UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Ténica Superior de Ingeniería Agronómica, Alimentaria y de Biosistemas

Interplay of mobile genetic elements and defense systems on the evolution of prokaryotic genomes

D O C T O R A L T H E S I S

Submitted for the degree of Doctor by:

Yang Liu
BSc and MSc in Plant Protection

Madrid, 2024


UNIVERSIDAD POLITÉCNICA DE MADRID
Escuela Ténica Superior de Ingeniería Agronómica,
Alimentaria y de Biosistemas

Doctoral Degree in Biotechnology and Genetic Resources of Plants and Associated Microorganisms

Interplay of mobile genetic elements and defense systems on the evolution of prokaryotic genomes

DOCTORAL THESIS

Submitted for the degree of Doctor by:

Yang Liu
BSc and MSc in Plant Protection

Under the supervision of:
Dr. Jaime Iranzo

Madrid, 2024

Title: Interplay of mobile genetic elements and defense systems on the evolution of prokaryotic genomes

Author: Yang Liu

Doctoral Programme: Programa Oficial de Doctorado en Biotecnología y Recursos Genéticos de Plantas y Microorganismos Asociados

Thesis Supervision:

Dr. Jaime Iranzo, Ramón y Cajal fellow, INIA-CSIC Centro de Biología y Genómica de Plantas(Supervisor)

External Reviewers:

Thesis Defense Committee:

Thesis Defense Date:

This thesis has been partially supported by China Scholarship Council (No.202008440425), European Molecular Biology Organization Scientific Exchange Grant (No.10347), the Ramón y Cajal Programme of the Spanish Ministry of Science (No. RYC-2017–22524); the Agencia Estatal de Investigación of Spain (No. PID2019-106618GA-I00), the Severo Ochoa Programme for Centres of Excellence in R&D of the Agencia Estatal de Investigación of Spain (No. SEV-2016–0672 (2017–2021) to the CBGP); and the Comunidad de Madrid (through a Research Grant for Young Investigators from Universidad Politécnica de Madrid, No. M190020074JIIS).

Acknowledgement

I am so grateful to my advisor, Dr. Jaime Iranzo. Without him, this dissertation would not have been possible. I feel incredibly lucky and privileged to have him as my advisor. He is not only an outstanding scientist, but also an interesting person and a hands-on mentor. He always supports and tolerates me, and guides me in the right direction. He taught me that a good scientist should be rigorous and focused, always valuing their reputation. What I have learned from him in the past three years will be a treasure for life. I am grateful to have met and spent time with Dr. Jorge Calle Espinosa. He is not only a friend and mentor, but also a true scientist whom I aspire to emulate.

I want to express my sincere gratitude to the dissertation committee and providing invaluable advice on my research projects and dissertation. I would also like to extend my thanks to external reviewers.

To Mario Rodríguez Mestre and Dr. Urvish Trivedi, thank you for making my stay in Copenhagen a liberating and enlightening experience. I would also like to thank Dr. Jaime Huerta Cepas and EMBO for supporting this stay.

To my colleagues, Carla Alejandre Villalobos, Adrián López Beltrán, Dr. João Botelho, Dr. Adrián Aguirre Tamaral, Saioa Manzano Morales, Dr. Ziqi Deng, Dr. Lu Yang, Dr. Xi Jiang, and Yingnan Tian. I greatly appreciate their help and support in many aspects, both within and outside the lab.

To my friends: Dr. Runze Feng, Dr. Shiyi Luo, Dr. Zhao Yao, Dr. Qiqi Fu, Dr. Yecheng Zhang, and Luis Alberto Suarez Bracho, thank you for the enjoyable friendship.

To my master's thesis supervisor, Dr. Lian-Hui Zhang, Dr. Stephen Dela Ahator, and Dr. Xiaofan Zhou, for introducing me to academia.

To the staffs at CBGP and ETSIAAB-UPM for their valuable support in creating a conducive working environment

To the CSC-UPM Cooperative PhD Programme for providing financial support during my doctoral studies.

Finally, I must sincerely acknowledge my parents, Jiaxing Liu and Yuhong Chen; my wife, Yiye Han; and my grandparents, Shuang-Zhen Han, Mei-Lan Guo, Ben-Shan Chen, and Si-Zhao Liu. Their endless love and support have been my greatest source of strength, and I am full of gratitude beyond words.

Abstract

Background: Horizontal Gene Transfer (HGT) is a crucial mechanism that drives microbial evolution and adaptation by facilitating rapid genetic change at population and evolutionary levels. The pivotal role of HGT in microbial evolution is exemplified by its major contribution to evolutionary rates through an increase of gene flow. Understanding HGT is central to explaining how microbes adapt to new environments, develop antibiotic resistance, and acquire pathogenicity factors. HGT is most often mediated by mobile genetic elements, such as plasmids and prophages. For that reason, defense systems targeting mobile genetic elements could, in principle, have a notable impact on genomic plasticity. Previous studies attempting to quantify the effect of defense systems on genomic plasticity have produced inconclusive or contradictory results, possibly due to methodological limitations. This thesis revisits this question using state-of-the-art tools and datasets.

Objectives: This thesis revolves around three specific topics: (i) the uncertainty of pangenome analysis in the presence of HGT, (ii) the impact of defense systems on genomic plasticity, and (iii) the interplay between defense systems and mobile genetic elements. The first study focuses on the methodological challenges of gene clustering within prokaryotic pangenomes due to HGT. The second investigates how CRISPR-Cas and other defense systems against mobile genetic elements affect genome composition across different species at short and intermediate evolutionary time scales. The final study focuses on phage satellites, a highly diverse family of mobile genetic elements, and investigates the co-localization of these with anti-phage defense systems.

Results: The first study assesses the influence of gene clustering criteria on pangenome analysis. It reveals that methodological inconsistencies can overshadow ecological and phylogenetic signals. In particular, estimates of genome plasticity based on binary (presence or absence) phyletic profiles may be more accurate if using synteny-based gene clusters, because they better capture the contribution of intra-species paralogs to the total gene flux. The second study demonstrates that the impact of CRISPR-Cas systems on genome composition varies depending on the species and the class of genes under consideration. The largest effects were found on genes that belong to mobile genetic elements. Strains with CRISPR-Cas systems exhibit less gene flow at intermediate evolutionary time scales, particularly lower rates of gene gain associated with mobile genetic elements. This indicates a constraining effect of CRISPR-Cas on (pan)genomic plasticity. Moreover, we found that the number and types of anti-CRISPR proteins are a factor that positively correlates with the abundance of mobile genetic elements in genomes that harbor a CRISPR-Cas system. The same analyses performed

for five different defense systems revealed a general association between defense systems and mobile genetic elements, although with significant variation across species. The last study implements a network-based approach to reclassify phage satellites into families. Additionally, it shows that phage satellites often harbor defense systems and the distribution of defense systems across phage satellites is taxa- and family-dependent.

Conclusions: Selection of appropriate gene clustering criteria is critical for an unbiased analysis of pangenomes, particularly in light of the role of HGT in microbial evolution. This thesis also highlights the significant but variable impact of defense systems, such as CRISPR-Cas, on the genomic plasticity of prokaryotic species. Finally, we show that defense systems are often co-located with phage satellites. Thus, phage satellites could be major vehicles for the transfer of defense systems across bacteria.

Resumen

Antecedentes: La transferencia genética horizontal (TGH) es un mecanismo crucial que impulsa la evolución y la adaptación microbianas. Permite un cambio genético rápido, que a menudo afecta a los microorganismos a nivel poblacional y evolutivo. La importancia vital de la TGH en la evolución microbiana queda ejemplificada por su contribución a las tasas evolutivas en forma de flujo genético. Este proceso dinámico es fundamental para entender cómo los microbios se adaptan a nuevos entornos, desarrollan resistencia a los antibióticos y adquieren islas de patogenicidad. Muchos genes adaptativos están asociados a elementos genéticos móviles (EGMs), como plásmidos y profagos, lo que pone de relieve la importancia de la plasticidad genómica. Este contexto sienta las bases para investigar el impacto de los sistemas de defensa, como los sistemas CRISPR-Cas, en la plasticidad genómica.

Objetivos: Los estudios presentados tinenen como objetivo aclarar las complejidades de la agrupación de genes, el impacto del sistema de defensa en la plasticidad genómica y la interacción entre los sistemas de defensa y los elementos genéticos móviles. El primer estudio se centra en los retos metodológicos de la agrupación (clustering) de genes en pangenomas procariotas debido a la TGH. El segundo investiga cómo los sistemas CRISPR-Cas, un mecanismo de defensa contra la TGH, afectan a la composición del genoma en diferentes especies a un nivel evolutivo corto. Además, se aplicaron los mismos análisis sobre otros cinco sistemas de defensa. El último estudio se centra en la familia de proteínas de los satélites de fagos, que presenta una baja similitud, e investiga la co-localización de los sistemas de defensa dentro de los satélites fágicos.

Resultados: El estudio inicial pone de relieve la influencia de los criterios de agrupación en el análisis del pangenoma. Revela que las incoherencias metodológicas pueden eclipsar las influencias ecológicas y filogenéticas. Los análisis de la plasticidad del genoma basados en perfiles filéticos binarios (presencia o ausencia) pueden ser más precisos si se utilizan agrupaciones génicas basadas en la sintenia, porque captan mejor la contribución de los paralogismos intraespecíficos al flujo genético total. El segundo estudio demuestra que el impacto de los sistemas CRISPR-Cas en la composición del genoma varía en función de la especie. Se encontró una influencia significativa en genes pertenecientes a elementos genéticos móviles. Las especies con estos sistemas muestran tasas más bajas de flujo genético en una escala temporal evolutiva más profunda, en particular tasas más bajas de ganancia de genes de elementos genéticos móviles, lo que indica un efecto limitante sobre la plasticidad (pan)genómica. Se descubrió que el número y los tipos de proteínas anti-CRISPR son un factor que se correlaciona positivamente con la abundancia de genes de EGM en genomas

que contienen CRISPR. Además, la aplicación de los análisis en los otros cinco sistemas de defensa reveló una asociación general entre los EGM y los sistemas de defensa, pero ésta varía dependiendo de la especie y los DF concretos. El tercer estudio identificó focos de actividad de sistemas de defensa en satélites de fagos y nuevas islas cromosómicas inducibles por fagos a través de redes de genes compartidos.

Conclusiones: La selección de criterios apropiados de agrupación de genes es crítica para el análisis imparcial de pangenomas, particularmente a la luz del papel de la TGH en la evolución microbiana. Los estudios destacan un impacto significativo pero variable de los sistemas de defensa como CRISPR-Cas en la plasticidad genómica entre especies. Los resultados obtenidos también sugieren que los satélites de fagos actúan como focos de actividad del sistema de defensa, facilitando la transferencia potencial de estos.

Table of Contents

List of Figures

List of Tables

Abbreviations and acronyms

Abi abortive infrction system

Acr anti-CRISPR

AIC Akaike information criterion

ANI Average Nucleotide Identity

ARG antibiotic resistance gene

BLAST Basic Local Alignment Search Tool

BREX Bacteriophage Exclusion (defense system)

CARD Comprehensive Antibiotic Resistance Database

CBASS cyclic-oligonucleotide-based anti-phage signaling system

CF-PICI capsid-forming PICI

CGAS core gene alignment similarity

COG Clusters of Orthologous Groups

CPR candidate phyla radiation

CRISPR Clustered Regularly Interspaced Short Palindromic Repeats

crRNA CRISPR RNA

dbAPIs anti-prokaryotic immune system databases

DDH DNA-DNA Hybridization

DFs defense systems

DIAMOND double index alignment of next-generation sequencing data

DISAM defence island system associated with restriction–modification

DM DNA methyltransferase

DMS DNA modification system

DPANN Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota.

DRTs defense-associated reverse transcriptases

dsDNA double stranded DNA

EBA Embedding-based Alignment

eggNOG evolutionary genealogy of genes: Non-supervised Orthologous Groups

EMBOSS European molecular biology open software suite

FTP File transfer protocol

GGD Genome to Genome Distance

GLMM Generalized Linear Mixed Model

G-neg-PICI Gram negative PICI

GTDB Genome Taxonomy Database

HGT Horizontal Gene Transfer

HMM Hidden Markov Model

HQ high-quality

ICEs Integrative Conjugative Elements

ILR isometric log-ratio

IMEs Integrative Mobilizable Elements

MAG Metagenomeassembled genomes

MCL Markov Cluster Algorithm

MGE Mobile Genetic Elements

MIMAG Minimum Information about a Metagenome- Assembled Genome

MLST Multi-Locus Sequence Typing

NCBI National Center for Biotechnology Information

NCS Negative correlated species

NS No Correlated Species

NVI normalized variation of information

OGC Orthologous Gene Cluster

ORFans Orphan genes

ORFs open reading frames

OSLOM Order Statistics Local Optimization Method

pAgo prokaryotic Argonautes

PAM protospacer-adjacent motif

PARIS Phage Anti-Restriction-Induced System

PCS Positive correlated species

Pycsar pyrimidine cyclase system for antiphage resistance

PGLMM Phylogenetic Generalized Linear Mixed Model

PICIs Phage-inducible chromosomal islands

PLE PICI-like elements

REs Restriction Endonucleases

RG representative genome

RGI Resistance Gene Identifier

RM Restriction-Modification

RNAi RNA interference

RTs reverse transcriptases

SAG single-amplified genomes

SaPI Staphylococcus aureus pathogenicity islands

VFs virulence factors

VFDB Virulence Factor Database

Chapter 1

Introduction and state of the art

1.1 Genomic plasticity and microbial pangenomes.

1.1.1 The concept pangenome and its properties.

Bacterial taxonomy originally began as a largely intuitive process (Williams, 1983). In 1989, DNA-DNA Hybridization (DDH > 70%) and difference in melting temperature (ΔTm threshold 5°C) were established as the gold standards for defining bacterial species (Wayne et al., 1987). With the advent of advanced sequencing techniques, the emphasis has shifted towards delineating species based on whole genomes, with Average Nucleotide Identity (ANI) >95% and Genome-to-Genome Distance (GGD) >70% as usual criteria (Goris et al., 2007; Konstantinidis and Tiedje, 2005; Thompson et al., 2013). The rapid increase in the number of sequenced genomes soon led to the realization that different isolates from the same bacterial species greatly vary in their gene content. The term *pangenome* was first coined in 2005 in a study of 8 *Streptococcus agalactiae* genomes and refers to the union of all genes in all the strains of a species. The pangenome includes *core genes* found in all strains and *accessory* (or cloud) genes that are missing in one or more strains (Tettelin et al., 2005). By 2015, the concept had expanded beyond species, being applied to broader taxonomic levels such as genus, class, phylum, and even superkingdom.

The pangenome can be dissected into two primary components:

- Core Genome: This encompasses the genes common to all strains of a given taxonomic level, be it species, genus, etc. These genes are generally assumed to be essential to the identity and function of the species. The relative size of the core genome ranges from 3% to 84% for well-sampled genomes (McInerney et al., 2017; Figure 1.1).

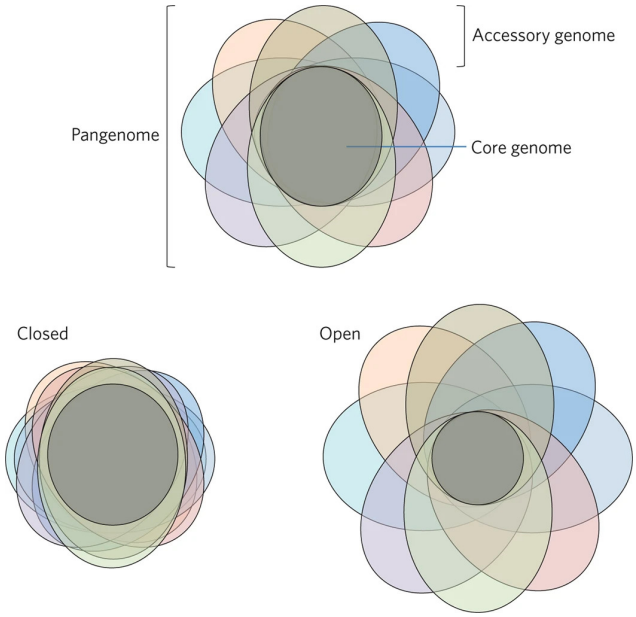

Figure 1.1: Schematic representation of pangenomes as Venn diagrams by (McInerney et al., 2017).

- Accessory (Cloud and Shell) Genome: This segment consists of genes present only in some strains or individuals of a species. Some of them are mobile genetic elements and other provide with adaptations to specific environmental conditions.

A pangenome can be classified as either open or closed by the scaling of the accessory genome with the number of sequenced genomes:

- An open pangenome expands indefinitely with the addition of new genomes, indicating that sequencing more strains will likely reveal new genes.
- A closed pangenome does not increase its coding capabilities when new genomes are incorporated.

Bacterial pangenomes are relevant to a wide range of studies, including: 1) Exploration of genomic diversity: This involves investigating the processes that shape pangenomes, inferring the dynamics of gene gain and loss over time (for instance, through horizontal gene transfer), and identifying specific genes that have been acquired or lost during evolutionary processes. 2) Gene clustering: This entails grouping genes based on sequence homology or synteny to discern functional, structural, or evolutionary parallels and disparities. Such clustering also considers the potential biases introduced by sample size. 3) Functional analysis: This seeks to unravel the roles of genes from different clustered families in encodiing protein functions, as the pangenome encompasses the set of all functional capabilities available to a species.

1.1.2 Difficulties in pangenome characterization.

To comprehensively study the pangenome of a species, it is essential to first gather all available high-quality genomes and identify their open reading frames (ORFs). These ORFs are then clustered into sets of "equivalent" genes, which serve as the basis for subsequent comparative analyses. For meaningful results, it's crucial that these gene clusters represent coherent units from both functional and evolutionary standpoints.

One straightforward approach is to cluster genes based on homology. Homologous sequences share a common ancestor. While all homologous sequences share this basic property, they can further be categorized based on their evolutionary histories. Paralogs, for instance, are genes that have duplicated either before (out-paralogs) or after (in-paralogs) speciation, evolving distinctively over time. Such genes can display functional divergence and accelerated evolutionary rates (Petitjean et al., 2017; (Ahrens et al., 2020). In contrast, orthologs are genes that share a common ancestor at the time of speciation (Fitch, 2000; Gabaldón and Koonin, 2013; Koonin, 2005). Given the intrinsic relationship between orthologs and speciation, they are often the primary focus in comparative genomics and phylogenomic studies.

Beyond homology and orthology, synteny conservation offers another layer of precision in determining gene equivalence. According to the synteny criterion, two genes (typically orthologs), are grouped together if they share the same genomic neighborhood across different genomes. This conservation helps distinguish vertically transmitted genes from those acquired horizontally and can differentiate between multiple orthologs resulting from gene expansions within a species. An illustrative example of these concepts is provided in Figure 1.2.

The outcomes of pangenome analyses can vary based on the method and parameters chosen for constructing the pangenome, especially in the gene clustering step. For example, as the sequence similarity threshold used to define homology becomes more stringent, the pangenome size tends to increase, while the core genome size decreases (Tonkin-Hill et al., 2020; Sitto and Battistuzzi, 2020; Bayliss et al., 2019). Despite this variability, many studies assume that qualitative trends remain consistent regardless of the specific criteria applied. This assumption requires careful consideration, especially given the distinctions among homology, orthology, and synteny, the variability in evolutionary rates, and the high rates of horizontal gene transfer (HGT) observed in some genomes (Puigbò et al., 2014; Hao and Golding, 2006; Treangen and Rocha, 2011). The possibility that homology, orthology, and synteny-based gene clustering produce incongruent results in comparative pangenome studies is not just a technical caveat. Instead, intraspecific HGT and gene duplications raise the fundamental question of what equivalence class (homology, orthology, synteny, or something else) best captures the essentially dynamic nature of pangenome evolution (Cummins et al., 2022).

To stress the fact that the optimal criterion is somewhat arbitrary and dependent on the main research goal, we adopt the method-agnostic term Operational Gene Cluster (OGC) as an umbrella that includes homologs, classical orthologs (possibly inferred through different methods), and vertically transmitted subsets of orthologs with conserved synteny. Depending on the particular choice, OGC may imply different degrees of functional equivalence and shared ancestry.

State-of-the-art methods for OGC construction implement different strategies to deal with sensitivity-vs-specificity trade-offs at manageable computational cost. Basic approaches, like CD-HIT (Fu et al., 2012) and MMseqs2 (Steinegger and Söding, 2017), group homologous genes based on a fixed similarity threshold. More advanced tools differentiate true orthology from other classes of homology either by constructing gene-level phylogenetic trees and applying heuristic rules (Altenhoff and Dessimoz, 2012) or by subclustering based on gene neighborhoods. While the first approach, used by databases like COG (Galperin et al., 2019) and tools like OrthoFinder (Emms and Kelly, 2019), is computationally demanding, it is also more accurate for assessing true orthology. Synteny-based methods, like Roary (Page et al., 2015) and PanOCT (Fouts et al., 2012), are faster but might deviate from the traditional concept of orthology. Phylogeny-aware and synteny-based methods are not mutually exclusive, but they are rarely used together due to computational constraints (Zhou et al., 2020). A comprehensive summary of these methods is provided in (Manzano-Morales et al., 2023).

Despite their distinct conceptual underpinnings, different types of OGC are sometimes just viewed as exchangeable heuristic approaches that approximate the concept of orthology in a computationally tractable manner. In this thesis, we tested to what extent such an assumption is true and found that some properties of the pangenome, mostly concerning its size and the identity of the core genome, are indeed robust. However, pangenome properties that are related to its fluidity (that is, the genomic variability among strains) can be greatly affected, leading to relatively poor correlation in the results of comparative genomic analyses conducted with different methods.

1.1.3 Pangenomes in microbial ecology and evolution.

Genomic plasticity is a fundamental aspect of microbial adaptation and evolution. It refers to the ability of a genome to undergo changes and reorganizations. HGT is the process by which organisms acquire genetic material from other organisms in the population that are not their direct ancestors (Soucy et al., 2015). As a major source of new genetic material, HGT rates are key to genomic plasticity. Several studies have shown that the rates of evolution by gene gain and loss are of the same order as those associated with nonsynonymous substitution

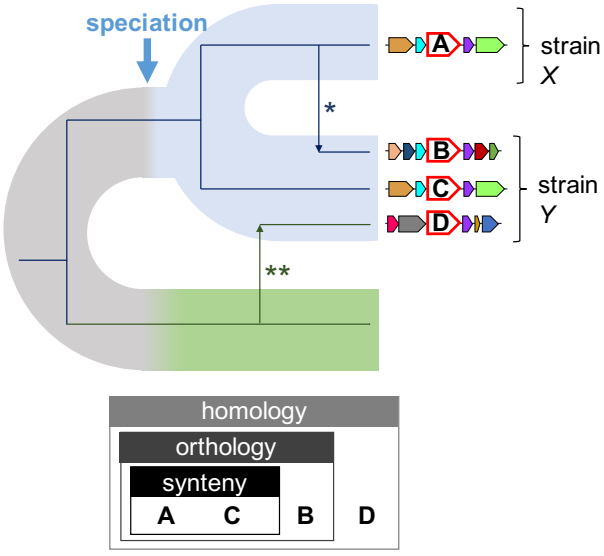

Figure 1.2: Homology, species-level orthology, and synteny conservation (adapted from Manzano-Morales et al., 2023)

rates (Puigbò et al., 2014; Iranzo et al., 2019).

The impact of HGT on genomic plasticity is highlighted by the connection of numerous adaptive genes with mobile genetic elements (MGE), including plasmids, transposons, and prophages. One of the most notable examples of the effects of HGT is the dissemination of antibiotic resistance. Antibiotic resistance genes are frequently located on MGE like plasmids and prophages, which can be easily transferred between bacteria, rapidly spreading resistance traits (Bennett, 2008; Partridge et al., 2018).

Pangenomes are a valuable source of information for eco-evolutionary studies, as they encompass all the genetic and functional diversity harbored by the strains of a species (Collins and Higgs, 2012). This comprehensive genetic repertoire is shaped by the interplay of gene acquisition, gene loss, and selection, providing a lens into genome flexibility, ecological specialization, and the adaptability of organisms to ever-changing environments (Liao et al., 2021; Whelan et al., 2021; Maistrenko et al., 2020; Shapiro, 2017).

