Salmonella* RefSeq.
- **Microreact project 2**, illustrating the serovar-wise distribution of the pathogenic determinants on *Salmonella* RefSeq.
- **Supplementary tables** 2.1 to 2.13-

  S2.1. Features of the 12,244 *Salmonella* genomes analysed in this study.

  S2.2. Features of the plasmids found across *Salmonella* RefSeq.

  S2.3. Prevalence of plasmid incompatibility groups across *Salmonella* RefSeq.

  S2.4. Features of the prophages found across *Salmonella* RefSeq.

  S2.5. List of pathogenicity genes analysed in this study.

  S2.6. Distribution and location of pathogenic determinants in *Salmonella* RefSeq.

  S2.7. Average of plasmids, prophages, and pathogenic determinants in *Salmonella*.

  S2.8. Prevalence (%) of pathogenic determinants across plasmids.

  S2.9. Core, persistent, shell and cloud genes and their predicted function.

  S2.10. Regions of genomic plasticity identified in *Salmonella* RefSeq.

  S2.11. RGP families and gene count in *Salmonella* spots.

  S2.12. Pathogenic determinants present on spots of integration in *Salmonella*.

  S2.13. Flanking genes defining the integration spot and their predicted function.

Chapter 3

- The **spacer arrangement** of 133 strains belonging to 26 serovars.
- The **supplementary table** provides a detailed list of all strains utilised in the study.

Chapter 4

- **Interactive network** showcasing the interactions between potential regulatory spacers and their gene targets for *Salmonella* serovars Typhi, Typhimurium and Enteritidis.

[APPENDIX III]

List of Publications, Quests, Conferences and Funding

Publication from Ph.D. Thesis

1. **Kushwaha, S.K.,** Bhavesh, N.L.S., Abdella, B., Lahiri, C., and Marathe, S.A. (2020). The phylogenomics of CRISPR-Cas system and revelation of its features in *Salmonella*. Sci Rep 10, 21156. 10.1038/s41598-020-77890-6.
2. **Kushwaha, S.K.,** Narasimhan, L.P., Chithananthan, C., and Marathe, S.A. (2022). Clustered regularly interspaced short palindromic repeats-Cas system: diversity and regulation in *Enterobacteriaceae*. Future Microbiol 17, 1249-1267. 10.2217/fmb-2022-0081.
3. **Kushwaha, S.K.,** Kumar, A.A., Gupta, H., and Marathe, S.A. (2023). The Phylogenetic Study of the CRISPR-Cas System in *Enterobacteriaceae*. Curr Microbiol 80, 196. 10.1007/s00284-023-03298-w.
4. **Kushwaha, S.K.,** Anand, A., Wu, Y., Avila, H.L., Sicheritz-Ponten, T., Millard, A., Marathe, S.A., and Nobrega, F.L. (2023). Genomic plasticity is a blueprint of diversity in *Salmonella* lineages. bioRxiv, 2023.2012.2002.569618. 10.1101/2023.12.02.569618 (Under revision in PLOS Biology).
5. **Kushwaha, S.K.,** Venkateswaran, S, and Marathe, S.A. Analysing self-targeting CRISPR spacers in *Salmonella* to understand their role in endogenous gene regulation (Article under preparation).

Other Publications

1. Wu Y., Garushyants S.K., van den Hurk A., Aparicio-Maldonado C., **Kushwaha S.K.**, King C.M., Ou Y., Todeschini T.C., Clokie M.R.J., Millard A.D., Gençay, Y.E., Koonin, E,V., and Nobrega, F.L. (2024). Bacterial defense systems exhibit synergistic anti-phage activity. Cell Host Microbe, 10.1016/j.chom.2024.01.015
2. Gambino, M., **Kushwaha, S.K.**, Wu, Y., Beajoui, S., Jensen, C.M., Bojer, M.S., Lutz, M., Klein-Sousa, V., Taylor, N.M.I., Song, W., Xiao, M., Nobrega, F.L, Brøndsted, L. Determinants of phage host range in porcine enterotoxigenic *Escherichia coli* (Under peer review in Applied and Environmental Microbiology).

3. Rothschild-Rodríguez, D., **Kushwaha, S.K.,** Hedges, M., King, C.M., Lawson, S., Wand, M., Sutton, M., and Nobrega, F.L. An OPEN collection of phages targeting *Klebsiella* spp. for your research. https://www.klebphacol.org (Article under preparation).