A large portion of accessory genes in pangenomes, especially orphan genes (ORFans) without known homologs, remain functionally uncharacterized. Due to their rarity and rapid evolution, these genes are often overlooked in both experimental and computational comparative studies. Yet, a significant fraction of better characterized accessory genes can be categorized into two main functional groups: (a) accessory genes involved in adaptation to specific environmental conditions (which would include metabolic genes with a locally adaptive role); and (b) genes that belong to genetic parasites or selfish genes, including mobile genetic elements (MGE),

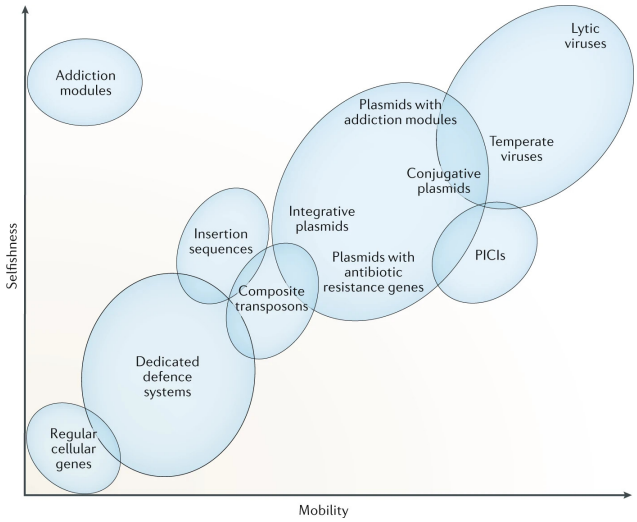

Figure 1.3: Distribution of mobile genetic elements and in the virtual space bounded by the axes of selfishness and mobility (taken from Koonin et al., 2020)

and defense systems (Figure 1.3). These selfish elements often encode the machinery required for their own horizontal transfer or exploit other MGE to spread across the population. On the other side, microbial genomes often harbor defense systems (DFs) that prevent the entry and proliferation of parasitic MGE. (Makarova et al., 2013; Makarova, Haft, et al., 2011; Makarova, Wolf, et al., 2011). Almost all microbial genomes, with the possible exception of highly streamlined ones, contain MGEs and DFs (Frost et al., 2005). Although some MGE, such as (pro)phages, plasmids, and transposons, have been known for decades, specialized defense systems have only recently been identified in significant numbers (Makarova et al., 2020; Doron et al., 2018; L. Gao et al., 2020; Rousset et al., 2022; Wang et al., 2023; Botelho, 2023; Tesson et al., 2022; Payne et al., 2021; Payne et al., 2022). Some of the best-studied defense systems include restriction-modification (RM), CRISPR-Cas, and toxin-antitoxins. Among them, CRISPR-Cas systems have gained significant attention due to their potential in gene editing technologies. In biological terms, the most remarkable property of CRISPR-Cas systems is their adaptivenature and wide spread, that includes most archaea and around 40% of bacteria (though these numbers could be lower owing to their absence in *Patescibacteria*).

1.2 CRISPR and other defense systems.

1.2.1 The history of the CRISPR-Cas system.

The CRISPR-Cas system's history traces back to 1987 when Yoshizumi Ishino discovered repeated DNA sequences in the genome of *Escherichia coli* (Ishino et al., 1987). These

sequences, initially not fully understood, were later observed in archaea in 1993, specifically in *Haloferax mediterranei* (Mojica et al., 1993). With the advent of high-throughput sequencing, Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) were identified in various bacterial and archaeal genomes. By the early 2000s, Mojica found that the sequences between spacers were linked to bacteriophages, archaeal viruses, and plasmids, revealing CRISPR's immune function (Mojica et al., 2000; Mojica et al., 2005; Pourcel et al., 2005; Bolotin et al., 2005). The role of spacers as the memory component of CRISPR systems is now well established (Shmakov et al., 2017). The term "CRISPR" was proposed in 2002 (Jansen et al., 2002), clarifying the confusion from diverse names associated with similar sequences.

Cas proteins, discovered alongside CRISPR, were found to function similarly to the eukaryotic RNA interference (RNAi) system (Makarova et al., 2002; Jansen et al., 2002; Makarova et al., 2006). Cas3 and Cas4, identified in 2002, were linked to various DNA-related processes (containing helicases, exonucleases motif), while Cas1 (endonuclease, helps in the integration of new spacers into the CRISPR locus) and Cas2 (endonuclease, often forming a complex with Cas1 to facilitate the integration of new spacers) were unknow at that moment (Jansen et al., 2002). Due to their association with CRISPRs it was proposed that Cas proteins play a role in the funcion of CRISPR loci (Jansen et al., 2002). The immune function of the CRISPR-Cas system was experimentally proven in 2007, demonstrating its ability to provide resistance against specific phages (Barrangou et al., 2007; Marraffini and Sontheimer, 2008). Further studies revealed that the CRISPR-Cas system could limit plasmid transformation and that RNA molecules derived from CRISPR transcription collaborated with nearby Cas proteins (Brouns et al., 2008). A pivotal discovery was that the cas9 gene alone could provide cross-species protection against plasmid transformation and phage infection (Sapranauskas et al., 2011), leading to the development of the CRISPR-Cas9 genome editing tool. Subsequently, research on CRISPR has experienced a significant expansion.

Building on the foundational discoveries of the CRISPR-Cas system, the scientific community rapidly developed tools to identify and analyze this system within genomes. Tools such as CRISPRCasTyper (Russel et al., 2020), CRISPRCasFinder (Couvin et al., 2018), CRISPRidentify (Mitrofanov et al., 2021), and CRISPRcasIdentifier (Padilha et al., 2020) emerged. These innovative tools not only facilitated the identification of CRISPR-Cas systems but also enabled large-scale computational studies. Researchers could now delve deeper into the evolutionary history of the CRISPR-Cas system, tracing its origins and understanding its diversification over time.

1.2.2 Mechanism and classification of CRISPR-Cas systems.

The CRISPR-Cas system operates through three primary stages: adaptation, expression, and interference. In the adaptation phase, the Cas protein complex recognizes and binds to the target protospacer-adjacent motif (PAM) and excises the protospacer. Once the repeat sequence at the CRISPR array's 5' end duplicates, the adaptation complex integrates the protospacer DNA into the array, converting it into a spacer. Notably, some CRISPR-Cas systems that acquire spacers from RNA utilize a reverse transcriptase encoded at the CRISPR–cas locus. During the expression phase, the CRISPR array undergoes transcription, producing mature CRISPR RNAs (crRNAs). Each crRNA encompasses the spacer sequence and parts of the adjacent repeats. In the interference phase, the crRNA detects the protospacer originating from prophages or plasmids. The Cas protein, either inherent to the effector or recruited during this phase, then cleaves and neutralizes the identified protospacer (Makarova et al., 2020). In the end, CRISPR-Cas integrates fragments of genetic material from potential invaders (e.g., prophages and plasmids) as spacers and uses those spacers as probes to recognize and prevent later invasions. The functionality of CRISPR-Cas systems can differ based on the distinct system types. Given the rapid divergence of spacers in the CRISPR system and the absence of marker genes, classifying CRISPR-Cas systems presents challenges. Historically, authoritative classifications were proposed in 2011, 2015, 2017, and 2020 (Makarova, Haft, et al., 2011; Makarova et al., 2015; Koonin, Makarova, and Zhang, 2017; Makarova et al., 2020), each based on varying combinations of Cas proteins concerning their range and similarity with an updated list of new systems. As of the latest publication, these classifications have identified 2 classes, 6 types, and 33 subtypes of CRISPR-Cas systems (Makarova et al., 2020; Figure 1.4). Moreover, the discovery of new CRISPR-Cas systems continues. Traditional classification methods relied on highly conserved proteins like Cas1 (Makarova et al., 2006). However, with the identification of numerous new CRISPR-Cas systems, the focus shifted to grouping Cas proteins into distinct protein families based on gene sequence or protein structure similarity. This grouping was achieved by analyzing the topologies of Cas protein families across genomes (Makarova et al., 2020). However, CRISPR systems diverge rapidly, and within a type, the composition of the system may not remain consistent (Makarova et al., 2020). Thus, a bipartite gene-sharing network has been proposed for low-level clustering of CRISPR-Cas systems (Iranzo, Krupovic, and Koonin, 2016). Additionally, neighborhood analysis, based on genomic synteny , has been instrumental in deciphering the architecture of specific variants of the CRISPR-Cas system (Shmakov et al., 2019).

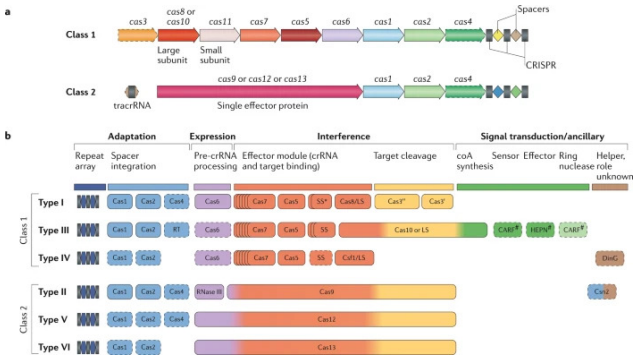

Figure 1.4: The two classes of CRISPR–Cas systems and their modular organization (taken from (Makarova et al., 2020)

1.2.3 The debate on the role of CRISPR-Cas in gene transfer.

In the last decade, there has been an increasing interest in evaluating the ecological and genetic factors that regulate HGT and determine the fate of recently acquired genes in microbial populations (Iranzo et al., 2019; Iranzo et al., 2017). Among those factors, special attention has been paid to the prokaryotic CRISPR-Cas immunity adaptive system (Koonin, 2019). This system can protect bacteria and archaea from viruses and other MGEs by recognizing, targeting, and degrading extraneous DNA and RNA molecules (Makarova and Koonin, 2015). Due to the ability to target and destroy genetic exchange vehicles such as viruses and conjugative plasmids, CRISPR-Cas system often interferes with HGT. In this line, experimental evidence has shown that CRISPR loci can limit the spread of antibiotic resistance genes (ARGs) (Marraffini and Sontheimer, 2008). Moreover, in *Pectobacterium atrosepticum*, CRISPR-Cas can inhibit the transduction of plasmid and chromosomal loci (Watson et al., 2018), while in Vibrio cholerae, a Type I-F CRISPR-Cas system encoded by phage ICP1 is used to target and overcome phage-inducible chromosomal island-like elements (O'Hara et al., 2017). Interestingly, Watson et al (Watson et al., 2018), based on experiments with *Pectobacterium atrosepticum*, suggested that CRISPR-Cas system can, in fact, enhance transduction of escape phages by limiting the spread of lethal phages. Albeit the evidence showing the potential of CRISPR-Cas systems in altering HGT, its impact on microbial populations on evolutionary periods remains a topic of debate. While some studies suggest that CRISPR-Cas systems can inhibit HGT, others propose that its effects may be negligible. A comparative study of genomes with and without CRISPR-Cas in *Pseudomonas aeruginosa* has shown that CRISPR-Cas is associated with smaller genomes, higher GC content, and lower abundance of prophage and ICE (integrative and conjugative elements) (Wheatley and MacLean, 2021). In *Enterococcus*, *Staphylococcus*, *Acinetobacter* and *Pseudomonas*, lower levels of MGE carrying ARGs were found in genomes that had a higher abundance

of CRISPR-Cas (Pursey et al., 2022). The association between CRISPR-Cas and ARGs in genomes can be both positive and negative depending on the type of ARGs, as seen in *Bacillus cereus*, *Neisseria meningitidis* and *Escherichia coli* (Shehreen et al., 2019). On the other side, CRISPR-Cas abundance was found to be strongly positively associated with viral abundance in several species (Meaden et al., 2022). Interestingly, a recent study found that type I-C, I-E, I-F, III-A, IV-A1, and IV-A3 CRISPR systems are encoded within MGEs (Botelho et al., 2023), suggesting idea that the prevalence of CRISPR could be positively related to MGE dynamics. Finally, other studies found that CRISPR-Cas systems do not alter the HGT rates measeured over long evolutionary periods by studying 269 groups of bacteria and archaea (Gophna et al., 2015). According to that, CRISPR-Cas may not impose overly stringent constraints on gene exchange in natural populations. To summarize, previous research has shown that the CRISPR-Cas system can favor or limit HGT.However, because most studies involved a single species, it is unclear what the most general scenario is (Westra and Levin, 2020).

1.2.4 CRISPR and other prokaryotic defense systems as a diverse arsenal.

While the CRISPR-Cas system is a well-recognized prokaryotic defense mechanism, it is just one of many such systems that prokaryotes employ (Bernheim and Sorek, 2020; Figure 1.5). As of 2017, several other defense systems were identified, including the restriction-modification system (RM, and DNA modification system: DMS) which targets specific sequences on invading phages (Oliveira et al., 2014), and the abortive infection system (Abi) that triggers cell death or metabolic arrest upon infection (Chopin et al., 2005). The RM system consists of two enzymes: Restriction Endonucleases (REs) and DNA methyltransferases (DM) (Bujnicki, 2001). REs recognize and cleave specific DNA sequences, typically 4-8 base pairs in length, called recognition sites. When a bacteriophage injects its DNA into a bacterial cell, the restriction endonuclease scans the DNA for its specific recognition site. Upon finding this site, the enzyme cleaves the phage DNA, rendering it non-functional and preventing the phage from replicating within the bacterium. While restriction endonucleases are responsible for cutting foreign DNA, the bacterial genome also contains the same recognition sites and is at risk of being cleaved. To protect its own DNA from being cut, the bacterium uses DNA methyltransferases. These enzymes add a methyl group to adenine or cytosine residues within the recognition sites of the host DNA. Methylation of these sites prevents the restriction endonuclease from recognizing and cutting the bacterial DNA, ensuring that only foreign, non-methylated DNA is targeted and cleaved (Bujnicki, 2001).

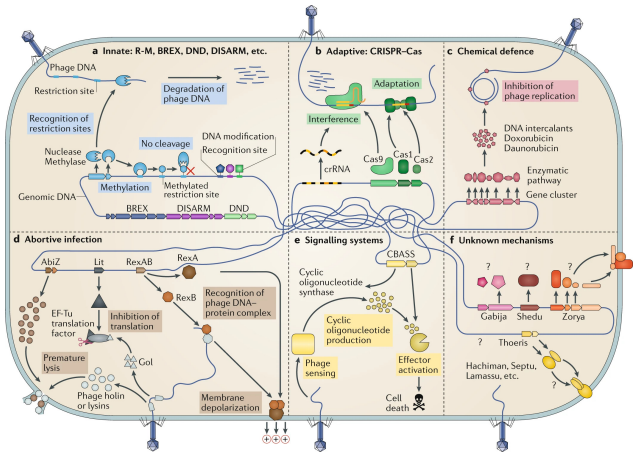

Figure 1.5: Defence systems that target nucleic acids encompass both innate and adaptive immunity (taken from Bernheim and Sorek, 2020).

Most Abi systems are toxin-antitoxin systems (Koonin, 2023). The Abi system is triggered by phage infections and initiates cellular responses, disrupting normal processes. This either induces programmed cell death or halts metabolic activities, preventing phage replication. While the infected bacterium may die, this action safeguards the surrounding bacterial population from further infections. Different Abi systems have unique triggers and responses, tailored to specific phage threats (Lopatina et al., 2020). Other systems, such as Bacteriophage Exclusion (BREX) (Goldfarb et al., 2015) and prokaryotic Argonautes (pAgos) (Makarova, Wolf, van der Oost, and Koonin, 2009) have mechanisms that remain to be fully elucidated.

Since 2018, more and more antiviral defense systems have been discovered, including Druantia, Gabija, Hachiman, Kiwa, Septu, Shedu, Thoeris, Wadjet, Zorya (Doron et al., 2018) and DISARM (Ofir et al., 2018). The cyclic-oligonucleotide-based anti-phage signaling system (CBASS) was identified by (D. Cohen et al., 2019), and a subsequent analysis of 38,167 bacterial and archaeal genomes revealed over 5000 CBASS systems (Millman, Melamed, et al., 2020). These systems generate a cyclic oligonucleotide signaling molecule upon phage infection, activating an effector that induces cell death through various mechanisms. Another discovery highlighted defense-associated reverse transcriptases (DRTs), a family of reverse transcriptases (RTs) that serve as active defense systems, with six candidates (UG1, UG2, UG3, UG8, UG15, and UG16) offering robust protection against double stranded DNA (dsDNA) phages (L. Gao et al., 2020). Recently, Retrons, which are elements composed of a reverse transcriptase and non-coding RNA and were previously of uncertain function, have been shown to play a role in defense against phages (Millman, Bernheim, et al., 2020).

Defense mechanisms in bacteria and archaea can be categorized by different criteria. Based on their operational principles: (a) resistance through altering virus receptors, (b) immunity,

and (c) triggering dormancy or initiating programmed cell death (Koonin, Makarova, and Wolf, 2017). Based on primary functional pathways: recombination, epigenetics, abortive infection, and adaptive (Rocha and Bikard, 2022). Recent advancements in the field have been further propelled by the development of tools like Defense-Finder and Padloc, which collectively integrate essential proteins as marker genes from over 60 defense systems or subsystems (Tesson et al., 2022; Payne et al., 2021). Defense-Finder, moreover, categorizes defense systems based on their functions into virus nucleic acid degradation, Abortive infection, inhibition of DNA/RNA synthesis, and an unknown category (Tesson et al., 2022).With these tools and discoveries, the study of prokaryotic defense systems has entered a new era, amassing a wealth of data. This positions researchers to delve deeper into the evolutionary implications of host-pathogen interactions and the roles played by these diverse defense systems.

1.3 Mobile genetic elements (mobilome).

1.3.1 The role of mobile genetic elements in shaping prokaryotic pangenomes.

MGEs play a pivotal role in shaping the pangenomes of prokaryotes, acting as agents of horizontal gene transfer. These elements, which include viruses, plasmids, transposons, ICEs and Integrative Mobilizable Elements (IMEs), among others, are segments of DNA capable of moving between different bacterial and archaea genomes (Burrus and Waldor, 2004; Frost et al., 2005; Guédon et al., 2017). Their mobility is facilitated through three primary mechanisms: transformation, conjugation, and transduction, with other mechanisms like transfection also playing a role (Thomas and Nielsen, 2005). The ubiquity of MGE activities across prokaryotes is well-documented (Gogarten and Townsend, 2005). For example, IMEs can independently encode their excision and integration, and can manipulate the mating machinery of conjugative elements, such as plasmids or ICEs, to aid their intercellular transfer (Guédon et al., 2017).

MGEs contribute significantly to the formation of pangenomes. They are part of accessory genes, which can augment the adaptive capabilities of the species, potentially enabling organisms to exploit new ecological niches or functions. MGEs can carry accessory genes that are advantageous to their hosts. When these genes boost the host's fitness, it can promote the MGE's transmission within the host lineage. This leads to an expansion of the host population that carries the MGE, driven by the benefits the MGE provides (Stevenson et al., 2017). However, not all MGEs operate this way. Some, like certain plasmids, carry very few accessory genes, indicating that they might be more parasitic and less beneficial to their hosts.

They might prioritize their own spread and survival over providing benefits to their hosts. This can lead them to become more infectious, focusing on their own propagation rather than carrying genes that are advantageous for their hosts (Lopatkin et al., 2016). While these elements often carry genes essential for their replication and transmission, they can also harbor a plethora of genes with yet-to-be-determined functions (Brockhurst et al., 2019).

The exchange of genes is predominantly mediated by MGEs. Bacteriophages, for example, often have restricted host ranges, typically infecting specific species or genera (N. L. Gao et al., 2018; Hyman and Abedon, 2010). In contrast, plasmids can exhibit broader host ranges, depending on the diversity of replication genes they possess, which are essential for their stable maintenance across different host taxa (Jain and Srivastava, 2013).

It's noteworthy that transposons can be found within ICEs, further highlighting the interconnected nature of these elements (Burrus and Waldor, 2004). Additionally, CRISPR systems, traditionally known for their defense functions, have been co-opted for various other roles. For instance, interplasmid competition by plasmids and antidefense and interviral conflicts by viruses (Koonin and Makarova, 2022). The dynamic interplay between MGE, defense systems, and prokaryotic genomes underscores the complexity and adaptability of microbial life, with MGEs acting as crucial drivers of genetic diversity and evolution.

1.3.2 Defense system in MGE.

All organisms employ various systems of innate and adaptive immunity to regulate the dissemination of mobile genetic elements (Koonin, 2016). Bacteria have evolved a range of intricate defense mechanisms to counteract the intrusion of MGE, and MGE have evolved mechanisms to escape bacterial defense systems. For instance, a class of anti-defense proteins was found in prophages, plasmids, and ICEs, known as anti-CRISPR (Acr), which can inhibit CRISPR-Cas systems (Bondy-Denomy et al., 2013; Pinilla-Redondo et al., 2020).

Defense systems in MGEs are intricate and have evolved to serve various purposes. Many defense systems remain poorly understood, and there's a belief that numerous systems are yet to be discovered. The expansion in the number and type of defense systems recently has been attributed to researchers identifying novel systems that co-localize with previously known ones. These genes often have specific locations in MGEs, such as near the cos site of P4-like satellites, which could be potential areas of interest for discovering novel defense systems (Rousset et al., 2022). A significant number of these systems might be awaiting discovery among the myriad of MGEs present across microbial genomes.

Two of the most well-known defense systems, toxin-antitoxin, and restriction-modification,

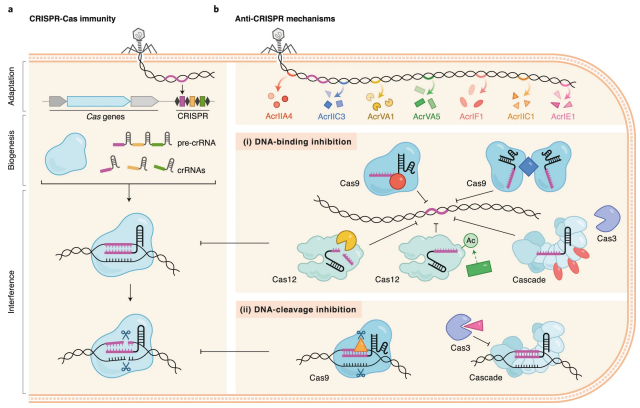

Figure 1.6: Stages of CRISPR-Cas immunity and mechanisms of Acr function (taken from Marino et al., 2020)

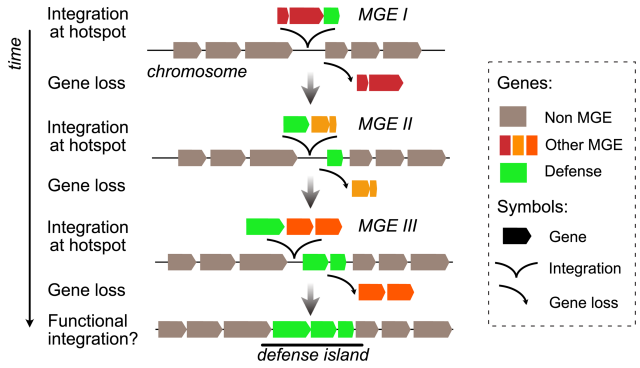

Figure 1.7: MGE turnover at hotspots may result in defense islands (Rocha and Bikard, 2022).

leverage mobile genetic elements for their dissemination (Makarova, Wolf, and Koonin, 2009) and CRISPR-Cas systems are also sometimes recruited by MGEs (Peters et al., 2017; Faure et al., 2019)A recent study found that almost half of the ICEs (42.9%) and one third of the IMEs (31.0%) in complete genomes contained least one defense system (Botelho, 2023). From an evolutionary viewpoint, it is tempting to propose that the presence of multiple defense systems may be more involved in the maintenance of MGEs than of the cell (Rocha and Bikard, 2022). Certain satellites can prevent phage transmission by releasing viral particles that exclusively contain the satellite genome (Seed et al., 2013).

Anti-defense mechanisms, including both anti-RM and anti-CRISPR/Cas (Figure 1.6), frequently group together, especially within identifiable MGEs (Pinilla-Redondo et al., 2020). The existence of defense mechanisms within MGEs explains the joint localization of defense and counter-defense systems in specific regions of the bacterial chromosome(Rocha and Bikard, 2022; Figure 1.7). Genes acquired through HGT, especially MGEs, often integrate in well-delimited chromosome hotspots, giving rise to such defense islands. (Touchon et al., 2009; Rodriguez-Valera et al., 2009; Oliveira et al., 2017; Hackl et al., 2023).

1.3.3 Phage satellites and their relationship to other MGE.

Phage satellites are emerging as significant actors in the realm of MGEs. Phage satellites can hijack viral particles produced by phages for their own dissemination(Seed et al., 2013). Although this induces cellular death, the inhibitory effect of the satellite on phage reproduction provides a protective shield to the microbial community (Rocha and Bikard, 2022). Moreover, satellites can employ mechanisms inherent to other MGEs for their propagation, such as the utilization of conjugative pili by mobilizable elements (Smillie et al., 2010).