Quests

1. Received special recommendations for developing "CRISPR-mediated antimicrobials" at the Antimicrobial Resistance Quest organised by C-CAMP, India, and CARB-X in April 2021.

Conferences Attended

1. **Kushwaha, S.K.,** Nobrega, F.L., and Marathe, S.A. "Repurposing native CRISPR-Cas system as therapeutic against *Salmonella".* In- iCRISPR 2021, organised by SRM Institute, India, November 25th-27th, 2021.
2. **Kushwaha, S.K.,** Marathe, S.A., and Nobrega, F.L. "Repurposing CRISPR against *Salmonella* spp*".* In- Oxford Bacteriophage Conference- Phages 2022, in Oxford, UK, September 5th-6th, 2022.
3. **Kushwaha, S.K.,** Anand, A., Avila, H.L., Wu, Y., Marathe, S.A., and Nobrega, F.L. "*Salmonella* displays functional specialisation in its regions of genomic plasticity". In- International Symposium 2023, New concepts in prokaryotic virus-host interactions, in Berlin, Germany, October 2nd-4th, 2023.

Funding

1. Received the esteemed "Newton Bhabha PhD Placement Fund", awarded by the British Council, UK, and the Department of Biotechnology, India, enabling research in the UK.
2. Received the DST-SERB "International Travel Scheme" grant for attending the International Symposium 2023, New concepts in prokaryotic virus-host interactions, in Berlin, Germany.

[APPENDIX IV]

Biography of Prof. Sandhya Amol Marathe

Prof. Sandhya Amol Marathe is working as an Associate Professor in the Department of Biological Sciences, Birla Institute of Technology and Science Pilani (BITS-Pilani), Pilani campus, Rajasthan, since April 2017. She obtained her Bachelor's degree from Pune University. She completed her MSc. - Ph.D. integrated doctoral degree from the Indian Institute of Science (IISc), Bangalore, in the area of Infection Biology. She received the Best Thesis award, MCB, IISc, in 2013.

After her Ph.D., she worked as a research associate at MCB, IISc. She worked as a visiting Assistant Professor at BITS-Pilani for four years starting in July 2013. Her broad areas of research interest include bacterial pathogenesis and host-pathogen interaction. Prof. Marathe has completed one research project funded by SERB-DST as Co-Principal Investigator and one as principal investigator. Currently, she has three research projects as a project coordinator and a principal investigator funded by DBT, SERB POWER Grant (approved) and ICMR (approved). She has published more than 29 research articles in peer-reviewed journals. She has successfully guided several undergraduate and postgraduate students in their research studies. She has successfully supervised the thesis of one Ph.D. student and co-supervised another Ph.D. student. Currently, she is supervising three more students for their Ph.D. studies.

Biography of Prof. Franklin L. Nobrega

Prof. Franklin L. Nobrega is a Microbiology Associate Professor at the School of Biological Sciences, University of Southampton, a position he has held since July 2020. He earned his Ph.D. from Wageningen Universiteit in the Netherlands in 2017. Following his doctoral studies, he pursued a Postdoctoral role at the Kavli Institute of Nanoscience in Delft, Netherlands. His academic journey also includes roles as a scientific consultant at SNIPR Biome in Copenhagen, Denmark, and at BGI: Shenzhen in Guangdong, China.

Prof. Nobrega's research team focuses on the arms race between bacteria and their viruses, the bacteriophages, from a biological, ecological, and therapeutic perspective. They seek to understand the impact of bacteriophages in shaping natural microbial communities, particularly their role in the evolution of defence and anti-defence mechanisms, and their capacity to modulate bacterial metabolism, especially in biofilm and gut communities. They also work to develop innovative phage therapy approaches to fight antibiotic-resistant bacterial infections.

Prof. Nobrega's research is supported by an array of funding sources, including the Royal Society, UK; King Abdullah International Medical Research Center, Saudi Arabia; Bowel Research, UK; University of Southampton, UK; Wessex Medical Research, UK; IfLS Research Stimulus Fund, UK; University Hospital Southampton NHS Foundation Trust, UK; ZonMw, Netherlands; and Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Netherlands.

Prof. Nobrega's impactful contributions extend to his publication record, which boasts over 35 research papers published in esteemed peer-reviewed journals. Furthermore, he actively participates as a peer reviewer for several notable publications and grants, including EMBO Journal, F1000 Research, ISME Communications, Microbiome, and Mobile DNA.