Currently, there are four well-described families of phage satellites (de Sousa et al., 2023). All satellites encode a set of core functions that are sometimes non-homologous but can be grouped into four major groups (integration, regulation, replication and hijacking) (de Sousa et al., 2023; Ibarra-Chávez et al., 2021):

- P4-like satellites refers to those that resemble the well-studied P4 satellites of *Enterobacteriaceae*. A large majority of P4-like satellites are present in *Enterobacteriaceae*, but other families harboring these elements include *Yersiniaceae*, *Pectobacteriaceae*, *Erwinaceae* and *Hafniaceae*. The well described "core" genes of P4-like satellites include:
  - Integrase, required for integration into chromosomes that usually occurs by one that is of the Tyrosine recombinase family.
  - Psu, Delta and Sid, which are involved in the hijacking of the capsid of the P2 helper phage.
  - Regulatory protein, typically homologous to AlpA. Ash (also called ϵ), which inactivates the repressor of the helper phage, causing its induction.
- PICIs are present in many more species than P4-like satellites. Most PICIs are found in *Escherichia* (two sub families), *Mycobacterium* and *Staphylococcus*, but they are also present in *Acinetobacter*, *Bacillus* (as well as *Lactobacillus*), *Burkholderia*, *Clostridium*, *Rhodococcus* or *Sinorizhobium*. The "core" genes of PICIs include:
  - Integrase.
  - Regulatory protein homologous to MerR of Stl.
  - Primase-replicase module
  - Capsid morphogenesis (more frequent in PICI from Proteobacteria), encoding a protein that is thought to modify the morphology of the hijacked capsids to block the encapsidation of phage DNA.

  - Small terminase subunit (TerS), responsible for redirecting the packaging of the capsid to the satellite's DNA

- cfPICIs are related to PICIs, but with a unique trait: they assemble their own cfPICI-specific capsid. Yet, cfPICI are incapable of forming viable phage particles because they lack other structural genes that they hijack from the helper phage, e.g. holins and tail-associated proteins. Five core components of cfPICI are homologous or analogous to the five core components of PICIs. But they also have their own unique components. The well described "core" genes associated with cfPICIs that are also associated with PICIs include:
  - Integrase.
  - Regulatory protein homologous to MerR of Stl.
  - Primase-replicase module.
  - Capsid morphogenesis (more frequent in PICI from Proteobacteria), encoding a protein that is thought to modify the morphology of the hijacked capsids to block the encapsidation of phage DNA.
  - Nuclease (HNH) that is essential for phage head morphogenesis (and DNA packaging) in fully functional phages.
  - Head decoration module (a serine protease).
- PLEs form very homogeneous groups and are specific to Vibrio. Meaning you will not find them in other hosts. The well described "core" genes associated with PLEs include:
  - Integrase.
  - CapR, which represses the capsid morphogenesis of the ICP1 prophage.
  - Replication initiation protein (RepA).
  - Nickase that hampers the replication of the hijacked phage (nixI).
  - Gene that accelerates the lysis of the bacterial host cell (lidI).
  - Sigma 70-like factor, a component of the specificity subunit of the bacterial RNA polymerase.
  - A profile with homology to a cyclin-dependent kinase-activating kinase (MAT1) suggested to be involved in nucleotide excision repair of damaged DNA.

PICIs have been on the spotlight for their profound impact on bacterial communities, especially

in the context of virulence and toxin production (Penadés and Christie, 2015). These PICIs, composed of gene fragments from the bacterial chromosome, become activated during prophage infections or inductions. They have the capability to replicate, be packaged within a phage, and subsequently be released, facilitating their integration into diverse bacterial hosts (Penadés and Christie, 2015). PICIs are comparable to phages, encompassing a core genome that includes modules for induction, integration-excision, replication, packaging, and accessory functions (Penadés and Christie, 2015). The key difference is that the spread of PICIs requires a helper phage (Penadés and Christie, 2015). Given their propensity to carry genes associated with virulence, host adaptation, antibiotic resistance, and biofilm formation, PICIs might play a pivotal role in the ongoing tug-of-war between parasites and hosts, often oscillating between phage and bacterial chromosomes (Penadés and Christie, 2015). Notably, certain antibiotic resistance genes, such as *fosB*, *ear*, and the multidrug exporter gene *mdr*, have been identified on PICIs in *Staphylococcus aureus*, suggesting their potential transfer via these islands (Novick et al., 2010).

Despite their dependency on helper phages for transmission, the host range of PICIs is not strictly confined to their phage counterparts (Ibarra-Chávez et al., 2022). Their prevalence in genomes remains enigmatic, primarily due to the lack of a systematic identification approach before 2022 (Ibarra-Chávez et al., 2022). However, the recent identification of approximately 5,000 phage satellites in complete bacterial genomes has fueled the study of these satellites across bacterial genetic repertoires (de Sousa et al., 2023; Figure 1.8). Also, promising bioinformatic tools will enable the identification of many PICIs among prokaryotes, forming the foundation for future study.

1.4 The interplay between defense systems and mobile genetic elements determines the evolutionary tug-of-war between hosts and parasites.

The concept of "arms race" refers to the continuous evolutionary struggle between organisms and their adversaries. This continuous struggle propels the evolution of adaptations in both parties, each striving to gain an upper hand. A prime example of this arms race is observed in the enduring conflict between viruses and prokaryotes. Such interactions, reminiscent of warfare, have been pivotal in driving evolutionary innovations (Forterre and Prangishvili, 2009). Viruses frequently co-opt host genes to bolster their defense mechanisms and obtain novel traits. Conversely, host cells often harness viral genes for diverse purposes (e.g., obtain antibiotic resistance or resist against another virus). Given the swift evolutionary trajectory of

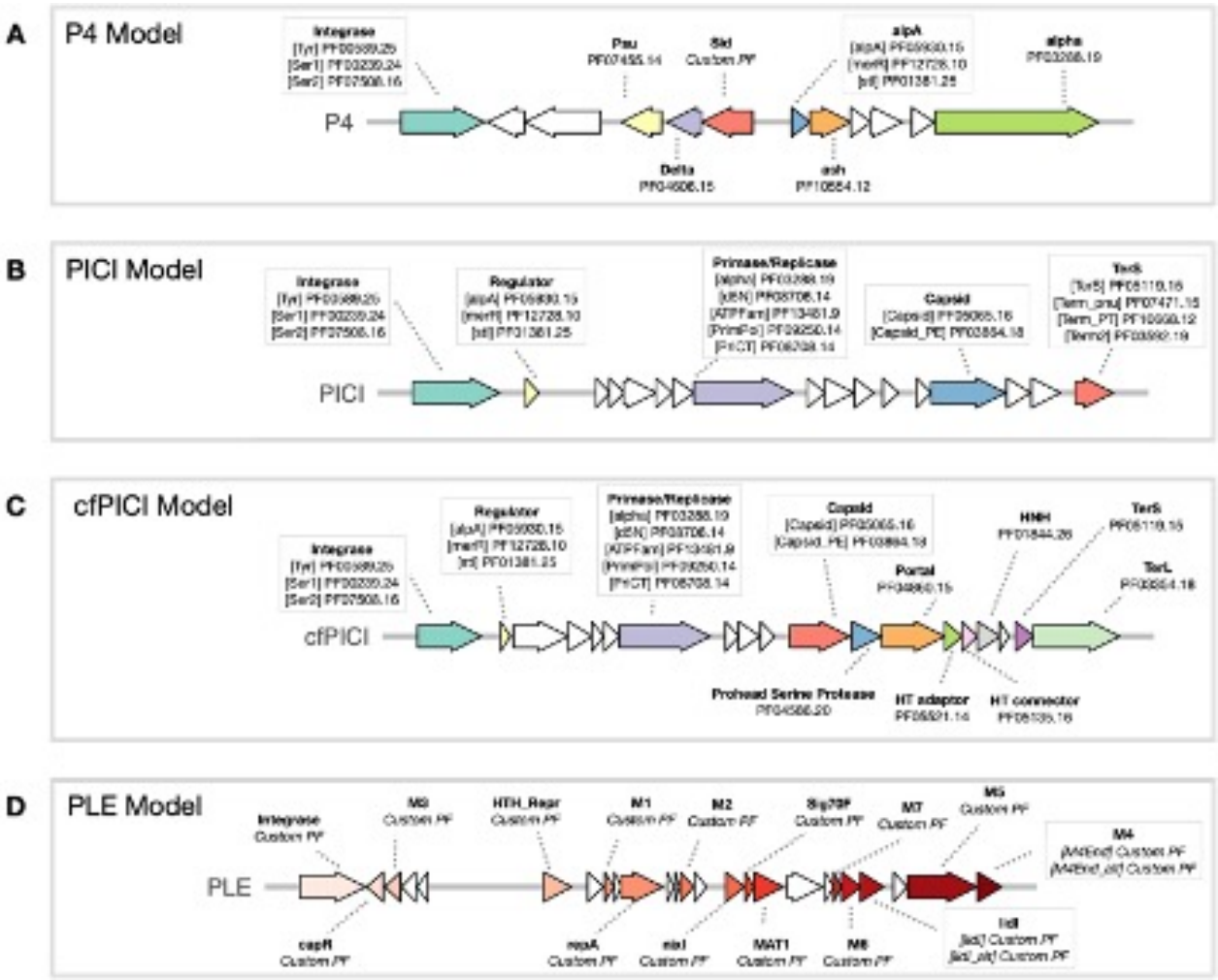

Figure 1.8: Core components of the models for each of the satellite families (taken from de Sousa et al., 2023)

viruses, they are postulated as major contributors to the genesis of new genes in the biosphere (Daubin and Ochman, 2004).

A sizable portion of prokaryotic genomes, up to 10%, is allocated to defense system components (Makarova et al., 2013). Many of these defense genes are highly mobile and bear hallmarks of selfish genetic elements (Koonin and Dolja, 2013). The evolutionary arms race extends beyond just the host's antivirus response (Koonin and Dolja, 2013), with layers of anti-defense mechanisms (and even anti-anti defense systems) continually being uncovered (Rocha and Bikard, 2022).

The relationship between MGEs and host defense systems is multifaceted (Figure 1.9). Beyond their antagonistic interactions, they also share components, such as nucleases, which has led to the 'guns for hire' concept (Koonin and Krupovic, 2015). This concept highlights the dynamic exchange of proteins between defense systems and MGEs. These molecular "weapons" can move between MGEs and defense systems, driven by ecological and evolutionary pressures. The entity offering the most conducive environment for their proliferation retains these genes. This fluid exchange of components for either defense or offense is facilitated by the underlying mechanistic parallels between the core biochemical activities of both processes. At the end, these molecular tools gravitate towards the "highest bidder" that maximizes their proliferation, regardless of their original function (Koonin et al., 2020).

Maintaining MGEs or defense systems is not without its costs. While these elements are crucial for certain adaptive responses, they come with significant burdens, especially when they are not immediately essential (Iranzo, Puigbò, et al., 2016). The high loss rates of such genes underscore the evolutionary pressures they are subjected to. However, the mere deleterious effects of MGEs do not capture the entirety of their persistence in genomes. A range of factors, from the intermittent benefits they offer to the host to the impact of external environmental conditions, play relevant roles in their retention (Iranzo and Koonin, 2018). As for the defense systems, mathematical analyses of genome evolution indicate that CRISPR-Cas systems predominantly face negative selection over extended evolutionary periods (Iranzo et al., 2017). Supporting that view, experiments have suggested that deactivating CRISPR–Cas systems in *Streptococcus pneumoniae* could be advantageous under specific scenarios (Bikard et al., 2012).

The burden of maintaining defense systems (e.g., due to their metabolic cost or the risk of autoimmunity) has led to a dynamic balance where such systems are frequently acquired and subsequently lost in bacteria. This observation motivated the 'pan-immune system' concept, which refers to the collective immune systems of genetically close microbial strains within a community (Bernheim and Sorek, 2020).

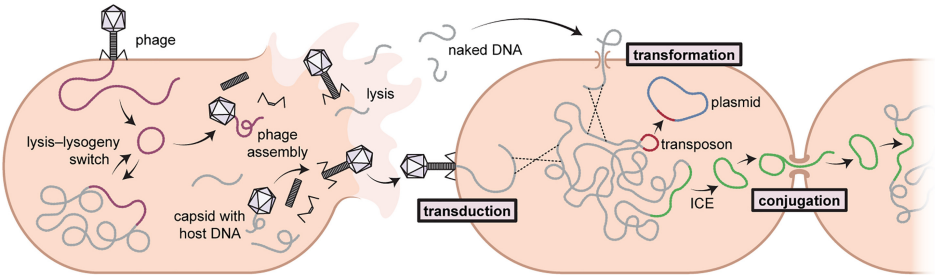

Figure 1.9: Major routes of HGT and some of the MGEs that drive it (taken from Mayo-Muñoz et al., 2023), exemplified for bacteria.

Chapter 2

Objectives

The overarching goal of this work is to understand the interplay among mobile genetic elements of different kinds and associated defense systems, and how such interplay affects genomic plascticity and the ability of the host to gain and lose genes. The ability of the host to gain and lose genes is going to determine prokaryotic ability to adapt to environmental stresses, compete with other bacteria, or produce pathology in plants, insects or animals. While different chapters delve into specific research topics —-pangenomes, CRISPR-Cas, and phage satellites—- they fundamentally reflect the dynamic nature of prokaryotic genomes as a result of their exposure to mobile genetic elements. From a methodological perspective, this thesis combines bioinformatic tools for comparative and functional genomics, advanced statistical methods, and large genomic datasets.

The thesis is organized around five primary themes that define a set of specific goals:

- The initial section delves into the nuances of pangenome research, emphasizing how the selection of pangenome tools (and more precisely gene clustering methods) can significantly affect the subsequent quantification of pangenome properties. Most of the contents (text, tables and figures) of this section have been adapted from (Manzano-Morales et al., 2023).
- In next section, the effects of the CRISPR-Cas system on gene gain and loss are examined across species, factoring in phylogeny. This exploration extends to the impact of CRISPR-Cas on accessory genes and MGE.
- The study is subsequently expanded to other defense systems, assessing similarities and differences with respect to the CRISPR-Cas system.
- The next section is dedicated to developing a method for clustering orthologous CRISPR

arrays while accounting for the horizontal transfer of CRISPR arrays.

- The last section traverses the domain of phage satellites and PICIs, current focal points in MGE research. The study includes the reclassification of phage satellites based on their gene content and the identification of protein families and bacterial defense systems potentially linked to these satellites.

The concluding chapter (Discussion) summarizes the insights reached throughout the thesis and outlines potential avenues for future research based on these findings.

Chapter 3

Materials and methods

3.1 Comparison of gene clustering criteria in pangenome analyses

3.1.1 High-quality genomic sequences

High-quality genomes (according to the Minimum Information about a Metagenome-Assembled Genome (MIMAG) criteria (Bowers et al., 2017)) with completeness >99% and contamination <1% assessed by CheckM (Parks et al., 2015), mean contig length >5kb, and contig count <500 were parsed from Genome Taxonomy Database (GTDB) release 95 (Parks et al., 2020). Genomes from species (sensu GTDB) containing at least 15 high-quality genomes were downloaded from NCBI FTP site by their corresponding IDs. Metagenome-assembled genomes (MAG) and single-amplified genomes (SAG) were not included in the analysis. The 321 bacterial and 1 archaeal species, belonging to 125 different genera (sensu GTDB) were used in following analysis. To minimize possible taxonomical biases, only one representative species per genus was selected for subsequent analyses. For those genera with >1 suitable species, the species were kept with the highest number of high-quality genomes, as they presumably were the most informative for pangenome studies.

To keep our analyses within a manageable computational cost, species with >100 high-quality genomes were subsampled to keep at most 100 genomes per species. To that end, separately, the amino acid sequences of 120 nearly universal marker genes employed by the GTDB were aligned with MAFFT (Katoh and Standley, 2013), concatenated the alignments, and ran IQ-Tree (Nguyen et al., 2015) to obtain phylogenetic trees including all the strains of the same species. Then, a heuristic subsampling strategy was applied that maximized the diversity of

the subsampled genomes whilst accurately reflecting the topology of the strain trees.

The final dataset comprised 6,796 bacterial genomes belonging to 124 species and 55 archaeal genomes belonging to 1 species. ORF were predicted with Prodigal v2.6.3 (Hyatt et al., 2010), using codon table 11 and the "single" mode as recommended for finished genomes and quality draft genomes. Additionally, for the species *Mycoplasmopsis bovis PG45* (GCF_000183385.1) and *Mycoplasma pneumoniae FH* (GCF_001272835.1), code table 4 was applied.

3.1.2 De novo OGC construction

Eight sets of de novo species-level OGC were generated, each one representing an alternative approach to identify homologous genes and discriminate within-species paralogs (Table 3.1). To facilitate comparison among methods, started in all cases from the same predicted ORF previously obtained with Prodigal, overriding any optional ORF prediction step provided by the OGC construction software.

Table 3.1: Parameters and options for comparing pangenome clustering tools.

Method	**Version**	**Non-default parameters**
Prodigal	2.6.3	-p single (*)
Roary (95% identity)	3.13.0	-i 95 -g 100000 (**)
Roary (80% identity)	3.13.0	-i 80 -g 100000 (**)
eggNOG-mapper	5.0.2	-m diamond –tax_scope_mode narrowest –usemem
CD-HIT (50% identity)	4.8.1	-c 0.5 -G 0 -aL 0.8 -g 1 -M 8000 -n 3
CD-HIT (80% identity)	4.8.1	-c 0.8 -G 0 -aL 0.8 -g 1 -M 8000 -n 5
MMseqs2 (50% identity)	14-7e284	easy-cluster –cluster-mode 1 –min-seq-id 0.5
MMseqs2 (80% identity)	14-7e284	easy-cluster –cluster-mode 1 –min-seq-id 0.8
panX	1.5.1	-ngbk -rt 60
OrthoFinder	2.5.4	
mOTUpan	0.3.2	-c [OGC_matrix_file] -k [checkM_file]

* Additional option -g 4 to set the right translation table for Mycoplasmatales.
** Additional option -t 4 to set the right translation table for Mycoplasmatales.

Four sets of homology-based OGC were built with the sequence clustering tools MMseqs2

(Steinegger and Söding, 2017) and CD-HIT (Fu et al., 2012), setting the minimum identity threshold to 50% and 80%. MMseqs2 was run with the options 'easy-cluster' for cascaded clustering and 'cluster-mode 1' to define clusters based on connected components. The options for CD-HIT were set so that sequence homology was calculated locally, the alignments covered >80% of the longest sequence, and sequences were assigned to the best-matching cluster (-G 0 -aL 0.8 -g 1 -M 8000). The word size for CD-HIT (option -n) was set to 5 for minimum identity 80% and 3 for minimum identity 50%, as recommended by the developers. MMseqs2 and CD-HIT do not perform any kind of paralog splitting; therefore, the resolution of the resulting OGC only depends on the similarity threshold used for clustering.

Two sets of synteny-based OGC were built with the software Roary (Page et al., 2015), setting the minimum identity threshold to 80% and 95%. The Roary algorithm starts by pre-clustering highly-similar protein sequences with CD-HIT to obtain a smaller set of representative sequences. Roary then conducts an all-against-all comparison with BLAST and filters the hits based on the user-provided identity threshold. Based on the network of hits, representative sequences are clustered with Markov Cluster Algorithm (MCL) (Enright et al., 2002) and the resulting clusters are merged with the pre-clusters. As a final step, Roary uses conserved gene neighborhood information to split homologous groups containing paralogs into groups of synteny-supported OGC.

Two sets of orthology-based OGC were built with the software panX (Ding et al., 2018) and OrthoFinder (Emms and Kelly, 2019). Both algorithms initially cluster sequences in orthologous groups by performing an all-against-all similarity search with DIAMOND (Buchfink et al., 2021) and posterior clustering with MCL. The hits retrieved by DIAMOND are filtered only in terms of statistical significance, regardless of sequence identity, which allows recovering relatively divergent homologs. In the panX algorithm, the sequences of these initial clusters are aligned with MAFFT (Katoh and Standley, 2013) and cluster-level phylogenetic trees are built with FastTree (Price et al., 2010). Finally, panX obtains orthology-supported OGC by examining the resulting trees and applying a set of heuristic rules to split paralogs from true species-level orthologs. OrthoFinder directly builds phylogenetic trees from the DIAMOND scores, infers the root based on gene duplication patterns, and generates OGC that are compatible with the rooted trees.

3.1.3 OGC construction by reference database mapping

Reference-based OGC were built by mapping translated ORF to the eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) database version 5.0 (Huerta-Cepas et al., 2019). For that purpose, the eggNOG-mapper v2 (Cantalapiedra et al., 2021) was

ran with command line options `-m diamond-tax_scope_mode narrowest` to search queries against eggNOG sequences with DIAMOND and transfer orthologous group annotations at the highest possible taxonomic resolution (which typically corresponds to the genus level).

3.1.4 Pangenome features

OGC presence-absence matrices (sometimes known as phyletic profiles) and gene-to-OGC relationships were used to estimate a set of pangenome features that intend to capture both the size and the diversity of the pangenome.

Pangenome size was calculated as the total number of OGC retrieved for each species. Because this measure is positively correlated with the number of genomes sampled, an unbiased estimate was obtained by randomly subsampling sets of 14 genomes and taking the average over 100 realizations (we refer to this and other pangenome features obtained with the same procedure as "14-mean"). To ensure that the results obtained with these two measures were sufficiently representative, three additional measures of pangenome size were calculated: Chao's lower bound (Chao, 1987); the normalized pangenome size, obtained by dividing the pangenome size P_{tot} by the sum of the harmonic series of the number of genomes (n) such that $P_{\text{norm}} = \frac{P_{\text{tot}}}{\sum_{i=1}^{n-1} \frac{1}{i}}$ (Bobay and Ochman, 2018); and the pangenome size of the 15 most dissimilar genomes in terms of genome content. After verifying that all those metrics were highly correlated across species (R>0.95), proceeded with the uncorrected and 14-mean pangenome sizes, which are simpler and more easily interpretable.

The core genomes were defined as the set of OGC contained in all the genomes of a species. This strict definition is appropriate given the almost full (>99%) completeness filter imposed on high-quality genomes. Pangenomes were also characterized in terms of the absolute number of single-copy core OGC, accessory OGC (those that are not core), and singleton OGC (those encompassing a single ORF). As measures of genome content diversity, the mean percentage of OGC were computed that are accessory and singleton with respect to all the OGC present in each genome, averaged over 100 random subsamples of 14 genomes per species. In addition, a measure of genomic fluidity that quantifies the average dissimilarity in gene content of randomly sampled pairs of genomes (Kislyuk et al., 2011) was obtained with the 'fluidity' function of the R package micropan (Snipen and Liland, 2015). To ensure that variations in fluidity were not simply due to the randomness of the subsampling procedure, a series of preliminary tests were conducted to determine the minimum number of sampled pairs required for convergence in fluidity estimates. Using 500 random pairs (instead of only 10 pairs, which is the default in micropan), the relative variability across repeated estimates was reduced to below 0.5%. Therefore, for all fluidity calculations, the option `n.sim = 500` was

set.

The species-wise core gene alignment similarity (CGAS) was calculated using ORFs that belong to single-copy core OGC. Pairwise global alignments of all the ORFs assigned to the same OGC were performed with the Needleman-Wunsch algorithm as implemented by the needleall program of the EMBOSS suite (Madeira et al., 2019). After removing all self-alignments (alignments of ORFs with themselves), the average identity of each pair of genomes was calculated as $\frac{\sum_i m_i}{\sum_i L_i}$ where m_i and L_i represent the number of matches and total alignment length for the pair of sequences from the i-th OGC and the sum extends over all single-copy core OGC. The species-wise CGAS was obtained as the average over all pairs of genomes of the same species. The nucleotide sequence divergence in single-copy core genes was defined as one minus the species-wise CGAS. Given the high computational cost of this procedure, calculation of CGAS and core gene divergences was restricted to reference-, synteny-, and orthology-based OGC.

3.1.5 High-resolution species trees and inference of gene gains and losses

Phylogenetic trees were built for each species based on the all single-copy core OGC obtained by Roary.

Amino acid sequences for each single-copy core OGC were aligned with mafft-linsi (L-INS-I Algorithm, default options, v7.475) (Katoh and Standley, 2013) and back-translated to nucleotide alignments with the pal2nal.pl (v14) using codon table 11 (except for species of the order Mycoplasmatales, which use codon table 4) (Suyama et al., 2006). The nucleotide alignments were concatenated and used as input for FastTree (-gtr -nt -gamma -nosupport -mlacc 2 -slownni, v2.1.10) (Price et al., 2010). The tree topologies produced by FastTree were subsequently provided to RAxML (`-f e -c 25 -m GTRGAMMA -p 344312987`, v8.2.12) for branch length optimization (Stamatakis, 2014). The ETE3 (Huerta-Cepas et al., 2016) was used for mid-point rooting and visualization.

The branch-specific gene gains and losses were inferred with the phylogenomic reconstruction software Gloome (O. Cohen et al., 2010) using the presence or absence of each OGC and the high-resolution species trees. The parameter configuration was set to optimize the tree branch lengths under a genome evolution model with 4 categories of gamma-distributed gain and loss rates and stationary frequencies at the root (isRootFreqEQstationary=1). To compare OGC generation methods, the input phyletic profiles were varied according to the desired method while keeping the species trees unchanged. The Gloome optimization algorithm

did not converge for Enterobacter himalayensis and Chlamydia muridarum; therefore, those species were excluded from all the analyses involving gene gains and losses.

3.1.6 Functional annotation and statistical analysis of functional profiles

Functional annotation at the gene level was done by mapping individual genes to custom-made HMM profiles of the 2020 release of the COG database (Galperin et al., 2021). Functional annotation at the OGC level was done by applying the majority rule to gene-level annotations. Coarse-grained pangenome functional profiles were built by counting the number of OGC assigned to each of the 21 major prokaryotic functional categories defined in the COG database.

The statistical analysis of functional profiles was conducted by applying the phylofactorization framework (Washburne et al., 2017). First, to account for the constant-sum constraint that complicates the statistical analysis of compositional data, the isometric log-ratio (ILR) transform was applied to the species-wise functional profiles. Informative ILR balances were defined following a guide tree that was obtained by calculating the mean differences between synteny and orthology OGC for each functional category and performing hierarchical clustering of the functional categories based on such differences. Then, linear mixed effects models were set out for each of the 20 ILR balances, with OGC clustering criteria as fixed effects and species as random effects. Model fitting was performed with the R package lmerTest (Kuznetsova et al., 2017) and contrasts among balances were conducted with the R package phylofactor (https://github.com/reptalex/phylofactor) , which ranks the balances based on the fraction of the total variance that is explained by the model. Statistical significance was calculated by using Satterthwaite's approach to estimate the degrees of freedom of the F-statistic and applying Bonferroni correction to account for multiple comparisons.

3.2 Quantifying the effect of CRISPR-Cas and other defense systems immunity on gene gain and loss

3.2.1 Genome collection

The 82,595 high-quality (HQ) genomes (according to the Minimum Information about a Metagenome-Assembled Genome (MIMAG) criteria (Bowers et al., 2017)) with completeness >99%, contamination <1%, and contig count <500 included in the Genome Taxonomy Database (GTDB) release 202 (Parks et al., 2020) were downloaded from NCBI FTP site

by their corresponding IDs. Note that genomes from species that have less than 10 HQ genomes were not downloaded. CRISPRCasTyper (default parameters, v1.2.4 (Russel et al., 2020) was used to predict which of these genomes contained complete CRISPR-Cas systems (define by CRISPRCasTyper which contain CRISPR and Cas proteins). For the following analyses, we classified a genome as having CRISPR systems if it contains at least one complete CRISPR-Cas system, under the assumption that a complete set of Cas genes is required for CRISPR-Cas activity (Makarova, Aravind, et al., 2011). The species with more than 10 genomes, at least 5 of which contain CRISPR-Cas system (CRISPR(+) genomes), were considered in the following analysis (we did not require that at least 5 genomes do not have CRISPR-Cas).

To reduce computational costs and include as many genomes containing the CRISPR system as possible, while also facilitating comparison to genomes without CRISPR, a random subset of 500 genomes per species was selected (seed: 19940421) according to the following rules:

a) If the total number of genomes exceeds 500:

- If the count of CRISPR(+) genomes is greater than 350 and CRISPR(-) genomes is greater than 150, we randomly select a subset comprising 350 CRISPR(+) genomes and 150 CRISPR(-) genomes (non-CRISPR-cas systems).

- If the count of CRISPR(+) genomes is greater than 350 and CRISPR(-) genomes is less than 150, we randomly select a subset comprising 350 CRISPR(+) genomes and keep all CRISPR(-) genomes (non-CRISPR-cas systems).

- If the count of CRISPR(+) genomes is less than or equal to 350, regardless of the count of CRISPR(-) genomes, we keep all CRISPR(+) genomes and randomly select a subset of CRISPR(-) genomes based on a 7:3 ratio of CRISPR(+) genomes to CRISPR(-) genomes.

b) If the total number of genomes is less than or equal to 500, we include all genomes.

c) The representative genome of the species was added if the selected genome not included.

After applying these filters, 19,323 high-quality genomes belonging to 196 bacterial and 1 archaeal species (sensu GTDB) were recovered. The distribution of species within each phylum is as follows: 76 in Proteobacteria, 68 in Firmicutes, 17 in Firmicutes_A, 16 in Actinobacteriota, 15 in Bacteroidota, 2 in Fusobacteriota, 1 in Campylobacterota, 1 in Chloroflexota and 1 in Methanobacteriota (archaea) according to GTDB taxonomy.

3.2.2 Gene prediction and annotation

Open reading frames (ORF) were predicted with Prodigal v2.6.3, using codon table 11 (prokaryotic genetic code) and "single" mode, as recommended for finished and draft quality genomes (Hyatt et al., 2010). Orthologous ORF were then clustered with Roary v3.13.0 (Page et al., 2015) setting an 80% identity threshold for initial clustering followed by synteny-based refinement (Codon table 11 and 80% sequence identity). The resulting gene clusters were functionally annotated by selecting a representative sequence, arbitrarily chosen among those with length greater than 0.95 times the average length of all sequences in the cluster (value generate by Roary). Representative sequences were annotated by mapping them to custom-made profiles of the Clusters of Orthologous Genes (COG) database (2020 release) (Galperin et al., 2021) with HMMER (e-value=0.001) (Eddy, 2011) available in eggNOG-mapper (v2.1.9) (Cantalapiedra et al., 2021). The 26 major prokaryotic functional categories defined in the COG database were assigned to the annotated genes. The functional categories (A, RNA processing and modification; B, chromatin structure and dynamics; W, extracellular structures; Y, Nuclear structure; and Z, cytoskeleton) were not further considered since they are exclusively eukaryotic or rarely occur in prokaryotic genomes (Galperin et al., 2021).

3.2.3 Species trees

Phylogenetic trees were built for each species based on a set of prokaryotic marker genes proposed by the GTDB, release 202 (see https://github.com/lyonliuyang/phd_thesis_supp). For each species, only those marker genes with prevalence >80% of our data-set were used for phylogenetic reconstruction. The codon alignment workflow was developed in a previous study (Y. Liu et al., 2022): Amino acid sequences for each marker gene were aligned with mafft-linsi (L-INS-I Algorithm, default options, MAFFT v7.475) (Katoh and Standley, 2013) and back-translated to nucleotide alignments with pal2nal.pl (v14) (Suyama et al., 2006) using codon table 11. The nucleotide alignments were concatenated and used as the input for FastTree (-gtr -nt -gamma -nosupport -mlacc 2 -slownni, v2.1.10) (Price et al., 2010). The tree topologies produced by FastTree were subsequently provided to RaxML v8.2.12(raxmlHPC -f e -c 25 -m GTRGAMMA -p 344312987) (Stamatakis, 2014) for branch length optimization. The ETE3 (Huerta-Cepas et al., 2016) was used for mid-point rooting and visualization.

Moreover, a super bacterial tree containing 19310 genomes was constructed by using GTDB marker genes and applying 80% prevalence cut off using the same procedure described above. A multispecies representative genomes (RG) tree was obtained by downloading the GTDB `bac120.tree` (release r202) and subsequently purging irrelevant genomes.

3.2.4 Comparison of gene gain/loss rates between CRISPR(+) and CRISPR(-) clades

For the sake of computational efficiency, a subset of 9,137 genomes was randomly chosen (using seed: 19940421). This subset was selected to include a maximum of 100 genomes per species, chosen randomly from a pool of 19,323 genomes. The selection process adhered to predefined criteria: each species included a maximum of 100 genomes, a minimum of 10 genomes were represented, at least 5 genomes per species contained CRISPR arrays, and no more than one-third of the selected genomes lacked CRISPR arrays, as described previously. Using ETE 3 (Huerta-Cepas et al., 2016), CRISPR(+) (respectively CRISPR(-)) clades were defined as clades with >80% of the corresponding genomes being CRISPR(+), and its counterpart sister clades CRISPR(-) where with <80% of the corresponding genomes being CRISPR(-). Sister CRISPR(+) and CRISPR(-) clades were excluded from further analysis if both were exclusively formed by a single genome. The branch-specific gene gain and loss rates of each clade were inferred with the phylogenomic reconstruction software Gloome (release May 2013) (O. Cohen and Pupko, 2010) using the gene presence/absence matrix obtained by Roary and species trees. The parameter configuration (see https://github.com/lyonliuyang/phd_thesis_supp) was set to optimize the tree branch lengths under a genome evolution model with 4 categories of gamma-distributed gain and loss rates and stationary frequencies at the root. Also, the function skewtest from the scipy python package v.1.10.1, that uses the method proposed by (D'agostino et al., 1990) was employed to test whether the skew is different from the normal distribution.

Additionally, we computed the gain and loss rates associated with each gene functional category and type of MGE. To determine gene function-specific gain and loss rates, we categorized genes based on NCBI COG annotation categories (Galperin et al., 2021). Furthermore, we partitioned clades into Positive correlated species (PCS) and Negative correlated species (NCS), based on a binomial Phylogenetic Generalized Linear Mixed Model (PGLMM) association analysis for the presence of CRISPR and the number of genes in the mobilome (functional category "X"; see section 3.2.6 for details). Additionally, for each MGE class, we based our analysis on the respective MGE genes, employing Poisson PGLMM. For MGE-specific gain and loss rates, we employed a straightforward approach by searching for prophages, plasmids, and transposons using keywords (no case distinction) such as 'phage,' 'plasmid,' and 'transpos' within the COG annotation profiles, respectively. The species-wise singleton genes (i.e. those representative genes that correspond to Roary Gene Clusters with only one gene) were not included in gain and loss analyses tailored to a particular gene functions or MGEs. This is because singleton genes are likely to evolve neutrally and their fitness effects of gene gain and

loss are negligible (Wolf et al., 2016). Finally, the Wilcoxon signed-rank test for matched pairs was used to check if there is a significant difference between CRISPR(+) and CRISPR(-) sister clades.

Phylogenomic analysis with GLOOME (see alsoANNEX):

Model selection. GLOOME allows two options for the inference of ancestral gene content in the root of the tree. With one option, it assumes that the gene gain and loss process is at equilibrium and the probability to find a given gene family at the root of the tree is given by its stationary probability. With the alternative option, there is no assumption of equilibrium but in turn it assumes that the probability at the root is the same for all gene families. We ran GLOOME with both alternative options in a subset of 100 species and compared their performance in terms of the log-likelihood of the model and the distribution of predicted gain and loss rates across species. In both cases, the stationary root model produced more consistent results (higher log-likelihood in most species and fewer extreme outliers in gain and loss rates). Therefore, we set the stationary root option for the rest of the analyses.

Calculation of branch rates. The rate of gene gain and/or loss per branch is calculated as the number of expected events in that branch divided by the original tree branch length, which represent evolutionary time in units of substitutions per site. The tree produced by Gloome cannot be utilized as it already represents evolutionary time in units of gain/loss events.

For example, consider the clade that includes *Pseudomonas aeruginosa* PA14, and their ancestor (N14). The gain rate for that clade can be calculated as following:

Gain rate is calculated as (Figure 3.1):

$$\text{gain rate} = \frac{event1 + event2 + event3}{branchLength1 + branchLength2 + branchLength3} = \left[\frac{\text{gain events}}{\text{substitutions/site}}\right]$$

To obtain the loss rate, the procedure is the same, but using the loss events instead of gain events in the numerator.

Additional notes:

- Branches are named based on the child node. For example, the branch named *Pseudomonas aeruginosa* PA14 in mydata.txt corresponds to the branch that goes to the ancestor N14 to the *Pseudomonas aeruginosa* PA14. Likewise, the branch named N14 goes from the root (N1) to N14.

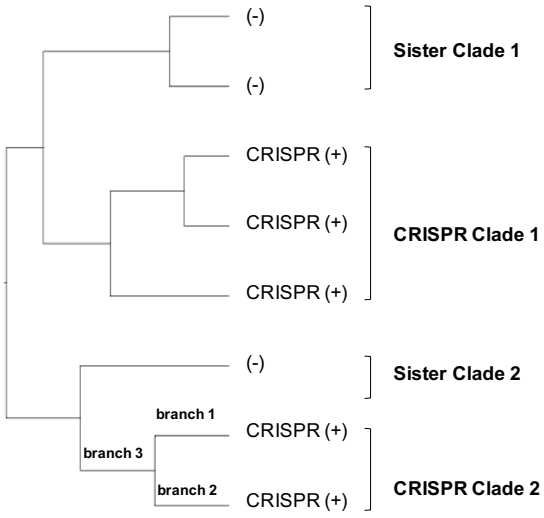

Figure 3.1: Definition of CRISPR and counterpart sister clades represented by a tree structure.

3.2.5 Identification of mobile genetic elements

Transposons: A heuristic sliding window search was applied to identify the number of transposons (mainly Insertion Sequence) in each genome. For each transposon, we searched for predicted genes whose COG annotation contained the keyword "transposase" (without case distinction). Our criteria for distinguishing individual transposons was to evaluate the genomic coordinates of these proteins. We considered proteins to belong to the same transposon if they were contiguous or separated by a single non-transposon-related protein. Conversely, if there were at least two intervening genes without any transposon-related annotations between two proteins with 'transposase' annotations, we classified them as separate transposons.

Prophages: The Prophages were predicted for 19,323 genomes using phispy (v4.2.21) (Akhter et al., 2012) with default parameters (the –phage_genes option is 2, look for two or more genes that are likely to be a phage in each prophage region).

Plasmids (only in complete genomes): The extra-chromosomal replicons labeled as 'plasmid' (without case distinction) in their NCBI genome file contig descriptions were counted.

ICE and IMEs (only in complete genomes): After removing the plasmid-type replicons, ICEfinder v1.0 (ICEberg 2.0 database) (M. Liu et al., 2019)was used to identify ICEs and IMEs.

Due to the inherent difficulty of discriminating among plasmids, ICE, and IME (Botelho et al., 2023), 2,964 complete genomes out of a total of 19,323 genomes were analyzed to determine the number of plasmids, ICEs, IMEs.

3.2.6 Statistical (Phylogenetic Generalized Linear Mixed Model) analysis of functional profiles

For each species, we counted the number of genes belonging to each functional category in genomes that do and do not harbor CRISPR-Cas systems (CRISPR(+) and CRISPR(-), respectively). Then, we compared both sets of values by applying five complementary approaches (also see code below):

a) To avoid potential phylogenetic biases, a PGLMM with Poisson distribution and canonical link function was fit using the pglmm_compare function resource code from the phyr (v1.1.0) (Li et al., 2020) package, each comparison considering the presence/absence of the CRISPR-Cas system as the predictor; the total number of genes in the genome (except those that are part of CRISPR-Cas) as the response variable, and the corresponding species tree for accounting the phylogeny.

b) Using method a), instead of total number of genes, we separated number of genes per functional category (excluding those in the CRISPR-Cas system that is being analyzed) vs presence/absence of the CRISPR-Cas.

c) Since genome size correlates with the presence/absence of CRISPR-Cas system in some species (eg. *Pseudomonas aeruginosa*) (Wheatley and MacLean, 2021), the PGLMM analysis described in b) was repeated using the fraction (instead of the number) of genes of the functional category, with respect to total number of functionally classified genes as the response variable (always excluding genes in the CRISPR-Cas system that is being analyzed). Due to the change in the nature of the response variable, a binomial distribution with its corresponding canonical link was selected for the fitting. we referred to it as 'genome size corrected PGLMM'.

d) We also incorporated supplementary analyses involving the quantity of genes corresponding to Prophage, Plasmid, Transposon (by keywords from NCBI COG annotation descriptions, 'phage,' 'plasmid,' and 'transpos', no case distinction). We fitted the PGLMM parameter as b) described above, with Prophage/Plasmid/Transposon genes used as independent variables separately.

e) Instead of conducting species-specific analyses as described in method b), we opted for a different approach. Here, we employed the total number of genes across 19310 bacterial genomes (excluding Archaeal genomes) as the response variable. We constructed a comprehensive super tree that encompassed all bacterial genomes included in this study. Subsequently, we utilized the presence/absence of the CRISPR-Cas system as a predictor variable and fitted

a Poisson distribution-based PGLMM to the data.

Then, for subsequent comparative analysis, we defined species as Positive Correlated Species (PCS), Negative Correlated Species (NCS), and No Correlated Species (NS) based on the association between the presence of CRISPR-Cas systems and the amount of MGE genes (X category) (c) or each type of MGE genes (d) according to effect size. The list of thresholds for the effect size are shown in the (Table 3.3).

Table 3.3: List of effect size thresholds for PCS and NCS by CRISPR-Cas Poisson PGLMM.

Category	Thresholds
X	±0.088014
Phage	±0.097307
Plasmid	±0.109001
Transposon	±0.192572

In our analysis, species categorization (sign of the correlation) was determined based on the sign and magnitude of the effect size from the MGE genes (X) PGLMM (binomial distributed). Specifically, NCS were defined as those species whose effect size was below the effect size of the first significant data point ($p<0.05$) Conversely, PCS were defined as those species whose effect size exceeded that of the first significant data point ($p<0.05$)

```
# The core code of implementation of Phylogenetic Generalized Linear
    Mixed Model:
# data table, tree and covariance matrix
dat <- read.csv(data_path,sep = '\t')
phy <- ape::read.tree(tree_path)
Vphy <- ape::vcv(phy)
Vphy <- Vphy/max(Vphy)
Vphy <- Vphy/exp(determinant(Vphy)$modulus[1]/ape::Ntip(phy))
re.1 <- list(covar = Vphy)

# sort data table and trim tree.
sp <- rownames(dat)
if(!all(is.element(sp, phy$tip.label)))  stop("\nSorry, but it
   appears that there are some species in the rownames of data that
   are not in phy")
if(!all(is.element(phy$tip.label, sp)))  {
  warning("\nIt appears that there are some species in phy are not
```

```
    contained in the rownames of data; we will drop these species")
  phy <- ape::keep.tip(phy, sp)
}
if(any(sp != phy$tip.label)){
  warning("\nThe data rows are resorted to match phy$tip.label")
  #dat <- dat[match(sp, phy$tip.label),]
  dat <- dat[match(phy$tip.label, sp),]
}

# Total number of genes (except those in the CRISPR-Cas systems that
    is being analyzed) vs presence/absence of the CRISPR-Cas system.
formula <- TOTAL_GENE_COUNTS ~ HAS_CRISPR
family <- 'poisson'
# Number of genes per category (excluding those in the CRISPR-Cas
   system that is being analyzed) vs presence/absence of the CRISPR-
   Cas system.
formula <- CATEGORY_GENE_COUNTS ~ HAS_CRISPR
family <- 'poisson'
# Binomial analysis for the number of genes per category with
   respect to total of functionally classified genes ("successe/
   failures", always excluding genes in the CRISPR-Cas system that
   is being analyzed).
formula <- cbind(CATEGORY_GENE_COUNTS,SUM_OF_CATEGORY_GENE_COUNTS -
   CATEGORY_GENE_COUNTS) ~ HAS_CRISPR
family <- 'binomial'
# Number of MGE genes vs presence/absence of the CRISRP-Cas system.
formula <- CATEGORY_MGE_GENE_COUNTS ~ HAS_CRISPR
family <- 'poisson'

# fitting model
mod <- phyr::pglmm(formula, data = dat, random.effects = list(re.1)
   , family = family, optimizer = c("nelder-mead-nlopt", "bobyqa", "
   Nelder-Mead", "subplex"), REML = TRUE, add.obs.re = TRUE, verbose
    = FALSE, cpp = TRUE, bayes=FALSE, reltol = 10^-6, maxit = 500,
   tol.pql = 10^-6, maxit.pql = 200, marginal.summ = "mean", calc.
   DIC = FALSE, prior = "inla.default", prior_alpha = 0.1, prior_mu
   = 1, ML.init = FALSE, s2.init = 1, B.init = NULL)
```

```
# Note that, we found a bug in the code of pglmm_compare. As a
   consequence, pglmm_compare produces wrong results if the rows of
   the original data table were not already sorted as the tips of
   the tree. To fix the bug, we reversed function ``dat <- dat[match
   (sp, phy$tip.label),]'' to ``data <- data[match(phy$tip.label, sp
   ),]''.(https://github.com/daijiang/phyr/issues/88)
```

Listing 3.1: The core code of implementation of Phylogenetic Generalized Linear Mixed Model

3.2.7 Location of CRISPR genes with respect to the MGEs

One possible explanation of the presence of a positive correlation between the presence of CRISPR-Cas systems and the abundance of MGEs is that the CRISPR-Cas genes are, in fact, contained in these MGEs. To test this hypothesis, for complete genomes, a Generalized Linear Mixed Model (GLMM) was fit by glmer() from lme4 (Bates et al., 2014) using the sign of the correlation between the presence of CRISPR-Cas systems and the abundance of MGE units as the predictor, the presence of CRISPR-Cas genes inside (1) or outside (0) the MGEs as the response variable (binomial distribution). In addition, species were added as a random effect to the model nested in sign of correlation since it is the cause of the correlation signs. The coordinates of prophages, ICEs, IMEs, and plasmids were previously obtained (see Section 3.2.5). Since the method for identifying transposons was targeted at single gene elements (short Insertion Sequences), and can not have a CRISPR system inside them, transposons are not included in this analysis.

3.2.8 Masking functional genes within MGEs

Some genes bolonging to MGE might be classified into other gene functional categories (Galperin et al., 2021). To identify and mask such genes, we focused on complete genomes and used previously detected MGEs (Plasmids, Prophages, ICEs, IMEs and Transposons) to mask the annotation profiles of all predicted genes whose coordinates overlap with previously identified MGEs. Moreover, we ran Phispy with parameter –phage_genes 0 (v4.2.21,) to identify other MGE, integrons, pathogenicity islands, and fragments of MGE that could have been previously missed in the complete genomes. Although it has been reported that Phispy occassionally identifies ribosomal RNA operons as MGE, such misidentification does not affect our analysis because ribosomal RNA genes are not included in any of the COG categories tested with PGLMM. The genome size-corrected binomial PGLMM was fitted as previously described (see (c) in Section 3.2.6), with CRISPR presence/absence as predicted variable and the number of genes which belong to NCBI COG annotation as response variable. By

performing a comparative analysis of the annotation profile with and without masking MGEs, we aim to discern whether the correlations with the presence of CRISPR in different COG classes are exclusively due to genes located within MGEs.

3.2.9 Exploring other factors that potentially lead to discrepancies in species-depended CRISPR effects

CRISPR type

First, genomes associated with ambiguous or unknown CRISPR-Cas systems were removed from the dataset. Second, a species was considered to contain a particular CRISPR type if at least two CRISPR(+) genomes belonging to that species were of that particular type. CRISPR types VI-B, III-B, V-A, II-B, I-G, and III-D were excluded from the statistical analysis because they were present in less than 5 of the 197 species considered in this study. Finally, a chi-squared test was applied to test whether the number of PCS, NCS, and NS was independent of the CRISPR-Cas type considered.

CRISPR plasticity

The most conserved Cas proteins in CRISPR-Cas system is Cas1, and it has been considered the signature for the presence of CRISPR-Cas systems in a genome (Makarova, Haft, et al., 2011; Makarova and Koonin, 2015; Haft et al., 2005; Makarova et al., 2006). The evolutionary plasticity of CRISPR was evaluated by the number of gains and losses of *cas1* genes. The *cas1* gene clusters were identified among the descriptors obtained after the gene function annotation process, and gain plus loss rates of *cas1* genes were previously inferred by GLOOME. Then two-side Mann–Whitney U test was used to evaluate the difference between PCS, NCS, and NS.

$$\text{CRISPR plasticity} = \frac{\sum_{\text{species}}(\textit{cas1}\ \text{gain events}) + \sum_{\text{species}}(\textit{cas1}\ \text{loss events})}{\sum_{\text{species}}\text{tree lengths}}$$

GC content, Genome size, Pan-genome size

Genome sizes and GC contents were obtained from GTDB metadata and averaged for each species. The pan-genome size of species that have more than 15 genomes were measured by randomly subsampling sets of 14 genomes and taking the average over 100 realizations (Manzano-Morales et al., 2023).

3.2.10 Defense systems

Proteins predicted by Prodigal were employed as input for identifying defense systems using Padloc v1.1.0 (DB v1.4.0). Our analysis focused on the most prevalent systems, including

RM, DMS, Abi, Gabija, DRT, and CBASS. We generated a matrix to represent the presence and absence of these defense systems. The PGLMM, detailed in the section 3.2.6 (refer to a, b, c, d), was used to analyze the presence or absence of each defense system:

(I) We used a Poisson distribution PGLMM for each comparison, considering the presence or absence of the defense system as the predictor. The response variable was the total number of genes, excluding those in the defense system under analysis, and the phylogenetic relationships were accounted for using the corresponding species tree.

(II) Using method (I), instead of total number of genes, we separated number of genes per category (excluding those in the defense system that is being analyzed) vs presence/absence of the defense system.

(III) The Binomial PGLMM analysis, as described in (II), was further adapted by using the proportion of genes within each functional category (relative to the total number of functionally classified genes), excluding genes in the analyzed defense system, as the response variable.

(IV) The quantity of MGE genes was determined by keywords search from NCBI COG annotation descriptions, 'phage,' 'plasmid,' and 'transpos' (no case distinction). The PGLMM parameter was fitted as a Poisson distribution, as described above, with the number of prophage, plasmid, or transposon genes as the independent variable and the presence or absence of the defense system as the dependent variable, analyzed separately.

Finally, we applied a binomial PGLMM for each species for each pair of defense systems. This model evaluated the presence (1) or absence (0) of one defense system as the predictor variable and the presence (1) or absence (0) of another defense system as the response variable. Additional parameters were utilized as previously describe in section (see Section 3.2.6)). Correlations were considered significant at a p-value threshold of less than 0.05.