Guiding the next generation of scientists is also a key facet of Prof. Nobrega's academic engagement. Having successfully supervised the thesis work of one Ph.D. student, he is currently supervising the research of seven more students pursuing their doctoral studies.

Biography of Simran Krishnakant Kushwaha

Ms Simran Krishnakant Kushwaha earned her Bachelor of Engineering in Biotechnology from Mumbai University in 2018. Her pursuit of doctoral studies led her to join the Birla Institute of Technology and Science, Pilani, in January 2019. Simran's research interest is understanding all aspects of the *Salmonella* bacteria, including its epidemiology, antimicrobial resistance, defence system against viruses, molecular biology, and host-pathogen interaction. She aims to identify new avenues for developing effective treatments against the ESKAPE pathogens using the CRISPR-Cas system.

During her undergraduate degree, Simran was honoured as the Valedictorian for securing the top position at Mumbai University. She has also achieved an All India Rank of 754 in GATE Biotechnology, 2018.

During her Ph.D. journey, she was awarded the esteemed sponsorship Newton Bhabha Placement Fund by the British Council, UK & the Department of Biotechnology, India, for enabling research in the UK. She also received special recommendations for developing CRISPR-mediated antimicrobials at the AMR Quest organised by C-CAMP, India, & CARB-X in 2021. Simran has presented her research work at various conferences in India, the UK and Germany and established research collaborations worldwide.

Simran has mentored more than ten students for their study projects and master theses and has been consistently involved in teaching microbiology & genetic engineering labs for undergraduate students.

[APPENDIX V]

Reprint of Publications

scientific reports

OPEN

The phylogenomics of CRISPR-Cas system and revelation of its features in *Salmonella*

Simran Krishnakant Kushwaha[1]✉, Narra Lakshmi Sai Bhavesh[1,4], Bahaa Abdella[2,3,4], Chandrajit Lahiri[2] & Sandhya Amol Marathe[1]✉

Salmonellae **display intricate evolutionary patterns comprising over 2500 serovars having diverse pathogenic profiles. The acquisition and/or exchange of various virulence factors influences the evolutionary framework. To gain insights into evolution of** ***Salmonella*** **in association with the CRISPR-Cas genes we performed phylogenetic surveillance across strains of 22** ***Salmonella*** **serovars. The strains differed in their CRISPR1-leader and** ***cas*** **operon features assorting into two main clades, CRISPR1-STY/*****cas*****-STY and CRISPR1-STM/*****cas*****-STM, comprising majorly typhoidal and non-typhoidal** ***Salmonella*** **serovars respectively. Serovars of these two clades displayed better relatedness, concerning CRISPR1-leader and** ***cas*** **operon, across genera than between themselves. This signifies the acquisition of CRISPR1/Cas region could be through a horizontal gene transfer event owing to the presence of mobile genetic elements flanking CRISPR1 array. Comparison of CRISPR and** ***cas*** **phenograms with that of multilocus sequence typing (MLST) suggests differential evolution of CRISPR/Cas system. As opposed to broad-host-range, the host-specific serovars harbor fewer spacers. Mapping of protospacer sources suggested a partial correlation of spacer content with habitat diversity of the serovars. Some serovars like serovar Enteritidis and Typhimurium that inhabit similar environment/infect similar hosts hardly shared their protospacer sources.**

Genus *Salmonella* is classified into two species, *Salmonella enterica (S. enterica)* and *S. bongori*. *S. enterica* evolved into six subspecies (subsp.) namely, *enterica*, *salamae*, *arizonae*, *diarizonae*, *houtenae* and *indica*[1]. The host-range for serovars of *S. enterica* subsp. *enterica* vary from broad-host-range to host-adapted and host-restricted[2] pertinent to within-host evolution[3]. Before divergence, *S. bongori* and *S. enterica* acquired *Salmonella* pathogenicity island 1 (SPI-1)[4] and later *S. enterica* laterally acquired SPI-2 thereby, enhancing its virulence potential[4]. As per the adopt-adapt model of bacterial speciation[5], the adopted lateral gene(s) divert the evolutionary path promoting bacterial adaptation and consequently increasing its fitness[6]. Over time, both species horizontally acquired multiple virulence factors progressively enhancing their pathogenicity[3].