3.2.11 Anti-defense system

Anti-defense system proteins, including anti-CRISPR, were identified using hmmsearch (Eddy, 2011) against the dbAPIS database (Yan et al., 2024), applying a threshold of E-value lower than 1×10^{-10}. The fraction of genomes having at least one anti-CRISPR was obtained from CRISPR(+) and CRISPR(-) genomes species-wise (PCS, NCS, and NS), and statistical significance was sassessed with the Chi-squared test. Next, the overall fraction of genomes harboring anti-CRISPR in the CRISPR(+) and CRISPR(-) genomes were compared between PCS and NCS. The analyses were conducted separately in genomes with and without prophages. Confidence intervals for the overall proportions of a particular anti-CRISPR protein were

calculated using the beta distribution.

3.2.12 Identification of CRISPR and DF inside MGE

The coordinates of CRISPR operons and other defense systems were obtained from CRISPRCasTyper (Section 3.2.1) and padloc (Section 3.2.10), respectively. The coordinates of MGEs were obtained previously (Section 3.2.5). The number of Cas proteins and defense system proteins was categorized based on their location inside or outside the MGEs.

3.2.13 Estimating the overall effects of the CRISPR-Cas system over phylogenetic time.

To evaluate the difference of gene gain and loss between CRISPR presence and absence over evolutionary time, we employed the subtree depth to identify recent and old events.

The tree depths were determined locally by navigating the tree from root to leaves. For each node, we considered the subtree with that node as the root and evaluated whether the distances from that node to all the leaves were smaller than the depth threshold. If this condition was met, the descent from that node was halted, and the current subtree was considered valid. This approach introduces a tolerance: if all node-to-leaf distances except one are shorter than the threshold, it is also considered a valid subtree.

Finally, the gains and losses between CRISPR and sister clades for those subtrees were evaluated. The study required that CRISPR and its related branches be part of the same subtree. In other words, only pruned subtrees containing both the CRISPR and sister clade were considered for calculating gain and loss rates. Any pruned subtrees that contained only leaves were ultimately discarded.

3.3 CRISPR arrays clustering

By using the CRISPR prediction, gene prediction, species tree, and Roary OGC, previously collected from 19,323 HQ genomes. We create an orthologous CRISPR database:

a) CRISPR array content analysis. For each species, we employed CD-hit to cluster all CRISPR spacers and repeats nucleotide sequences. Specifically, we used cd-hit-est with a 90% similarity and 90% coverage cut-off for spacers (-g 1 -d 0 -c 0.9 -aL 0.9) and a 100% similarity and 100% coverage cut-off for repeats (-g 1 -d 0 -c 1 -aL 1) (Fu et al., 2012). Prior to the CD-hit clustering of repeats, CRISPRs with reverse complement repeats were identified and

classified as having identical repeats. We then assessed the Jaccard distance between pairs of arrays from the same species, which have identical repeats, by evaluating their spacer content conservation.

b) CRISPR array neighbor genes analysis. - we collected the gene family IDs (produced by Roary) from the 10 genes located immediately upstream and downstream of the arrays, respectively.

- We then excluded CRISPR arrays located in short contigs, defined by having fewer than 10 neighboring genes both upstream or downstream. Arrays situated at the beginning or end of a contig, characterized by the absence of upstream or downstream neighboring genes, were also filtered out.

- To identify and merge tandem arrays, we constructed a species-wise spacer sharing network, including CRISPR arrays in the same species that have the same repeats (based on Jaccard distance), identified the connected components (NetworkX), and merged arrays from the same genome if they belong the same connected component.

- At this point, we refine the Jaccard distance matrix for CRISPR arrays by incorporating information about tandem arrays

c) Bidirectional best match and community detection.

Incorporated with reverse complement repeats and tandem CRISPR arrays information, we built a neighborhood similarity (Levenshtein distance) network using the "bidirectional best match" method (see ANNEX). The goal of the method is to find pairs of CRISPR arrays, one from each of the two genomes, that are each other's best match in terms of conserved synteny and, in consequence, likely share a common evolutionary origin. We combined all pairwise best matches into a neighborhood similarity network and generated clusters of arrays by applying a information theory-based community detection algorithm (Infomap, https://mapequation.org) to the network (num_trials=100, seed=42, flow_model=undirected).

d) Cluster refinement. We used undirected Order Statistics Local Optimization Method (OSLOM) network (-uw -r 10 -hr 0) v2.5 (Lancichinetti et al., 2011) to refine the communities previously found by Infomap based on statistical significance.

e) Tree-based refinement. Finally, we subdivided the CRISPR Orthologous Groups by tree-based refinement, that is, we split candidate clusters of arrays that were not inferred as monophyletic (i.e., if the last common ancestor of two genome harboring a CRISPR arrays from the same candidate cluster lacked CRISPR-Cas) based on ancestral reconstruction of CRISPR-Cas presence/absence with PastML v1.9.34 (Ishikawa et al., 2019).

We then quantified the activity of CRISPR arrays. For each cluster of orthologous CRISPR arrays, the level of activity was quantified by comparing the array divergence (Jaccard distances measured through differences in the spacer content) with the phylogenetic distances (nucleotide-level) among the genomes. We adopted a constant divergence model, according to which the similarity in spacer content S decreases with phylogenetic distance t as $\frac{dS}{dt} = -r \cdot S$, where the divergence rate r is a proxy for the activity of the CRISPR array. Accordingly, the divergence rate was inferred by fitting the empirical data to the curve: $D = 1 - \exp(-rt)$, where $D = 1 - S$ is the Jaccard dissimilarity index.

3.4 Comprehensive functional and evolutionary analysis of a large collection of Phage-inducible chromosomal islands

3.4.1 Datasets

Our approach to selecting datasets for this study was driven by a desire to use the most recent and comprehensive data available. To this end, we included our own dataset (Michael Widdowson), enriched by the inclusion of recently published data from Dr. Eduardo P. C. Rocha (de Sousa et al., 2023). Including PICI: phage inducible chromosomal islands, CF-PICI: capsid-forming PICI, SaPI: Staphylococcus aureus pathogenicity islands, PLE: PICI-like elements, and P4-like satellites.

3.4.2 Structure based phage satellites protein clustering

The protein language models were used to determined similarity of satellites proteins, the analysis involved several key steps :

(a) Protein Sequence Filtering: The satellite protein sequences exceeding 3000 amino acids in length were removed from dataset.

(b) Redundancy Reduction: CD-HIT was used to eliminate redundant proteins, setting a threshold of 90% sequence identity and 95% alignment coverage.

(c) Protein Embedding-based Alignment (EBA): This step, detailed in (Pantolini et al., 2022), involved embedding satellite proteins into a matrix and comparing matrices as per the method and guidelines provided in the GitHub repository (https://git.scicore.unibas.ch/schwede/EBA). For each protein sequence, we extracted amino acids into a matrix (one vector per residue) using the protT5_ext.extract() function. Protein pair

similarities were then computed using the sm.compute_similarity_matrix() function and scored using eba.EBA(). Comparisons with an eba_max value above 3 were retained.

(d) Diamond Blastp Analysis: All-against-all protein sequence comparison was performed using parameters: –threads 40 –ultra-sensitive -k0 –unal 1 –no-self-hits.

(e) The final similarity score was a summation of the min-max normalized EBA average score and BLASTP bitscore. Comparisons of protein pairs with a similarity score below the threshold were excluded from the analysis.

$$\text{Similarity} = \text{normalized}\left(\frac{1}{2} \times (\text{eba\_max} + \text{eba\_min})\right) + \text{normalized}(\text{Blastp Bitscore})$$

(f) Protein Clustering: The undirected and weighted protein network clustering was conducted using the Infomap default parameters with Cytoscape (Shannon et al., 2003) plugin clustermaker (Morris et al., 2011), with parameter 10 number of trials. Protein family clusters were annotated using hmmsearch (Eddy, 2011)), with HMM profiles obtained from Dr. Eduardo P. C. Rocha (de Sousa et al., 2023) and Pfam (Mistry et al., 2021) (see https://github.com/lyonliuyang/phd_thesis_supp).

3.4.3 Phage satellites gene sharing network

To eliminate redundant phage satellites, we computed their average nucleotide identity (ANI) using skani (Shaw and Yu, 2023). We set specific parameters for this analysis, including -E, –slow, -t 40, -m 220, -s 75, and –robust, to ensure accurate and robust comparisons. The redundant sequences were identified by 90% coverage and 95% ANI.

To calculate the similarity of each pair of phage satellite genomes, we used the protein clusters to compared the protein families shared between them, the similarity score will calculation using equation below (Bin Jang et al., 2019):

$$P(X = c) = \frac{\binom{a}{c}\binom{n-a}{b-c}}{\binom{n}{b}}$$

where, c is the number of shared protein clusters between two phage satellites, a and b are the total numbers of protein clusters (including singletons) in each of the two phage satellites, and n is the total number of protein clusters (including singletons) in the dataset.

The Similarity Score is calculated by taking the negative logarithm (base 10) of the product of the probability P and the total number of comparisons T in the dataset.

$$\text{Similarity Score} = -\log_{10}(P \times \text{T})$$

Where P is the probability of observing a specific number of proteins shared between a pair of satellite sequences and T is the total number of comparisons. T can be calculated as $\frac{\text{N}\times(\text{N}-1)}{2}$, where N is the number of phage satellites in the dataset.

The satellite gene sharing network was constructed by treating each phage satellite genome as a node. The weight of each edge was determined based on the similarity score using the Weighted Louvain Clustering method (Blondel et al., 2008).

3.4.4 Phage satellites and adjacent gene prediction and annotation

We predicted satellite host genes using Prodigal-GV (Camargo et al., 2023) applied to host genomes. We extracted these satellites and their adjacent 10 upstream and 10 downstream genes using their genomic coordinates. To detect defense systems in satellites and their neighboring genes, we employed Padloc (Payne et al., 2021) and Defense-Finder (Tesson et al., 2022). ARGs were identified using the Comprehensive Antibiotic Resistance Database (CARD) through the Resistance Gene Identifier (RGI) tool (Alcock et al., 2023), setting the cutoff to retain Perfect/Strict labeled hits. For virulence proteins, we conducted a Blastp search using the Virulence Factor Database (VFDB) full protein dataset (B. Liu et al., 2022). Criteria for retention included a bitscore greater than 100, and for proteins with multiple annotations, only the highest-scoring hit was considered. Anti-CRISPR proteins were identified using Blastp against the anti-CRISPRdb (Dong et al., 2018). Lastly, anti-defense systems were identified using hmmsearch (Eddy, 2011) against dbAPIS (Yan et al., 2024), with a filter set for an E-value less than 1×10^{-10}. For the visualization, we utilized genus trees downloaded from the Genome Taxonomy Database (GTDB) release 207.

Chapter 4

Results

4.1 Operational Gene Clusters and intrinsic uncertainty in pangenome analyses

4.1.1 Method-dependent variation and intrinsic uncertainty in pangenome size and diversity

Five primary methods were investigated for *de novo* species-specific OGC formation, each with different strategies for gene clustering and differentiation of paralogs (see Table 4.1). All *de novo* methods begin with a set of open reading frames (ORFs) and cluster them based on a preset identity threshold that can be adjusted to include closer or more distant homologs. The resulting clusters are then processed to separate paralogs into distinct OGCs. The reference-based method maps the ORF to a reference database of orthologous groups, but it does not account for variability between strains. Additionally, the eggNOG database requires orthologous groups to contain sequences from at least three species. Mapping ORFs to eggNOG automatically excludes sequences without known homologs and sets a hard limit on the taxonomic resolution of OGC at the genus level. However, reference-based orthology assignments are now highly efficient and scalable for large (meta)genomic datasets.

The study covers techniques showcasing three different approaches used by several leading tools for pangenome analysis:

- orthology-based clustering (implemented by panX and OrthoFinder).
- synteny-based clustering (implemented by Roary).
- homology-based clustering (implemented by CD-HIT and MMseqs2, which is the OGC

Table 4.1: OGC generation strategies and tools used in this study.

Clustering	Paralog splitting	Identity threshold	Software	Ref.
de novo	orthology-based	none (e-val < 0.001)	panX	(Page et al., 2015)
de novo	orthology-based	none (e-val < 0.001)	OrthoFinder	(Ding et al., 2018)
de novo	synteny-based	80%, 95%	roary	(Fouts et al., 2012)
de novo	none	50%, 80%	MMseqs2, PanACoTA	(Steinegger and Söding, 2017; Altenhoff and Dessimoz, 2012)
de novo	none	50%, 80%	CD-HIT	(Fu et al., 2012)
reference db	n.a.	none (e-val < 0.001)	eggNOG-mapper	(Buchfink et al., 2021)

construction module used by PanACoTA (Perrin and Rocha, 2021) and PPanGGOLiN (Gautreau et al., 2020).

- Reference-based clustering (implemented by eggNOG-mapper)

For Roary, CD-HIT, and MMseqs2, we investigated the impact of implementing two distinct identity thresholds during the initial clustering stage. It is important to note that panX and OrthoFinder do not filter gene clusters based on sequence identity, but rather on their e-value. (Table 3.1) provides a summary of the optional settings used with each tool.

By applying these methods, we obtained nine alternative sets of species-wise OGC for 124 bacterial and one archaeal species. These species were selected to cover every genus in the Genome Taxonomy Database (GTDB), as defined by the phylogenetically consistent classification scheme established by GTDB (Parks et al., 2018), with the condition that there were at least 15 high-quality genomes available per species. The results can be found at https://dx.doi.org/10.5281/zenodo.7093013. The normalized variation of information (NVI) was utilized to measure the discordance between the OGC generated by each pair of methods. NVI takes values between 0 (if the two sets of OGC exhibit a one-to-one correspondence) and 1 (if they are entirely independent). The hierarchical clustering of methods based on this inidicates shows that the discordances, although small, are reproducible across species (refer to Figure 4.1). The dissimilarities between reference-based and *de novo* OGC are most notable. The resulting OGC is determined to a greater extent by the strategy for paralog

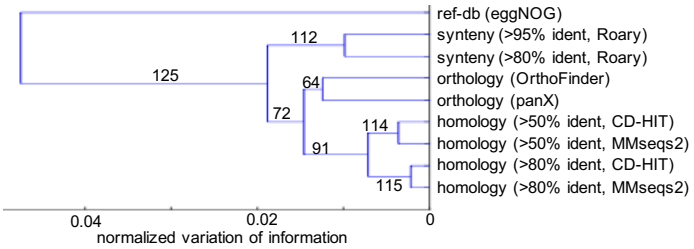

Figure 4.1: Consensus similarity tree of OGC building methods based on the species-wise normalized variation of information for the assignation of ORF to OGC. Labels indicate the number of species (out of 125) that support each branch.

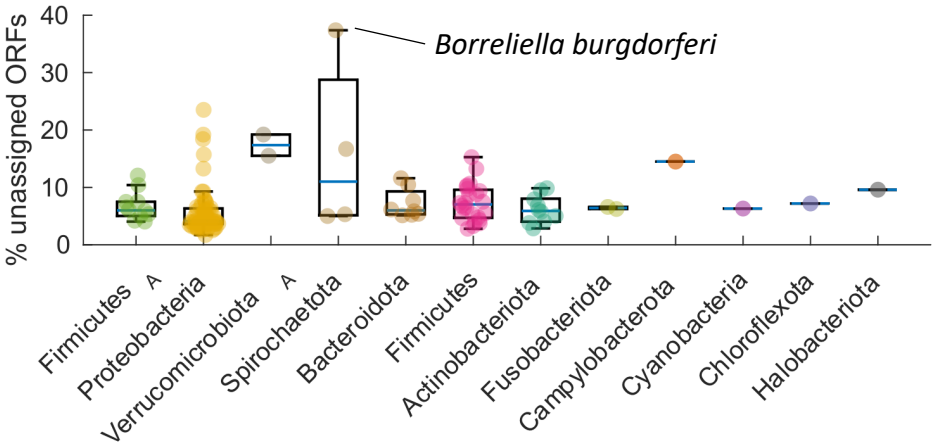

Figure 4.2: Fraction of ORFs per species that could not be classified into OGC by mapping to the eggNOG database. Species are grouped by phylum based on GTDB taxonomy.

discrimination than by the particular tools and identity thresholds. This is especially true for the relatively permissive thresholds implemented in the study.

One major limitation of constructing OGCs through reference-based approaches is their dependence on a limited diversity of the reference database. In most species, 5-10% of the ORF could not be mapped to the eggNOG database and were not assigned to any reference-based OGC (see Figure 4.2). The proportion of missing ORFs is greater in certain taxa that are either underrepresented or absent from the reference database. The most extreme case, with >30% unmapped genes, corresponds to B. burgdorferi, the causal agent of Lyme's disease. The poor performance of reference database mapping in B. burgdorferi is explained by the unique structure of its genome, which consists of a linear chromosome and >20 linear and circular plasmids without homologs in other species (Casjens et al., 2012; Fraser et al., 1997). Despite these limitations, reference-based OGC provides reasonably good estimates for the number of core genes per genome and allows for the retrieval of 85-90% of the single-copy core gene families identified by *de novo* approaches.

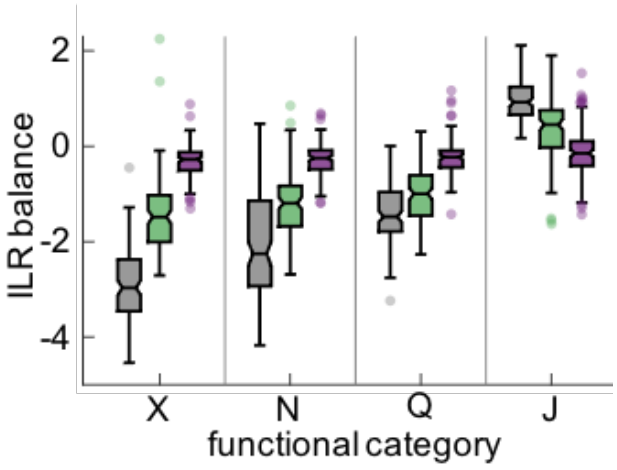

Figure 4.3: Functional differences among single-copy core OGC supported by different criteria. Each set of box plots represents the balance (measured as the isometric log-ratio) between the relative frequencies of a given functional category (x-axis) and the remaining categories not considered in previous sets (e.g., the second set of box plots corresponds to the balance between functional category N and all the rest except X). The figure shows the 4 ILR balances with the greatest variation across methods. Abbreviations of functional categories, X: mobilome; N: cell motility; Q: 2º metabolites biosynthesis, transport and catabolism; J: translation, ribosomal structure and biogenesis. Each data point corresponds to one species; boxes span the 25-75 percentiles; the central line indicates the median; whiskers extend to the most extreme data points that are not outliers; isolated points denote outliers; notches (only in f) show the 95% confidence interval of the median.

4.1.2 Systematic and species-specific biases in functional profiles

Among the pangenome features considered in this work, special attention has been paid on single-copy core genes by a closer inspection of core gene families. Although single-copy core genes have an impact on phylogeny resolution, there are also significant differences in the functional profiles of single-copy core genes that are supported by synteny and orthologs based methods (Figure 4.3). The overrepresentation of genes associated with mobile genetic elements, cell motility, and secondary metabolism, and underrepresentation of genes involved in translation among method-exclusive single-copy core genes were observed (linear mixed effects model for isometric log-ratios; $F(2,248) > 190$, $q <$ 10-20 in all cases).

Estimates of genome content diversity can be significantly influenced by the method used to distinguish paralogs when constructing *de novo* OGC, whether it be orthology- or synteny-based. To better comprehend the reasons and consequences of these variations, we categorized ORF and OGC into 21 broad functional categories that represent the primary molecular and cellular processes in prokaryotic cells. For each functional category, we calculated

the agreement between orthology- and synteny-based OGC by determining the NVI, the fraction of fully equivalent OGC. We also calculated the fraction of ORF assigned to fully equivalent OGC (see Figure 4.4a and Figure 4.5). The most significant inconsistency is found in mobile genetic elements, with only 25% of the OGC (encompassing 25% of the ORF) being equivalent. Moderate levels of inconsistency are also observed in defense systems, intracellular trafficking/secretion, and replication/recombination/repair. On the other hand, central cellular functions such as translation, transcription, nucleotide metabolism, and coenzyme metabolism exhibit the highest level of agreement, with 64-70% of the OGC (encompassing 77-84% of the ORF) being fully equivalent. These trends are also evident when examining the absolute and relative numbers of OGC per category (Figure 4.4b), with synteny-based paralog discrimination producing a disproportionate excess of OGC associated with the mobilome. The proportion of OGCs containing ORFs from multiple functional categories is consistently higher in orthology-based OGCs, although the absolute differences are modest (around 0.5-1% in most categories; see Figure 4.5). OGCs with diverse functions are frequently linked to signal transduction, cell cycle control/cell division, mobile genetic elements, and functions that are unknown or poorly characterized.

The analysis of functional profiles, which takes into account the variability between species, confirms that discriminating paralogs based on synteny leads to a significant increase in the fraction of OGC associated with mobile genetic elements (Figure 4.4c). The ILR-balance difference was 0.49, and the linear mixed effects model for isometric log-ratios showed that this difference was statistically significant ($F(2,124) = 189$, $q < 10\text{-}20$). Functional profiles derived from synteny and orthology-based OGC differ in the balance between central cellular functions, such as transcription, translation, cell cycle, nucleotide, and coenzyme metabolism, and other functional categories (ILR-balance difference = -0.11, $F(2,124) = 132$, $q < 10\text{-}20$); and between a set of functions including secondary metabolism, carbohydrate metabolism, secretion, defense and recombination, and the remaining functional categories (ILR-balance difference = 0.08, $F(2,124) = 62$, $q < 10\text{-}10$). Apart from these general trends, other significant differences between synteny- and orthology-based functional profiles are restricted to specific categories in one or a few particular species (Figure 4.4d), such as defense in *Legionella pneumophila* ($Z = 5.9$, $q < 10\text{-}5$), *Borreliella burgdorferi* ($Z = 5.4$, $q = 5\times10\text{-}5$) and *Bordetella pertussis* ($Z = 4.4$, $q = 0.002$), secondary metabolism in *Bacillus anthracis* ($Z = 4.0$, $q = 0.008$), and signal transduction in *Brachyspira hyodysenteriae* ($Z = 3.8$, $q = 0.019$).

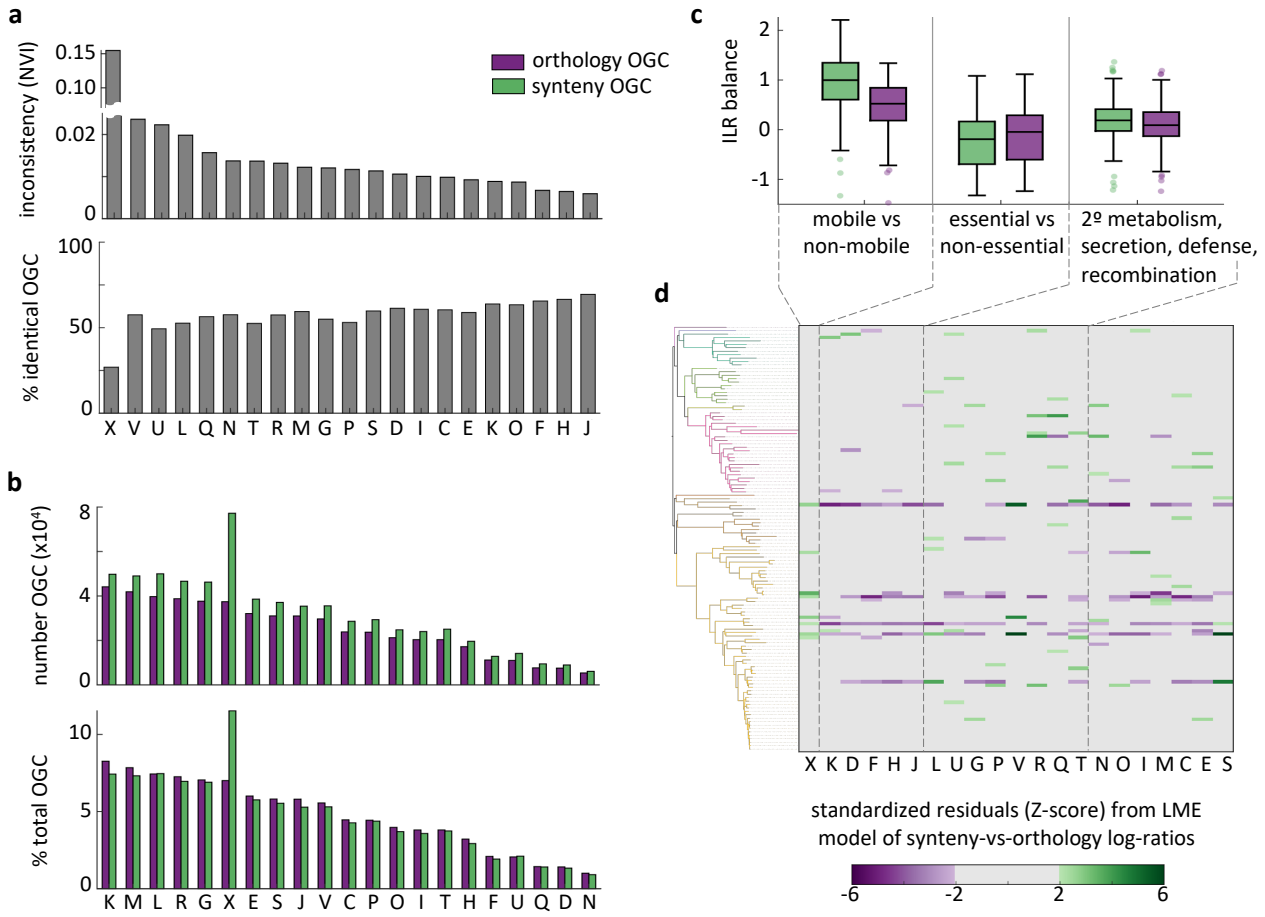

Figure 4.4: Systematic and specific biases in functional profiles associated with paralog splitting criteria. (a) Inconsistency of ORF assignations into OGC (normalized variation of information, top), and fraction of OGC that exactly contain the same ORFs (bottom) under synteny and orthology splitting criteria, stratified by functional category. (b) Absolute number (top) and relative fraction (bottom) of synteny- and orthology-based OGC associated with each functional category. (c) Balances (quantified as isometric log-ratios) for the functional categories that show the greatest systematic variation between paralog splitting criteria. Each set of boxplots represents the balance between the relative abundances of a group of functional categories (shown below) and all the remaining categories not considered in previous sets. Each data point corresponds to the pangenome of one species; boxes span the 25-75 percentiles; the central line indicates the median; whiskers extend to the most extreme data points that are not outliers; isolated points denote outliers. (d) Standardized residuals (Z-scores) of the linear mixed effects model used to infer the systematic differences shown in (c). Each row corresponds to the pangenome of one species, sorted according to the GDTB species tree (Parks et al., 2022) (phyla colored as in Figure 4.6). Colored cells indicate a significant excess of synteny- (green) or orthology-based (purple) OGC from a given category in a specific pangenome that is not explained by the general trends in (c). Abbreviations of functional categories, see Section 4.3.

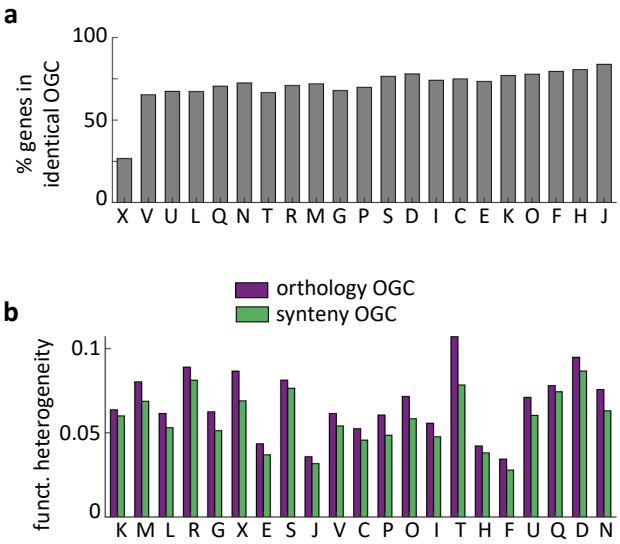

Figure 4.5: Gene-level agreement and functional heterogeneity between synteny- and orthology-based OGC. (a) Gene-level agreement between synteny- and orthology-based OGC, measured as the fraction of ORFs assigned to identical OGC and stratified by functional category. Identical OGC are those that contain exactly the same ORFs in both methods. (b) Functional heterogeneity in orthology- and synteny-based OGC. An OGC is considered functionally heterogeneous if it contains genes from >1 functional category (based on COG2020 functional annotation scheme). Abbreviations of functional categories, see Section 4.3.

4.1.3 Variability of gene flux estimates

To evaluate whether method-dependent variation in pangenome composition affects downstream analyses, we examined the impact of distinguishing between synteny- and orthology-based paralogs discrimination on a quantitative study of genome dynamics. To that purpose, we utilized the software Gloome, which infers gene gain and loss events along a lineage using a strain-level phylogenetic tree and a binary matrix of each OGC's the presence and absence profiles of each OGC. (Figure 4.6a) demonstrates that utilizing synteny-based OGC instead of orthology-based OGC leads to a 60% increase in the estimated number of gene gains and losses per lineage. Additionally, the genome versus vs gene change ratios, measured as the number of expected gains and losses per gene per core nucleotide substitution are also higher, with 40% increase when using synteny-based OGC. In contrast, the ratio between gene gains and losses, which determines the short-term dynamics of genome size (Sela et al., 2016), displays a more complex response. Synteny-based OGC produces lower or higher estimates than those obtained with orthology-based OGC depending on whether a species is dominated by gains or losses. Method-dependent uncertainties account for 15%, 18%, and 30% of the between-species variability in the total flux, the genome versus gene change ratio, and the gain versus loss ratio, respectively. These results suggest that comparative analyses of short-term

genome dynamics are highly sensitive to methodological choices for paralog discrimination, particularly when evaluating the balance between gene gain and loss.

A more thorough examination of gene flux by functional categories indicates that inconsistencies between gene clustering criteria (quantified as one minus the squared rank correlation of species-wise estimates to account for the numerous outliers) are more pronounced in mobile genetic elements and genes related to secondary metabolism and inorganic ion transportation (see Figure 4.6b). Although high inconsistencies are also observed in central functional categories, such as translation, it is important to note that the practical relevance of those is lesser due to the relatively low fluxes associated with those categories. When considering all species together by calculating their median, flux estimates obtained from synteny OGC display a systematic deviation of more than one additional event per gene in all functional categories, which is evenly distributed between gains and losses (see Figure 4.6c).

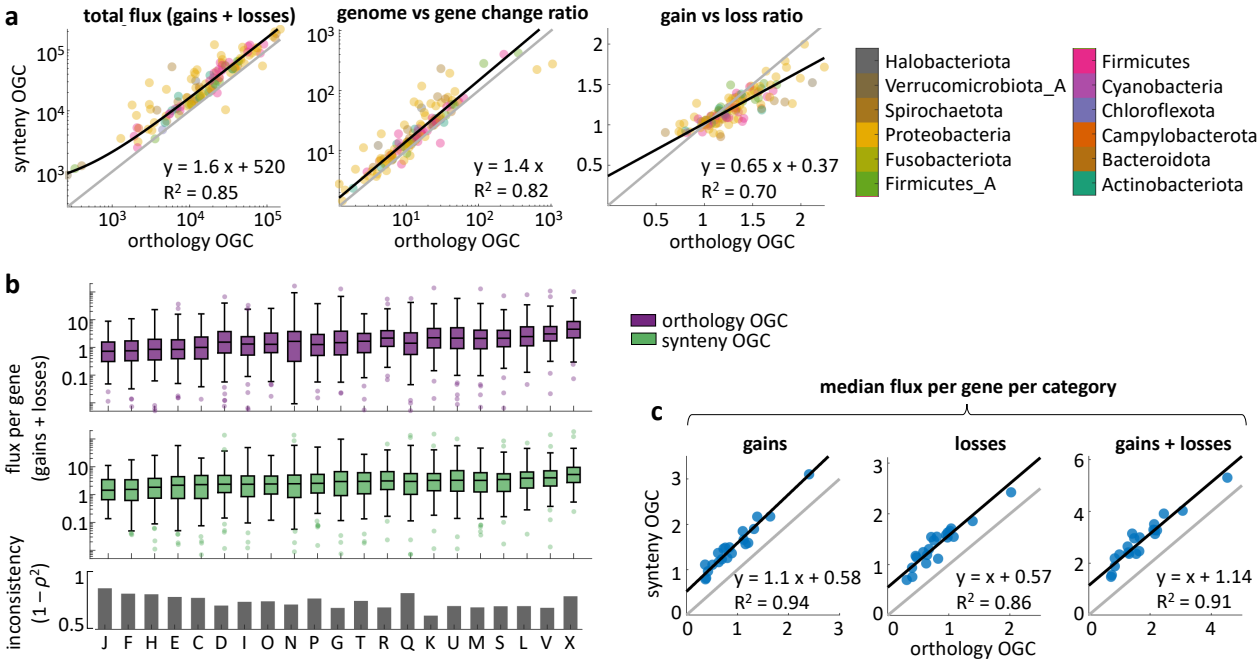

Figure 4.6: Effect of paralog splitting criteria on the inference of gene flux. (a) Species-wise comparison of the total gene flux (gains and losses along the species tree, left), genomic content vs gene sequence evolution ratio (middle), and total gain vs loss ratio (right) inferred from synteny- and orthology-based OGC. The method-dependent uncertainty is equal to 1 – R2. (b) Flux per gene per functional category inferred from orthology- (top) and synteny-based (middle) OGC. Each data point corresponds to one species; boxes span the 25-75 percentiles; the central line indicates the median; whiskers extend to the most extreme data points that are not outliers; isolated points denote outliers. The bar plot at the bottom shows the inconsistency between methods, quantified as one minus the squared rank correlation. Abbreviations of functional categories as in Section 4.3. (c) Median flux per gene per category, calculated over all the species, for orthology-based (x-axis) and synteny-based (y-axis) OGC. Similar trends are observed for gains, losses, and the combination of both.

4.2 Quantifying the effect of CRISPR-Cas immunity on gene gain and loss

4.2.1 CRISPR presence and absence affects genome content in a gene- and species-specific way

Our previous pan-genome study revealed that, when tracking vertically transmitted genes, the best methods for gene clustering are those based on synteny criteria (e.g., Roary) because they minimize the potential contamination by horizontally transferred genes (Manzano-Morales et al., 2023). Hence, we chose to apply the insights obtained from our previous study to investigate the impact of CRISPR-Cas systems on within-species genome fluidity at different time scales. The presence of CRISPR-Cas systems has been described to positively or negatively impact horizontal gene transfer (HGT) in some species (Wheatley and MacLean, 2021; Meaden et al., 2022; Shehreen et al., 2019) while not having a perceptible impact on this process in others (Gophna et al., 2015). To systematically address this issue, a total of 19,323 high-quality genomes belonging to different bacterial species from the GTDB database (release 202), were used to assess in which cases the presence of CRISPR-Cas systems correlates/anti-correlates with HGT. In particular, a total of 196 bacterial species and 1 archaeal species were considered. The distribution of species within each phylum is as follows: 76 in Proteobacteria, 68 in Firmicutes, 17 in Firmicutes_A, 16 in Actinobacteriota, 15 in Bacteroidota, 2 in Fusobacteriota, 1 in Campylobacterota, and 1 in Chloroflexota and 1 in Methanobacteriota (archaea) according to GTDB taxonomy.

The effects of CRISPR-Cas on gene abundance were analyzed by species and gene functional categories. To that purpose, we used the NCBI COGs database (release 2020) to categorize genes into broad functional classes (see below), and conducted a comprehensive analysis to examine the association between the presence of CRISPR-Cas systems in genomes and the gene counts within each category. This enabled a detailed characterization of how CRISPR-Cas systems impact these particular functional groups.

As a first step, we compared the gene abundances between two groups of species based on the presence or absence of the CRISPR-Cas system in the genome: those containing CRISPR-Cas are denoted as CRISPR(+) and those lacking it CRISPR(-). This comparative analysis intends to elucidate the genetic disparities linked to the presence or absence of the CRISPR-Cas system. To avoid biases, CRISPR(+) genomes were required to carry at least one *cas* gene along with the CRISPR array (see Methods) and genes located in the Cas operon were not counted as part of any functional category. It is important to note that

Table 4.3: NCBI COG Categories.

Category	Description
J	Translation, ribosomal structure and biogenesis
K	Transcription
L	Replication, recombination and repair
D	Cell cycle control, cell division, chromosome partitioning
Y	Nuclear structure
V	Defense mechanisms
M	Cell wall/membrane/envelope biogenesis
N	Cell motility
U	Intracellular trafficking, secretion, and vesicular transport
O	Posttranslational modification, protein turnover, chaperones
X	Mobilome: prophages, transposons
C	Energy production and conversion
G	Carbohydrate transport and metabolism
E	Amino acid transport and metabolism
F	Nucleotide transport and metabolism
H	Coenzyme transport and metabolism
I	Lipid transport and metabolism
P	Inorganic ion transport and metabolism
Q	Secondary metabolites biosynthesis, transport and catabolism
R	General function prediction only
S	Function unknown

phylogenetically proximate species often exhibit similar traits (Ives and Garland Jr, 2010; Ives and Helmus, 2011). Therefore, genomes with close phylogenetic relationships are likely to display similar patterns regarding the presence or absence of the CRISPR-Cas system and genome composition. To address this phylogenetic bias, a phylogenetic generalized linear mized-effects model (PGLMM) was fitted for each species and functional category. The presence or absence of CRISPR-Cas systems was defined as the predictor variable, and the number of genes in that particular gene category was defined as the response variable. We classified a species as Positively (respectively Negatively) Correlated Species (PCS, respectively NCS) if it displayed a significant correlation ($p < 0.05$) or an absolute effect size greater than the effect size of the first significantly correlated species.

Functional categories X, V, K, L, U, G, and Q are the ones with the highest numbers of PCS or NCS, indicating that the presence of CRISPR-Cas systems is associated with differences in gene content for those particular functional categories. The X functional category, containing

genes from mobile elements, shows the highest number of species displaying a significant correlation: 23 PCS, 15 NCS ($p < 0.05$) (Figure 4.7), and more than 100 species exhibit higher effect sizes (correlation) but not significant p-value.

When analyzing the fraction (rather than the absolute number) of genes from each functional category, we observed similar results and effect sizes, especially for those categories whose abundances are most strongly associated with CRISPR-Cas(Figures 4.8 and A.1). Therefore, the choice of absolute or relative units to quantify gene abundances has a minimal influence on the subsequent analysis of the mobilome.

We next summed up the genes from all categories and used them as the response variable. The results of the PGLMM indicate a slightly larger number of species in which the presence of CRISPR positively correlates with genome size. Specifically, 15 NCS and 27 PCS were found to be significant ($p < 0.05$) (Figure 4.9). These results led us to ponder whether this the excess of PCS could affect the direction of the overall correlation in bacteria. In other words, when we analyze PCS and NCS collectively from a statistical standpoint, is it possible that the correlation will be canceled out? Or does it produce an overall significant positive correlation? To explore this question, we implemented the PGLMM by considering all species as a whole. The results show minimal positive effect size (0.0047) but statistically significant ($p = 4 \times 10^{-7}$), suggesting that when all species are considered as a whole, the effects become less notorious.

To demonstrate species-specific correlations, we displayed the results of the PGLMM analysis of genome size on the phylogenetic tree of the considered species. We observed a significant association between the presence of CRISPR and certain species across different phyla (Figure 4.10). For example,there is a negative association between CRISPR in *Vibrio cholerae* and the abundance of V genes (other defense systems), as well as a positive association with the abundance of X genes (mobilome). The positive association can be explained by the previously discovered fact that the Type I-F CRISPR-Cas system, encoded by phage ICP1, counters Phage-inducible chromosomal islands-like elements (O'Hara et al., 2017). The co-occurrence of CRISPR presence with ICP1 phage genes in this context supports the idea of its positive association with the mobilome. In *Pseudomonas aeruginosa*, we observed a negative association between the presence of CRISPR and defense systems and transcription genes. This finding is consistent with previous research where CRISPR was found to be associated with a higher abundance of MGEs (Wheatley and MacLean, 2021). Furthermore, our findings support the idea that the CRISPR-Cas system can both negatively and positively correlate with gene abundance in different species, which aligns with previous findings (Shehreen et al., 2019). Additionally, most of the species that we have identified

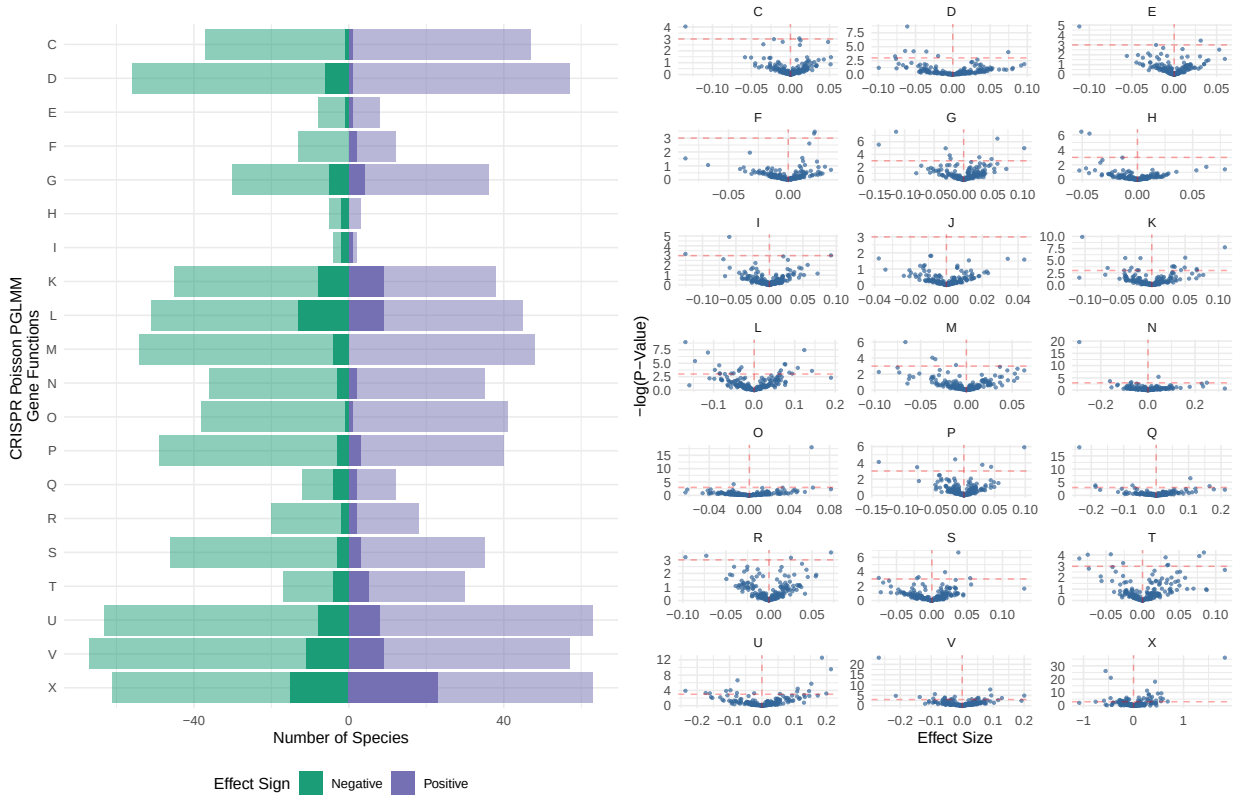

Figure 4.7: Correlation between presence of CRISPR-Cas and the number of genes from different functional categories (Poisson-distributed PGLMM). The bar plot displays the number of species that showed a significant correlation between the presence of the CRISPR-Cas and the number of functional-specific genes. Solid bars correspond to a $p < 0.05$ significance cutoff. Semitransparent bars correspond to an effect-size cutoff defined by the smallest absolute effect size that has a significant p-value. The scatter plot displays the distribution of p-values at different effect sizes (one point per species).

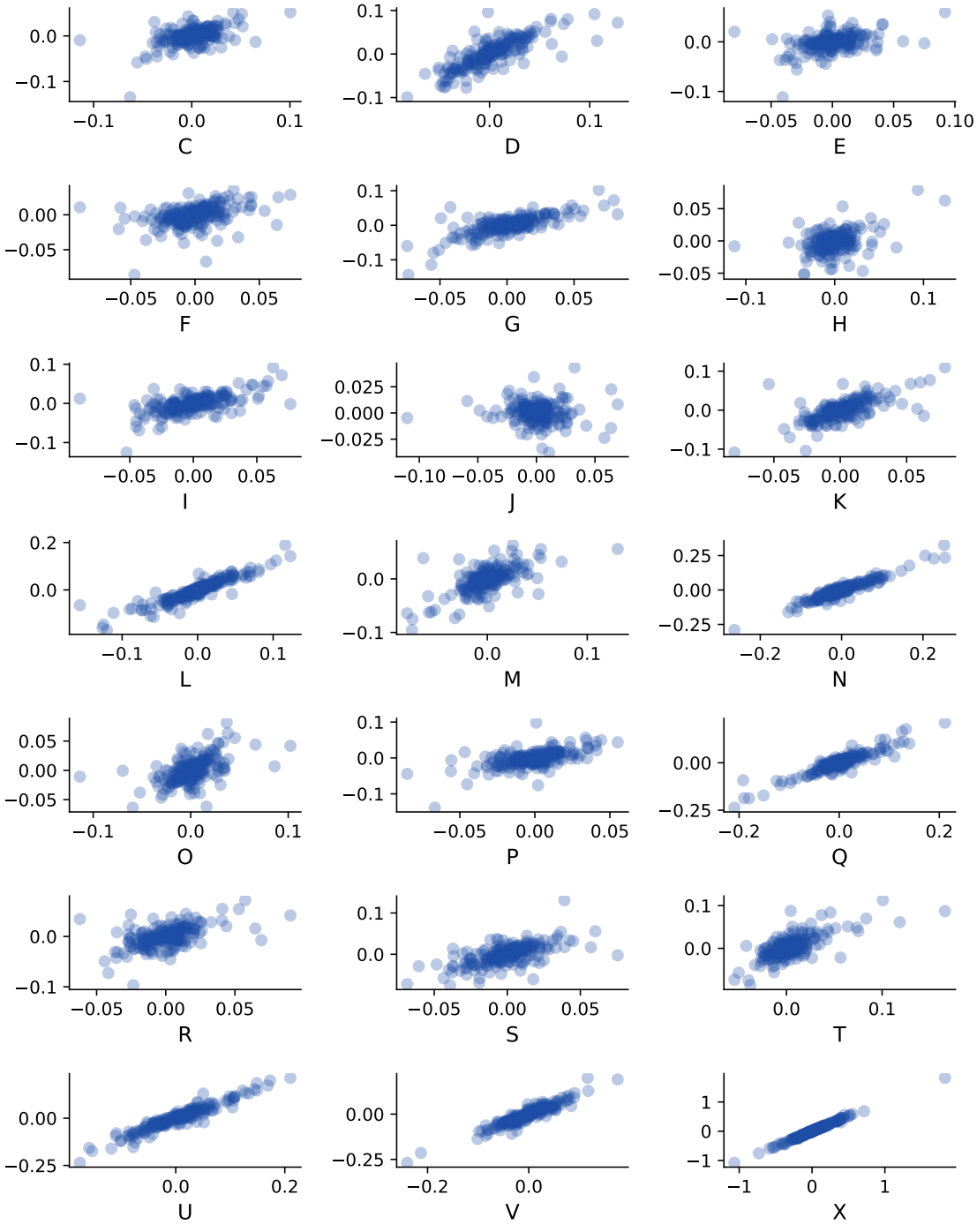

Figure 4.8: Comparison of the effect size of binomial and Poisson distributed PGLMM organized by functional categories. The x-axis represents effect size of the Binomial PGLMM; y-axis represents effect size of the Poisson PGLMM. Each subplot is a functional category and each dot is a species.

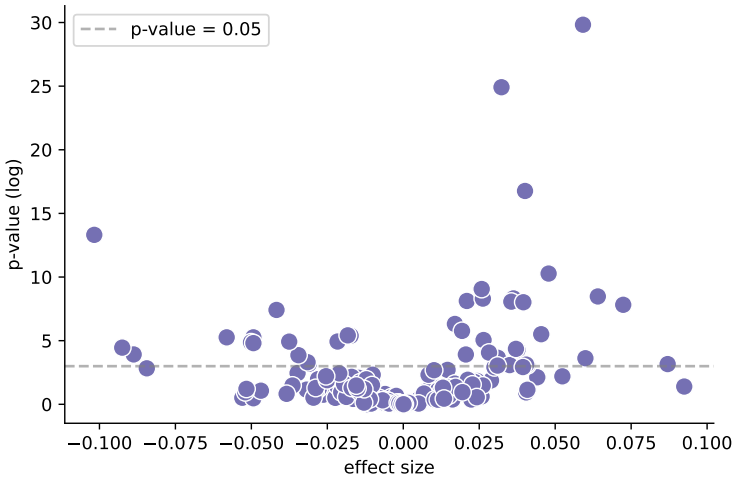

Figure 4.9: Correlation between the presence of CRISPR and the total number of genes on the genome. The dash line indicates a p-value of 0.05. Each point represents a species.

had not previously been reported as showing positive or negative correlations between gene abundance and CRISPR.