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and a set of CRISPR-associated (*cas*) genes are suggested to be acquired by horizontal gene transfer (HGT) event[7,8]. The Cas1 and Cas2 proteins are essential for spacer acquisition from invading mobile genetic elements (MGE)[1] while all Cas proteins participate in primed adaptation to update the invaders' memory[9]. The newly acquired spacers are added at the leader proximal end of the CRISPR array[1]. Cas proteins work in conjunction with the CRISPR-RNA to carry out the interference step[2]. CRISPR-Cas system has been related to the bacterial virulence potential[10–13]. The number of CRISPR array are negatively correlated with pathogenic potential of *Escherichia coli* where, the reduction in CRISPR activity is proposed to promote HGT favouring its evolution[14]. Conversely, some reports demonstrate a positive correlation between the CRISPR and pathogenicity owing to virulence genes regulation[10,13,15]. In *S. enterica* subsp. *enterica* serovar Enteritidis, Cas3 modulates biofilm formation and virulence by regulating quorum sensing genes[13]. Further, in *Salmonella* and *E. coli*, 53% of CRISPR protospacers traced to chromosomes[8] suggesting a potential role of the CRISPR-Cas system in endogenous gene regulation[16] and possibly pathogenesis[13].

S. enterica possesses type I–E CRISPR system comprising a *cas* operon and two CRISPR arrays, CRISPR1 and CRISPR2[17], separated by ~ 16 kb[18]. The *cas* operon present in proximity to the CRISPR1 array[19] contains 8 *cas* genes. Two distinct *cas* gene profiles has been observed with reported incongruence between the *cas* and whole

[1]Department of Biological Sciences, Birla Institute of Technology and Science (BITS), Pilani, Rajasthan, India. [2]Department of Biological Sciences, Sunway University, Petaling Jaya, Selangor, Malaysia. [3]Faculty of Aquatic and Fisheries Sciences, Kafrelsheikh University, Kafrelsheikh, Egypt. [4]These authors contributed equally: Narra Lakshmi Sai Bhavesh and Bahaa Abdella. ✉email: p20180406@pilani.bits-pilani.ac.in; sandhya.marathe@pilani.bits-pilani.ac.in

Future
MICROBIOLOGY

Review

For reprint orders, please contact: reprints@futuremedicine.com

Clustered regularly interspaced short palindromic repeats-Cas system: diversity and regulation in Enterobacteriaceae

Simran K Kushwaha[1], Lakshmi P Narasimhan[‡,1], Chandrananthi Chithananthan[‡,1] & Sandhya A Marathe[*,1]

[1] Department of Biological Sciences, Birla Institute of Technology & Science (BITS), Pilani, Rajasthan, 333031, India

[*] Author for correspondence: Tel.:+91 159 625 5614; sandhya.marathe@pilani.bits-pilani.ac.in

[‡] Authors contributed equally.

Insights into the arms race between bacteria and invading mobile genetic elements have revealed the intricacies of the clustered regularly interspaced short palindromic repeats (CRISPR)-Cas system and the counter-defenses of bacteriophages. Incredible spacer diversity but significant spacer conservation among species/subspecies dictates the specificity of the CRISPR-Cas system. Researchers have exploited this feature to type/subtype the bacterial strains, devise targeted antimicrobials and regulate gene expression. This review focuses on the nuances of the CRISPR-Cas systems in Enterobacteriaceae that predominantly harbor type I-E and I-F CRISPR systems. We discuss the systems' regulation by the global regulators, H-NS, LeuO, LRP, cAMP receptor protein and other regulators in response to environmental stress. We further discuss the regulation of noncanonical functions like DNA repair pathways, biofilm formation, quorum sensing and virulence by the CRISPR-Cas system. The review comprehends multiple facets of the CRISPR-Cas system in Enterobacteriaceae including its diverse attributes, association with genetic features, regulation and gene regulatory mechanisms.

First draft submitted: 1 April 2022; Accepted for publication: 22 July 2022; Published online: 00 August 2022

Keywords: CRISPR-Cas diversity • CRISPR-Cas noncanonical functions • CRISPR-Cas regulation • Enterobacteriaceae

Prokaryotic viruses (phages) are the most copious forms of biological life on Earth [1]. Bacteria and viruses often occupy the same niches and can remarkably defy each other [2]. Bacteria are equipped with various defense mechanisms, including restriction-modification systems and the sugar-nonspecific nucleases to degrade the invading mobile genetic elements (MGE) [3,4]. In 1987, Ishino *et al.* identified an 'unusual structure' at the 3′ end of the *iap* gene locus of *Escherichia coli* [5]. The structure was subsequently named clustered regularly interspaced short palindromic repeats (CRISPR), and the ancillary proteins were termed as CRISPR-associated proteins (Cas). Later, the system was proposed to act as guardians of the bacterial genome, regulating the tolerance of bacteria against environmental stresses and MGE attacks [6].