4.2.2 Comparison of CRISPR effect on different mobile genetic elements

In previous PGLMM analyses, we counted the total number of genes annotated in the X category. However, in this PGLMM analysis, we focused specifically on the number of genes annotated as prophages, transposons, and plasmids genes. This allowed us to gain insights into each type of MGE. We fit a Poisson-distributed PGLMM to prophages, transposons, and plasmids and found 9 NCS and 18 PCS for transposons, 14 NCS and 14 PCS for prophages PCS, and 5 NCS and 11 PCS for plasmids ($p < 0.05$) (Figure 4.11). Thus, the number of PCS was greater than NCS on transposons and plasmids, whereas the number of NCS was greater than PCS on prophages.

We used the effect size of Poisson PGLMM to compare the effect of CRISPR on the number of genes from different classes of MGE. We found that the effects on genes from prophages and transposons are highly correlated with those previously reported for the X (mobilome) category (0.69 Spearman coefficient). In contrast, there is a much weaker correlation between X-category genes and plasmids, and between transposons and prophages, indicating that there is a decoupling in the effect of CRISPR on these MGE (Figure 4.12).

Taken together, these results underscore the importance of CRISPR in preventing the invasion by MGE in several species, as evidenced by the anticorrelation between CRISPR presence

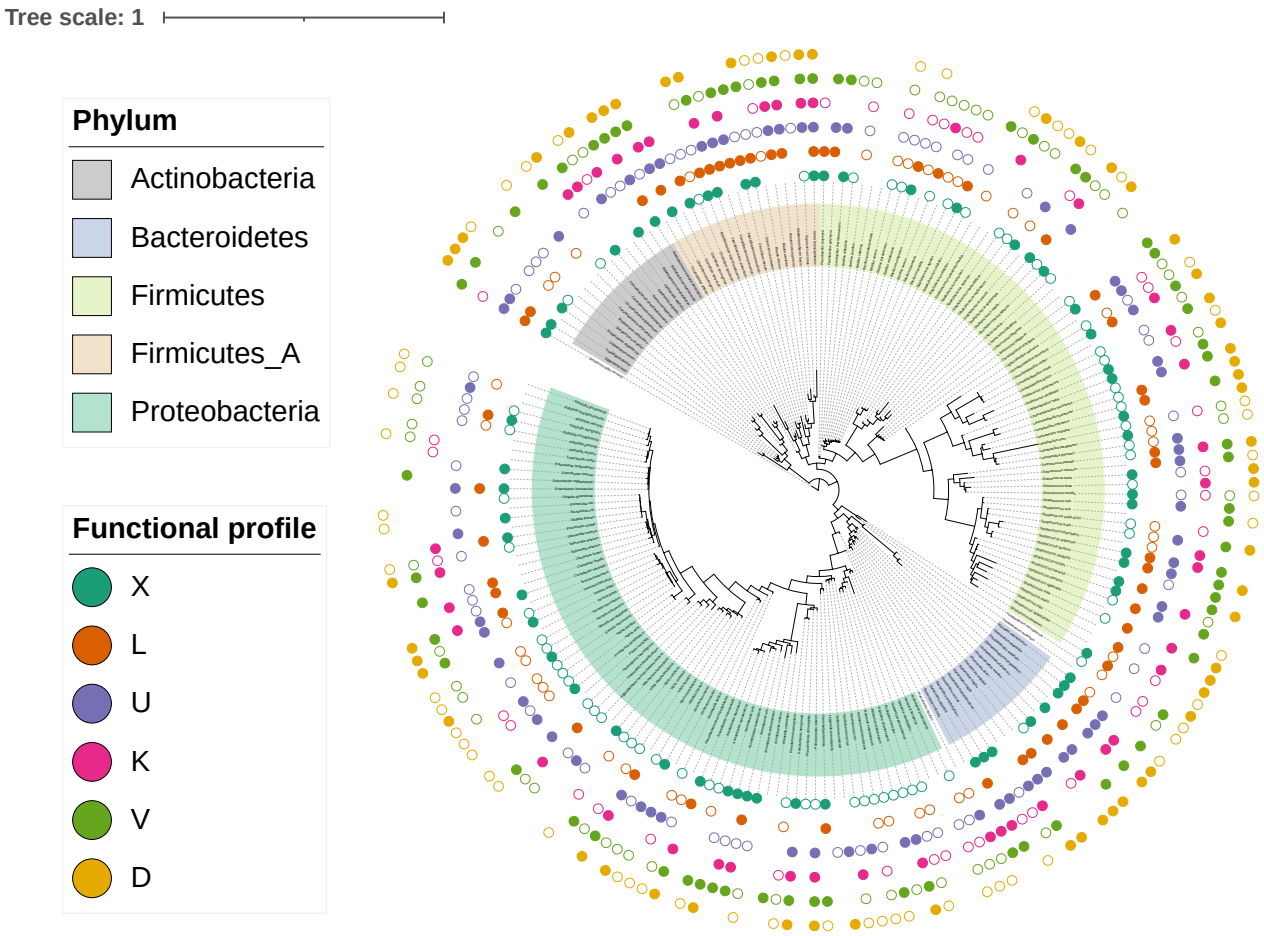

Figure 4.10: Distribution of the correlation along the GTDB species tree. The solid cycle indicates negative correlation between presence of CRISPR and gene abundance of the functional profiles; The hollow cycle indicates positive correlation between presence of CRISPR and gene abundance of the functional profiles. The correlation were generated by Poisson PGLMM with the absolute effect size greater or small than the effect size of the first significantly correlated species. Abbreviations of functional categories, see Section 4.3. Firmicutes_A includes Clostridiales and Lachnospirales.

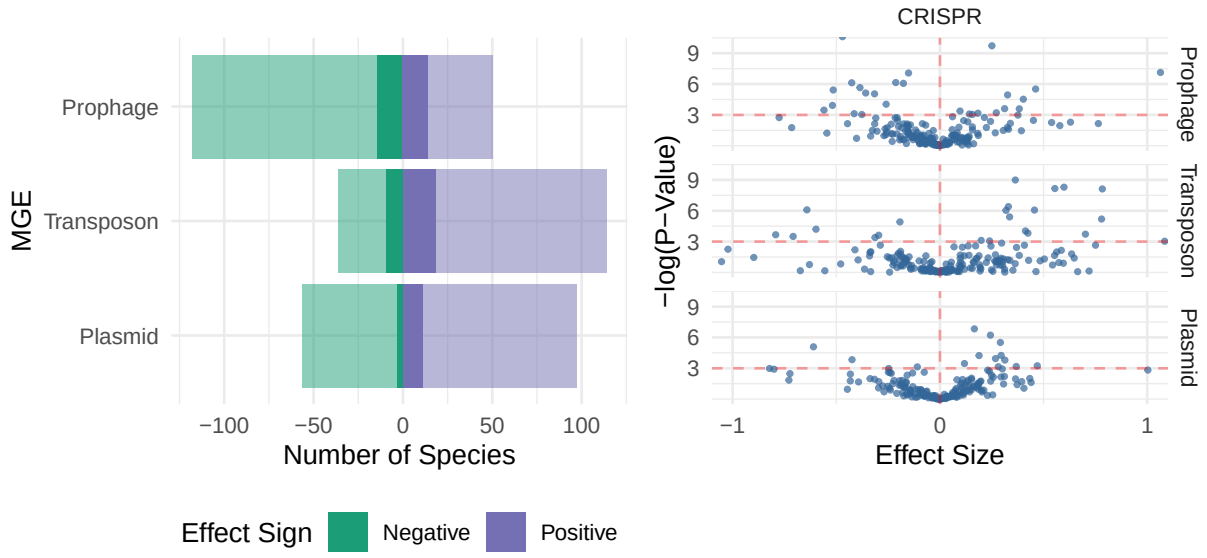

Figure 4.11: Correlation between presence of CRISPR-Cas and the number of genes from different MGE (Poisson-distributed PGLMM). The bar plot displays the number of species that showed a significant correlation between the presence of the CRISPR-Cas and the number of MGE-specific genes. Solid bars correspond to a $p < 0.05$ significance cutoff. Semitransparent bars correspond to an effect-size cutoff defined by the smallest absolute effect size that has a significant p-value. The scatter plot displays the distribution of p-values at different effect sizes (one point per species).

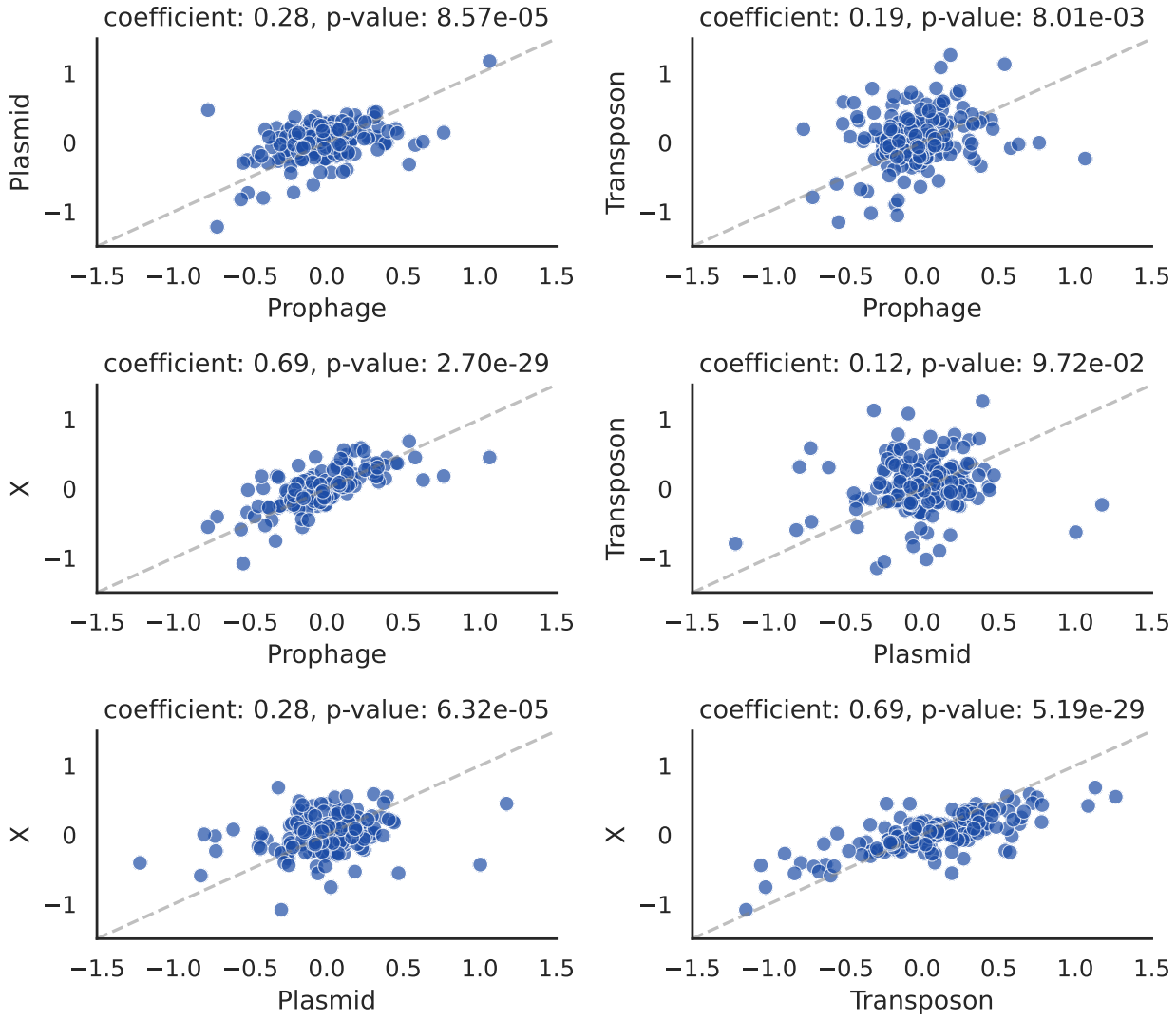

Figure 4.12: Cross-comparison of effect sizes across different types of MGE (PGLMM Poisson). The dash line represents the $x = y$. Statistical analysis was performed using Spearman's correlation coefficient with associated p-values.

and MGE abundance, although this is by no means a general rule. In fact, the higher number of PCS suggests a more complex relation between CRISPR presence and MGE abundance in a significant number of species.

4.2.3 The effect of CRISPR-Cas on genes from K, L and U categories is mediated via its actions over the mobilome

One potential reason explaining the effect of CRISPR-Cas systems on genes classified in non-mobilome (non-X) COG categories is that these genes can be carried by MGEs. To test this hypothesis, we replicated the PGLMM from the preceding section. A Binomial analysis, considering the fraction of gene numbers per category, was conducted on the set of 2,964 complete genomes. This analysis was performed both before and after the removal of identified MGEs, specifically transposons, prophages, plasmids, ICEs, and IMEs. Note that the analyses were performed only using complete genomes due to the difficulty in identifying plasmids, ICE, and IME on incomplete genomes.

After removing the MGEs from the analysis, we were able to recover only 4 PCS (out of 18) and 5 NCS (out of 16) related to the X functional category (Figure 4.13). Moreover, two species changed the sign of their correlation after masking. A visual inspection of the genes from the X category that showed significant correlation after masking revealed that most of these are likely parts of prophage components, although they were not predicted as whole prophages by MGE identification tools (Figure A.2). Ideally, the X category should show zero NCS and PCS after removing the MGEs. However, there are difficulties in recognizing new types of MGEs, such as Phage-Inducible Chromosomal Islands which which also contain phage genes (Penadés and Christie, 2015).

As hypothesized, the removal of MGE units generally led to a reduction in the number of species showing a correlation between the presence of CRISPR-Cas systems and the gene count in non-mobilome related functional categories (Figure 4.13). This effect was particularly noticeable in genes annotated as L (Replication/recombination/repair), U (Intracellular trafficking/secretion/vesicular transport), and K (Transcription) types, and the number of PCS within these categories decreased more significantly than the NCS. Regarding defense systems (V), three species previously categorized as PCS became statistically insignificant, suggesting that these defense-related genes might be located within the MGEs. In conclusion, these results clearly indicate that the effect of CRISPR-Cas systems on L, U, K genes (in both NCS and PCS) and V genes (in PCS) is mediated by its action over the MGEs that contain them.

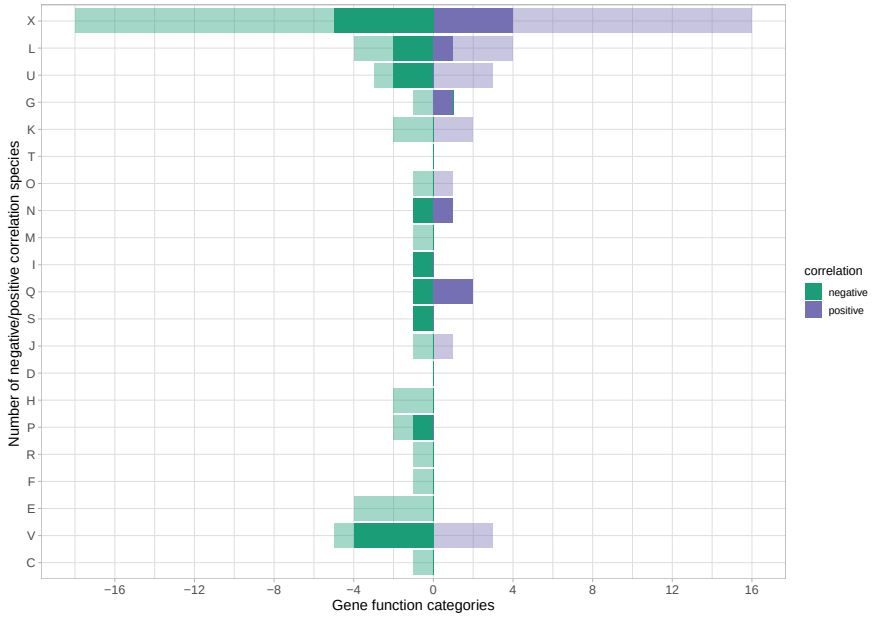

Figure 4.13: Correlation between presence of CRISPR-Cas and gene relative abundances (binomial-distributed PGLMM) with and without masking MGEs in complete genomes. Semi-transparent bars indicate the number of species showing significant correlations ($p < 0.05$) before masking MGEs. Solid bars indicate the number of species showing significant correlations ($p < 0.05$) after masking MGEs.

4.2.4 Estimating CRISPR gain and loss rates through the analysis of synteny-based orthologous gene clusters

In the previous sections, we explored the impact of CRISPR on gene abundance. Now, we will focus on the influence of CRISPR presence or absence on gene gain and loss. This evolutionary process is characterized by extensive gene turnover, which profoundly affects gene abundance (Iranzo et al., 2019). Mechanisms such as HGT and the involvement of MGEs facilitate this process and influence their evolutionary pathways (Iranzo et al., 2019). Our aim was to clarify the influence of CRISPR on the historical patterns of gene gain and loss, and enhance our understanding of how CRISPR affects the evolutionary forces that shape gene abundance over time.

Using single-species trees as references, we compared the gene gain and loss rates between sister clades, one harboring CRISPR–Cas systems (clades composed of >80% CRISPR(+) genomes) and the other lacking it (clades composed of >80% CRISPR(-) genomes). We calculated the rates of gene gain and loss with the software Gloome. When jointly analyzing all species, we obtained 549 pairs of CRISPR and sister (the counterpart of CRISPR presence) branches across 179 species. The distribution of clades and species is shown in Figure 4.14:

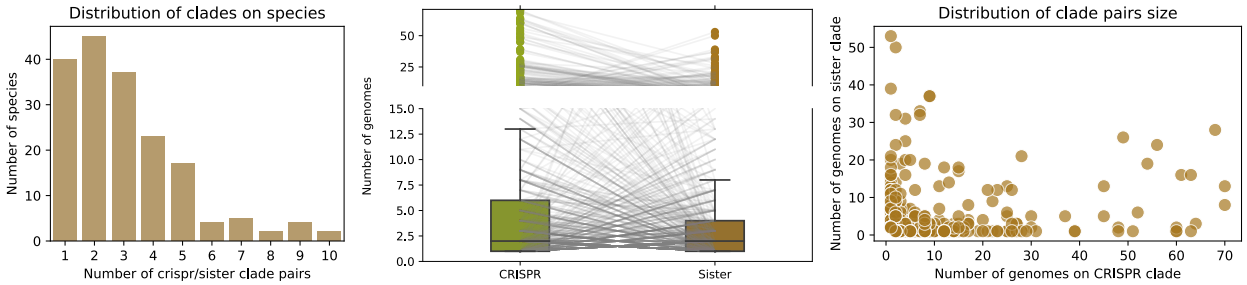

Figure 4.14: Statistics on gene gain and loss analyses. The bar plot displays the distribution of the CRISPR and sister clades among species. The box plot compares the sizes of CRISPR and sister clade pairs. Scatter plot showing the number of genomes between CRISPR and sister clade pairs.

- Most species contain 1-3 pairs of clades.
- The median number of genomes in a clade is 2 for both CRISPR and sister clades.
- If the number of genomes in a CRISPR clade is high, then its sister clade tends to have a low number, and vice versa.
- There are no instances where both a CRISPR clade and its sister clade contain more than 30 genomes each.

The overall distributions of the differences in gain, loss and gain plus loss rates between CRSPR(+) and CRISPR(-) sister clades has almost zero negative median, some pairs show higher rates in CRISPR(+) than in sister clades and others show the opposite trend (Figure 4.15). This observation coincides with CRISPR-Cas systems do not significantly restrict overall HGT on evolutionary timescales (Gophna et al., 2015). However, the distribution is significantly asymmetric and heavily left-tailed (sample size: 549 clades. skewness statistic: gain -3.45, loss -3.73, gainLoss -3.64, $p < 0.01$). Furthermore, it is important to note that the depth of CRISPR clades is higher on the negative axis than on the positive axis, which is a significant observation. We concluded that at moderate and long evolutionary times, CRISPR is more likely to have a negative effect on gene gain and loss. However, in very short time periods, the intrinsic effect of CRISPR-Cas on HGT cannot be separated from the very same HGT events that led to the divergence between CRSIPR(+) and CRISPR(-) branches.

The loss rate usually reflects the selection process, while the gain rate is commonly linked to HGT (Iranzo et al., 2019; Iranzo et al., 2017). Next, we focused gain rates by species (Figure 4.16), and observed that the asymmetry depends on the species, with rates in some CRISPR clades being lower than in sister clades. Furthermore, it is important to note that CRISPR clades with high depth are specific to certain species, resulting in varying rates of

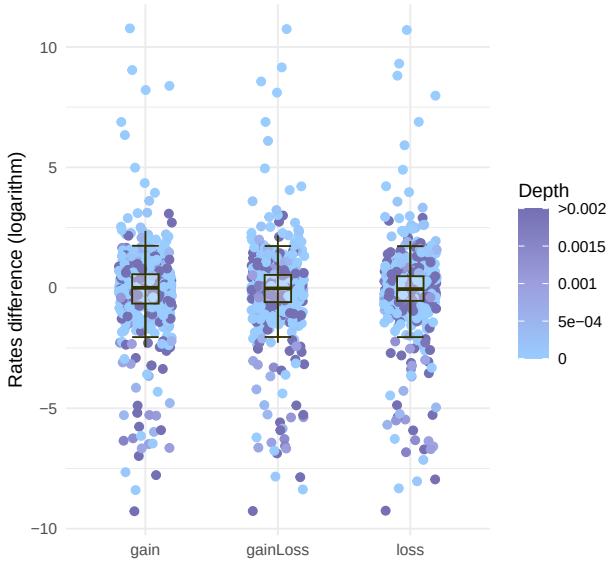

Figure 4.15: The difference of overall gene gain and loss rates between CRISPR(+) clade and sister CRISPR(-) clade. The y-axis is equal to the logarithm of the gene gain or loss rate of the CRISPR(+) clade minus the logarithm of the gain or loss rate of the sister CRISPR(-) clade. The color represents the depth of the CRISPR(+) clade in the strain tree (measured as the average number of substitutions per site in GTDB marker genes). Each point is a pair of CRISPR(+/-) clades.

low gain depending on the species.

In conclusion, the asymmetry of the distributions suggests that CRISPR leads to reduced rates of gene gain and loss on the deeper CRISPR clades, indicating a tendency to prevent HGT on an evolutionary timescale. The magnitude of the CRISPR-Cas effect on gene gain and loss varies among species. When considering all species together, a comparison of clades with and without CRISPR shows a significant but small effect size.

Focusing further, we identified that the presence or absence of CRISPR-Cas systems particularly influences the gene abundance of MGEs. To evaluate the effect of CRISPR-Cas systems on gene gain an loss rates in MGEs, we compared the specific rates in mobilome (X functional category, defined as PCS and NCS based on an effect size cutoff described previously, using Poisson PGLMM) between CRISPR(+) clades and their corresponding CRISPR(-) sister clades. Within the mobilome, it was observed that NCS and PCS associated with CRISPR(+) clades exhibited similar loss rates to those of CRISPR(-) clades, however, the gain rates were significantly different ($p < 0.05$) (Table 4.5). The result showed that the gain vs loss are significantly different, as the median of the loss rates are close to 0 while the gain rates are not (Figure 4.17).

Table 4.5: P-value of the Figure 4.17

Correlation	**GainLoss P-val**	**Type**	**CrisprSister P-val**
Negative	0.000261	Gain	0.003152
		Loss	0.140904
Positive	0.002567	Gain	0.021936
		Loss	0.948995

Upon analyzing the gene gain and loss rates within specific categories of the mobilome, we observed that, in general, there was no significant effect on gene loss rates. The presence of the CRISPR-Cas system affected gene gain rather than gene loss. The PCS showed a higher gene gain, while the NCS showed a lower gene gain when compared to the median of 0 (Figure 4.18). Significant impacts on gene gain rates were found from prophages, transposons in both NCS and PCS, and plasmids in PCS (Table 4.6). The gain to loss rates show differences in all comparisons (Figure 4.18), confirming that the effect of CRISPR on gene gain rates is very different from its effect on loss rates.