The CRISPR-Cas system prevails in ∼90% archaea and 30–40% bacteria [7,8], consisting of three critical attributes: a set of *cas* genes, a leader sequence and a succeeding CRISPR array [6]. The CRISPR array comprises partially palindromic direct repeat (DR) [9] and the spacers. Generally, the spacers are derived from MGEs like the bacteriophages and plasmids when they first invade the bacteria [6,10,11]. According to the 2019 classification of the CRISPR-Cas system by Makarova *et al.*, the system is highly diverse and categorized into two classes, six types and 33 subtypes [12]. The CRISPR-Cas system belonging to both the classes (class 1 and class 2) have been utilized for multiple applications ranging from gene manipulations, diagnostics and antimicrobial therapy to recombinant protein production [13–23]. Toward the end of this review, we briefly discuss some of these applications pertaining to the Enterobacteriaceae CRISPR-Cas system.

The CRISPR-Cas system and its mechanisms have been thoroughly explored within members of the Enterobacteriaceae family. Medina-Aparicio *et al.* and Xue and Sashital have discussed characteristics and mechanistic understandings of this adaptive immune system in their respective reviews [24,25]. Our review provides a detailed

Future
Medicine

Current Microbiology (2023) 80:196
https://doi.org/10.1007/s00284-023-03298-w

The Phylogenetic Study of the CRISPR-Cas System in *Enterobacteriaceae*

Simran Krishnakant Kushwaha[1] · Aryahi A. Kumar[1] · Hardik Gupta[1] · Sandhya Amol Marathe[1]

Received: 29 December 2022 / Accepted: 7 April 2023 / Published online: 28 April 2023

Abstract

The Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-CRISPR-associated (Cas) system is a bacterial and archaeal adaptive immune system undergoing rapid multifaceted evolution. This evolution plausibly occurs due to the genetic exchanges of complete loci or individual entities. Here, we systematically investigate the evolutionary framework of the CRISPR-Cas system in six *Enterobacteriaceae* species and its evolutionary association with housekeeping genes as determined by the *gyrB* phenogram. The strains show high variability in the *cas3* gene and the CRISPR1 locus among the closely related *Enterobacteriaceae* species, hinting at a series of genetic exchanges. The CRISPR leader is conserved, especially toward the distal end, and could be a core region of the leader. The spacers are conserved within the strains of most species, while some strains show unique sets of spacers. However, inter-species spacer conservation was rarely observed. For a considerable proportion of these spacers, protospacer sources were not detected. These results advance our understanding of the dynamics of the CRISPR-Cas system; however, the biological functions are yet to be characterised.

✉ Sandhya Amol Marathe
sandhya.marathe@pilani.bits-pilani.ac.in

[1] Department of Biological Sciences, Faculty Division-III, Birla Institute of Technology & Science, 3277-B, Pilani Campus, Pilani, Rajasthan 333031, India

Genomic plasticity is a blueprint of diversity in *Salmonella* lineages

Simran Krishnakant Kushwaha[1,2], Abhirath Anand[3], Yi Wu[2], Hugo Leonardo Ávila[4], Thomas Sicheritz-Pontén[5], Andrew Millard[6], Sandhya Amol Marathe[1], and Franklin L. Nobrega[2*]

[1] Department of Biological Sciences, Birla Institute of Technology & Science (BITS), Pilani, Rajasthan, India

[2] School of Biological Sciences, University of Southampton, Southampton, United Kingdom

[3] Department of Computer Sciences and Information Systems, Birla Institute of Technology & Science (BITS), Pilani, Rajasthan, India

[4] Laboratory for Applied Science and Technology in Health, Instituto Carlos Chagas, FIOCRUZ Paraná, Brazil

[5] Center for Evolutionary Hologenomics, Globe Institute, University of Copenhagen, Denmark, Centre of Excellence for Omics-Driven Computational Biodiscovery (COMBio), AIMST University, Bedong 08100, Kedah, Malaysia

[6] Centre for Phage Research, Department of Genetics and Genome Biology, University of Leicester, Leicester, United Kingdom

*Correspondence: F.Nobrega@soton.ac.uk

Cell Host & Microbe

CellPress
OPEN ACCESS

Article

Bacterial defense systems exhibit synergistic anti-phage activity

Yi Wu,[1,7] Sofya K. Garushyants,[2,7] Anne van den Hurk,[1] Cristian Aparicio-Maldonado,[1] Simran Krishnakant Kushwaha,[1,3] Claire M. King,[1] Yaqing Ou,[4] Thomas C. Todeschini,[1] Martha R.J. Clokie,[5] Andrew D. Millard,[5] Yilmaz Emre Gençay,[6] Eugene V. Koonin,[2] and Franklin L. Nobrega[1,8,*]

[1]School of Biological Sciences, University of Southampton, Southampton SO17 1BJ, UK
[2]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
[3]Department of Biological Sciences, Birla Institute of Technology and Science (BITS), Pilani, Rajasthan, India
[4]Wellcome Centre for Cell-Matrix Research, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK
[5]Department of Genetics and Genome Biology, University of Leicester, Leicester, UK
[6]SNIPR biome, Copenhagen, Denmark
[7]These authors contributed equally
[8]Lead contact
*Correspondence: f.nobrega@soton.ac.uk
https://doi.org/10.1016/j.chom.2024.01.015

SUMMARY

Bacterial defense against phage predation involves diverse defense systems acting individually and concurrently, yet their interactions remain poorly understood. We investigated >100 defense systems in 42,925 bacterial genomes and identified numerous instances of their non-random co-occurrence and negative association. For several pairs of defense systems significantly co-occuring in *Escherichia coli* strains, we demonstrate synergistic anti-phage activity. Notably, Zorya II synergizes with Druantia III and ietAS defense systems, while tmn exhibits synergy with co-occurring systems Gabija, Septu I, and PrrC. For Gabija, tmn coopts the sensory switch ATPase domain, enhancing anti-phage activity. Some defense system pairs that are negatively associated in *E. coli* show synergy and significantly co-occur in other taxa, demonstrating that bacterial immune repertoires are largely shaped by selection for resistance against host-specific phages rather than negative epistasis. Collectively, these findings demonstrate compatibility and synergy between defense systems, allowing bacteria to adopt flexible strategies for phage defense.

INTRODUCTION

Bacteria evolved numerous, diverse lines of active immunity as well as abortive infection mechanisms to withstand phage predation.[1] Recent systematic screening uncovered numerous anti-phage defense systems that widely differ in protein composition and modes of action.[2–7] The mechanisms employed by bacterial defense systems include phage genome or protein sensing followed by degradation,[8–10] introduction of modified nucleotides that abrogate phage replication,[11,12] as well as multiple sensing mechanisms leading to abortive infection that results in the host cell dormancy or death.[4,13–21] However, for many, perhaps, the majority of the bacterial defense systems, the mechanism of action remains unknown.

A bacterial genome carries, on average, about five distinct (currently identifiable) defense systems.[22] The remarkable variability of immune repertoires was observed even within the same species.[22–24] Genes encoding components of these systems tend to cluster together in specific genomic regions known as defense islands, sometimes associated with mobile genetic elements (MGEs) integrated into distinct hotspots in the bacterial genome.[24–26] Defense systems are believed to undergo frequent horizontal transfer between bacteria, and close proximity of the respective genes could facilitate simultaneous transfer of multiple systems.[27]

Despite the recent burst of bacterial defense system discovery, the causes of their clustering in defense islands remain poorly understood. It has been argued that co-localization of defense systems in MGEs and the resulting joint horizontal gene transfer (HGT) could provide fitness advantages to recipient bacteria, especially in phage-rich environments.[28] Additionally, it has been suggested that synergistic interactions between defense systems could drive their co-localization and favor their joint transfer,[29,30] as supported by the conservation of certain sets of defense systems.[31] For example, CRISPR-Cas systems of different subtypes often co-occur and the CRISPR arrays interact with Cas proteins across different systems.[32] Furthermore, toxin-antitoxin (TA) RNA pairs[33] and possibly other TA modules[34] safeguard CRISPR immunity by making cells dependent on CRISPR-Cas for survival. CRISPR-Cas and restriction-modification (RM) systems,[35] as well as BREX and the restriction enzyme BrxU,[30] co-occur resulting in expanded phage protection. However, these