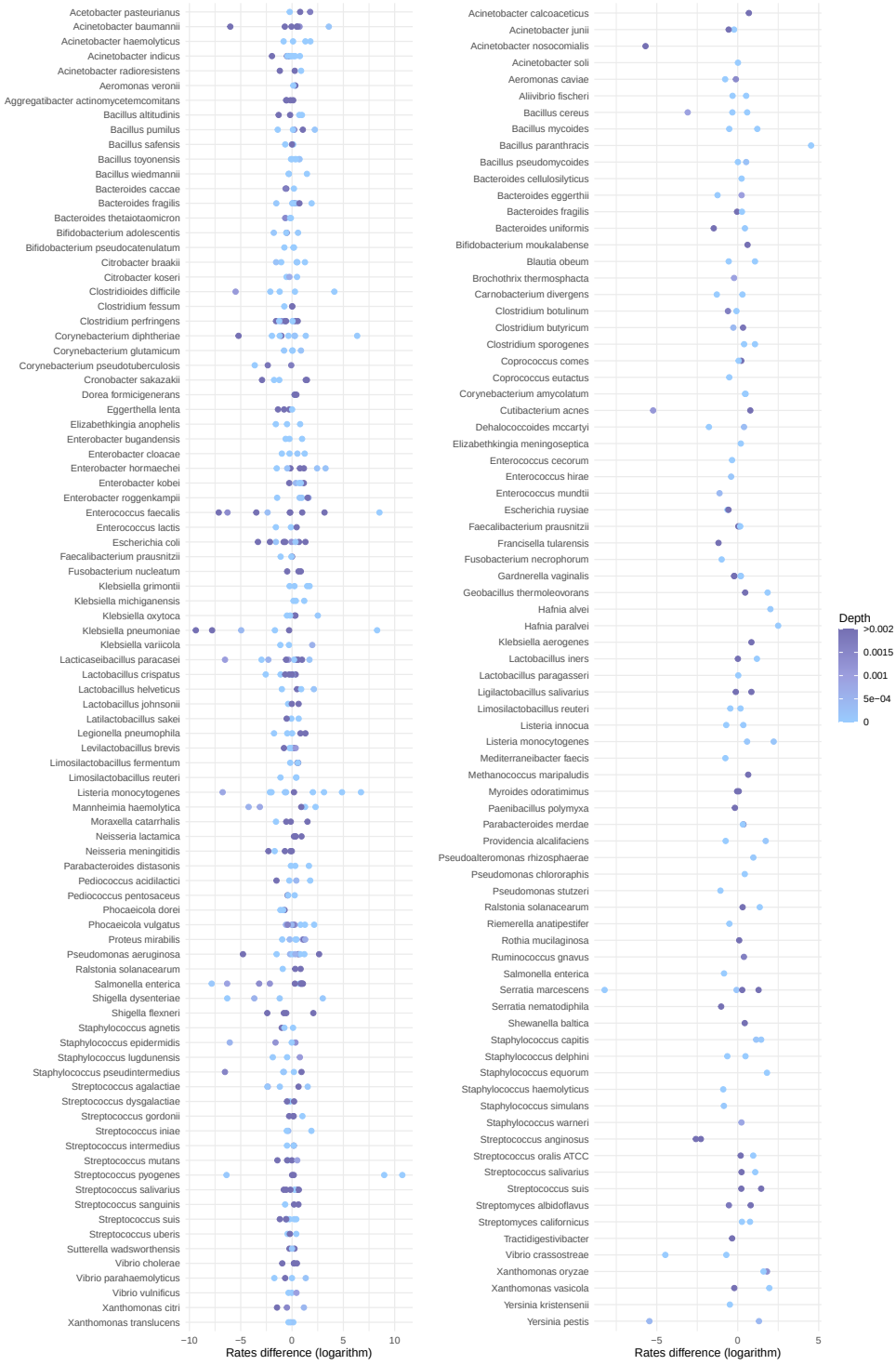

Figure 4.16: Gene gain rates between the CRISPR clade and its sister clade for each species. The x-axis is the rates of CRISPR clade minus corresponding sister clade. Each point is a pair of CRISPR/sister clade.

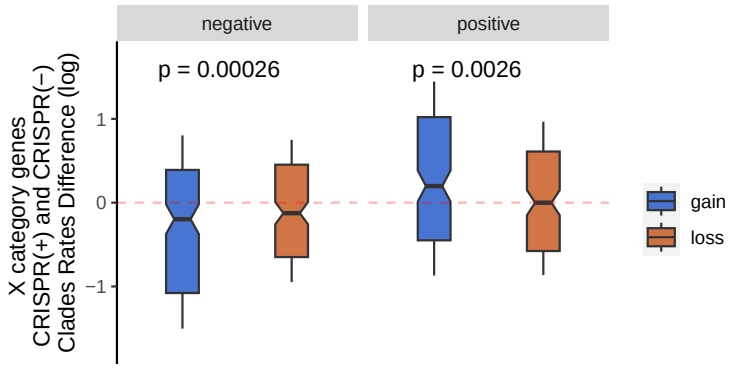

Figure 4.17: Log-difference in gene gain and loss rates between CRISPR(+) and CRISPR(-) sister clades. Gene gain and loss rates correspond to the X functional category (mobilome); "positive" and "negative" refers to the correlation between the presence of a CRISPR-Cas and the number of genes in the mobilome (PGLMM-based PCS and NCS). Notches indicate 95% confidence intervals for the median; the box covers percentiles 25 to 75. P-values correspond to the comparison between differences in gain and loss rates (Wilcoxon test for paired samples). Separate p-values testing the deviation of each median from zero are provided in Table 4.5

.

Table 4.6: P-value of the Figure 4.18

Category	**Correlation**	**GainLoss p-val**	**Type**	**CRISPR Sister p-val**
Prophage	Negative	0.000069	Gain	0.001454
			Loss	0.568690
	Positive	0.024870	Gain	0.007212
			Loss	0.697965
Plasmid	Negative	0.016295	Gain	0.062121
			Loss	0.759844
	Positive	0.000027	Gain	0.005779
			Loss	0.513733
Transposon	Negative	0.004163	Gain	0.034051
			Loss	0.469181
	Positive	0.001671	Gain	0.011706
			Loss	0.510489

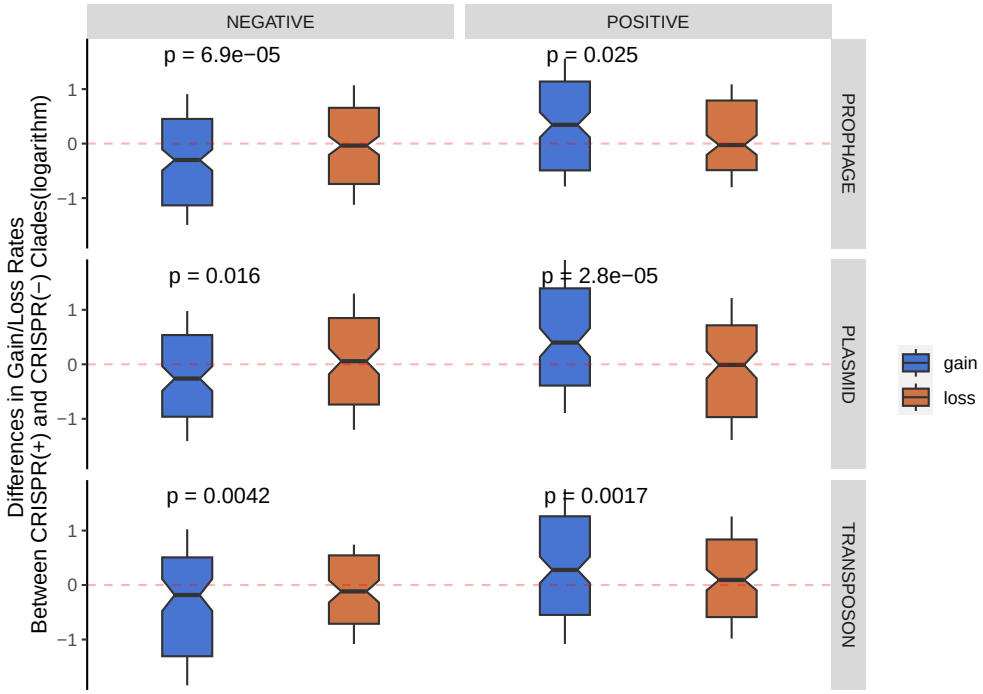

Figure 4.18: Log-difference in gene gain and loss rates between CRISPR(+) and CRISPR(-) sister clades for different MGE. Gene gain and loss rates correspond to prophages, plasmids, and transposons, as indicated on the right; "positive" and "negative" refers to the correlation between the presence of a CRISPR-Cas and the number of genes in the MGE of interest (PGLMM-based PCS and NCS). Notches indicate 95% confidence intervals for the median; the box covers percentiles 25 to 75. P-values correspond to the comparison between differences in gain and loss rates (Wilcoxon test for paired samples). Separate p-values testing the deviation of each median from zero are provided in Table 4.6

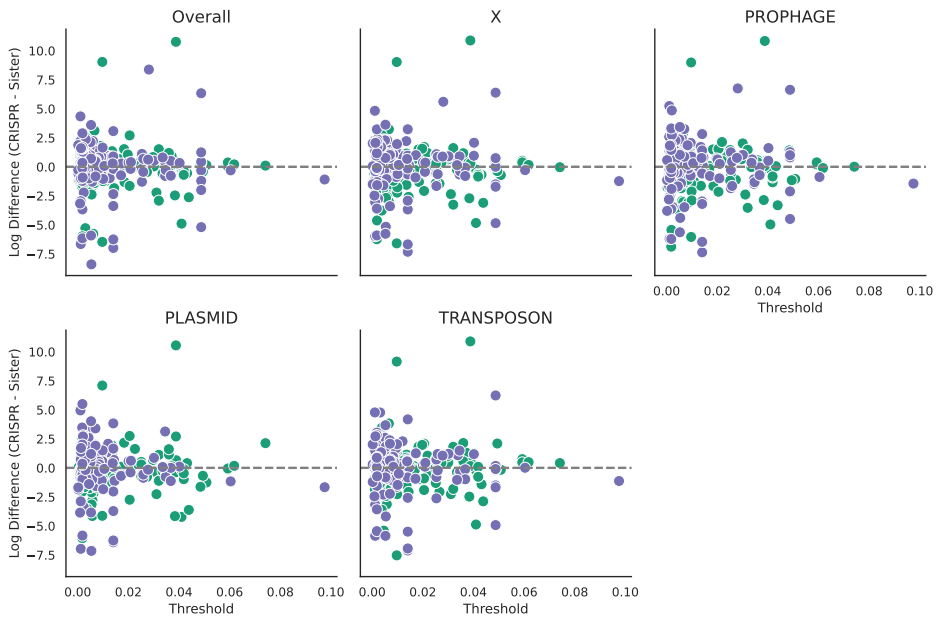

Figure 4.19: Difference in gene gain rates between CRISPR(+) and CRISPR(-) over evolutionary time. The NCS (green dots) and PCS (purple dots) cutoff defined by the smallest absolute effect size that has a significant p-value. The x-axis represents the tree depth of the CRISPR clades, y-axis is the corresponding gene gain rates of CRISPR(+) clades minus their sister CRISPR(-) clade. Each point represents a pair of the CRISPR and sister clade.

To delve deeper into the impact of CRISPR on various timescales, we segmented the species tree into subtrees by depth and analyzed the divergence in gain and loss rates between CRISPR clades and their sister clades across distinct evolutionary timescales. We found that the differences in gain and loss rates between CRISPR and sister clades tend to: (1) show large positive and negative dispersion at short evolutionary times, and (2) become more negative overall at longer evolutionary times.(Figure 4.19).

In summary, our study highlights the significant impact of CRISPR-Cas systems on the gain of MGE genes.

4.2.5 Factors that are not significantly involved in the association of CRISPR and MGE abundance

In our exploration of potential factors that could influence the variability in the effects of CRISPR-Cas across species, we employed the Mann–Whitney U test to compare GC content, genome size, and pan-genome size among NCS, PCS, and NS groups. We found no significant differences in these genomic attributes among the three groups (Figure A.3).

The type of CRISPR-Cas system is a crucial factor that determines its functional behavior (Makarova and Koonin, 2015). We conducted an analysis to examine the prevalence of different CRISPR types across species classified as NCS, PCS, and NS as detailed in Table A.1. However, the analysis did not detect any significant differences (chi-squared test $p = 0.6372$).

We used the *cas1* gene as a proxy for the presence of a functional CRISPR-Cas system. Accordingly, we tracked gains and losses of *cas1* to assess the propensity of a species to maintain, loss, and regain CRISPR-Cas systems. In total, 517 gene clusters (synteny-based OGC, generated by Roary) annotated as *cas1* were collected across 197 species. The distribution of *cas1* gain plus loss rates does not differ when comparing NCS and PCS (Mann–Whitney U test $p = 0.97209$), indicating that CRISPR plasticity does not explain the sign of the correlations.

Finally, we investigated if the positive correlation between CRISPR-Cas and gene abundance observed in a large number of species might be attributed to CRISPR systems being located within MGEs. To test this hypothesis, we collected the coordinates of CRISPR and MGEs in complete genomes and fitted a generalized linear mixed-effects model to test if the probability to classify a species as PCS (response variable) could be predicted by the fraction of CRISPR-Cas systems located inside MGE (predictor variable). Our results indicated that the fraction of CRISPR-Cas systems in MGE is a very poor predictor of the type of association between CRISPR-Cas and MGE abundance (effect size=-0.0611, $p = 0.923$). This finding suggests that a more complex interaction is at play, potentially involving other genomic factors or external selective pressures in driving the association between CRISPR-Cas and MGE abundance.

4.2.6 Anti-CRISPR proteins modulate the association between CRISPR-Cas and MGE abundance

Given the negative results from previous section, we decided to investigate the possible role of anti-CRISPR proteins. Anti-CRISPR proteins can suppress the effect of CRISPR-Cas systems and, therefore, they may explain the lack of major inhibitory effects of CRISPR-Cas on gene transfer (Davidson et al., 2020; (Camara-Wilpert et al., 2023; Mahendra et al., 2020). To investigate that possibility, we searched 19,323 genomes for anti-defense proteins. We found a total of 151,680 anti-defense proteins, of which 43,345 could be assigned to a specific defense system. Of those 24,677 are related to Restriction-Modification (RM) systems, 16,058 to CRISPR-Cas systems, and 591 to CBASS systems.

Here, we focus on anti-CRISPR proteins (other DFs are analysed in later sections). We first evaluated the likelihood of finding anti-CRISPR elements in genomes with and without

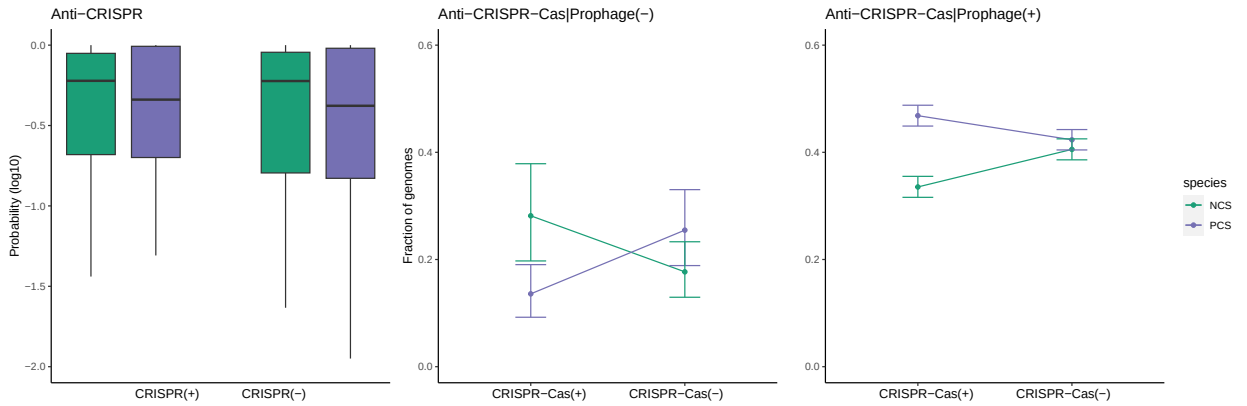

Figure 4.20: Anti-CRISPR proteins in PCS and NCS. Left: box plots indicate the fraction of genomes with anti-CRISPR per species (species with no anti-CRISPR were excluded). The central line corresponds to the median; boxes cover the data between percentiles 25 and 75. Middle and right: whisker plots represent the overall fraction of genomes harboring anti-CRISR (pooled over all species). Genomes with and without prophages were separately considered (right and middle plots, respectively) when calculating the fracton of anti-CRISPR. Whiskers represent the 95% confidence intervals. Data for genomes with prophages and without prophages are separately shown. All figures compare the prevalence of anti-CRISPR in CRISPR(+) and CRISPR(-) genomes, and in PCS and NCS (defined by setting an effect size threshold equal to the minimum significant effect size for the number of MGE genes).

CRISPR-Cas systems, both in PCS and NCS. This analysis revealed no significant differences between PCS and NCS ($p > 0.05$; Figure 4.20 left). However, a closer evaluation raised the possibility that these results were compromised by the confounding effect of prophages. Indeed, because anti-CRISPR proteins are typically found within prophages (Camara-Wilpert et al., 2023), differences in the abundance of prophages between PCS and NCS (or between CRISR+ and CRISPR- genomes) could mask the underlying differences in the prevalence of anti-CRISPR. To correct for that, we conducted a stratified analysis by separately considering genomes with and without prophages (Figure 4.20 middle and right).

As expected, a higher (approximately two-fold) prevalence of anti-CRISPR was observed in genomes that contain prophages than in those without prophages. When considering only genomes with prophages, two markedly different trends are observed in PCS and NCS. In PCS, anti-CRISPR are significantly more prevalent in genomes with CRISPR-Cas than in those without CRISPR-Cas. In NCS, the opposite trend is observed. Furthermore, the prevalence of anti-CRISPR in genomes with CRISPR-Cas is significantly higher in PCS than in NCS. These results strongly suggest that anti-CRISPR could be mediating positive associations between CRISPR-Cas and MGE abundance.

We further aimed to ascertain whether the type of anti-CRISPR proteins could determine

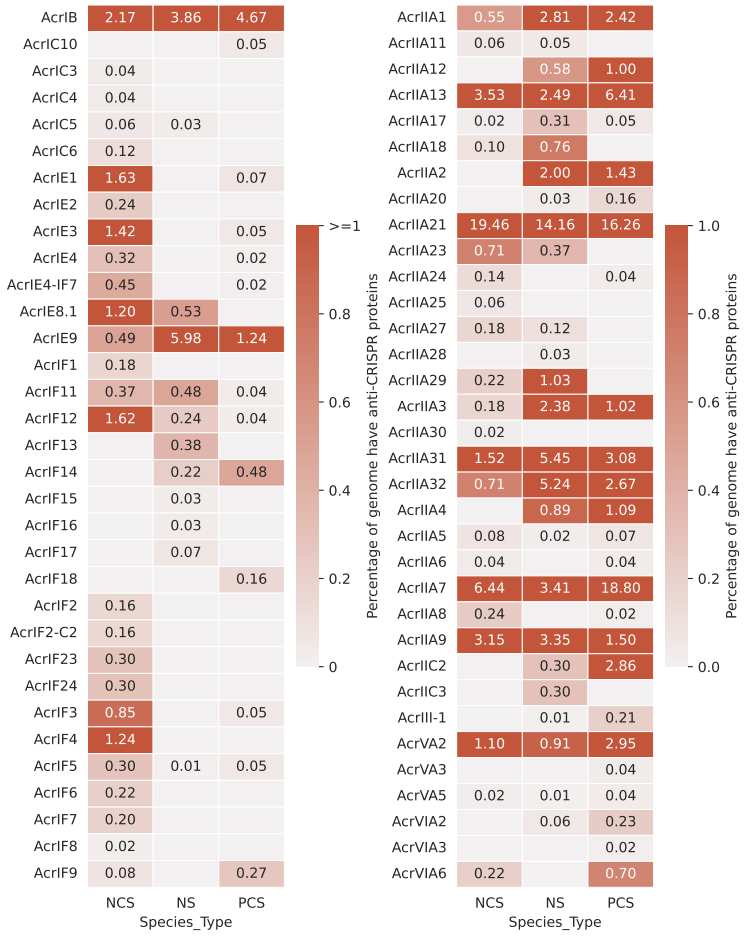

Figure 4.21: Fraction of genomes containing different types ofanti-CRISPR proteins in NCS, NS and PCS.

the sign of the correlation between CRISPR-Cas systems and MGE abundances. The diversity of anti-CRISPR types significantly differs between PCS and NCS (chi-squared test statistic=1702.60, $p < 10^{-8}$). PCS exclusively possess a variety of anti-CRISPR types including AcrIIA12, AcrIIA2, AcrIIA4, and AcrIIC2 (Figure 4.21). In contrast, NCS exhibit a broader range of anti-CRISPR proteins with low ($< 1\%$ prevalences. Despite these differences, most anti-CRISPR types have low prevalence ($< 2\%$), so the biological relevance is unclear. Moreover, we cannot rule out that the differences are driven by a narrow taxonomic span of most known anti-CRISPR types.

To account for these limitations, we focused on AcrIIA7, which shows a high prevalence and the largest difference between PCS and NCS (18.80% vs 6.44%). Recent studies have shown that AcrIIA7 is widespread and has undergone interphylum HGT events. Moreover, experimental work has confirmed that AcrIIA7 inhibits CRISPR-Cas function without binding to Cas proteins (Uribe et al., 2019). Motivated by these observations, we explored the possibility

that AcrIIA7 contributes to the positive correlation between CRISPR-Cas and MGE gene abundance. In turn, we found that the presence of AcrIIA7 is not associated with a higher abundance of genes from the mobilome in genomes harboring CRISPR-Cas (Figure A.4).

Taken together, our results suggest that anti-CRISPR proteins could be involved in the positive association of CRISPR-Cas and MGE abundance observed in many taxa, ameliorating the expected negative effects of CRSPR-Cas on gene acquisition. However, we could not identify a single anti-CRISPR type responsible for that effect. Such negative result could be partly due to the incompleteness of current anti-CRISPR databases. Therefore, further research will be required to better understand the role of anti-CRISPR in modulating the evolutionary implications of CRISPR-Cas for genome plasticity.

4.3 Effect of other defense systems on genome plasticity

4.3.1 Correlation of defense systems and genome content across species and functional categories

We expanded our research to include five prevalent defense systems (DFs): RM (Restriction-Modification Systems), Abi (Abortive Infection), Gabija, DRT (Defense-associated Reverse Transcriptases), CBASS (Cyclic oligonucleotide-based antiphage signaling system), and the DMS (a collection of potential new DNA-Modification based Systems, e.g. BREX, DISARM). Next, we studied the association between the presence of each DF and the number and fraction of genes in each functional category, using Poisson and binomial PGLMM, respectively (Figures 4.22 and 4.23). The results show that defense systems correlate with gene abundances to varying degrees, especially in the case of functional categories X, L, U, V. These findings are consistent with what we saw for CRISPR-Cas systems, where genes associated with the mobilome (X) were the most affected by the presence of DFs. In general, the new DFs show a higher number of significant positive associations with gene abundance than CRISPR-Cas. A recent report has found that many DFs are located within prophages, transposons, ICEs/IMEs or plasmids (Rocha and Bikard, 2022). In particular, Abi systems, that display the largest number of PCS for X genes are often encoded by MGEs such as plasmids, ICEs/IMEs and prophages (Samson et al., 2013). Additionally, transposons can act as carriers for defense systems (Benler et al., 2021). Such nested organization of DFs within MGE could possibly explain the high frequency of positive associations found in our analysis.

Because DFs are often part of MGE, one could expect that genomes subject to higher rates of HGT contain more DFs. On the other side, DFs could in principle block the entrance of

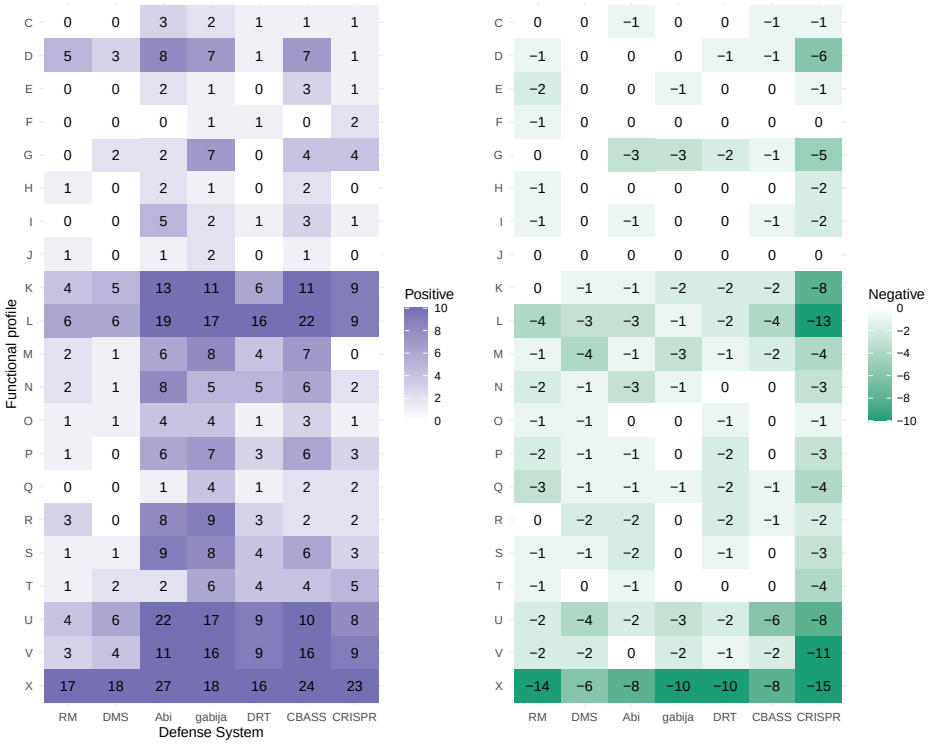

Figure 4.22: Correlation between presence of DFs and the number of genes from each functional category (Poisson-distributed PGLMM). The color map indicates the number of species that exhibited a significant correlation ($p < 0.05$) between the presence of a defense system and gene abundance.

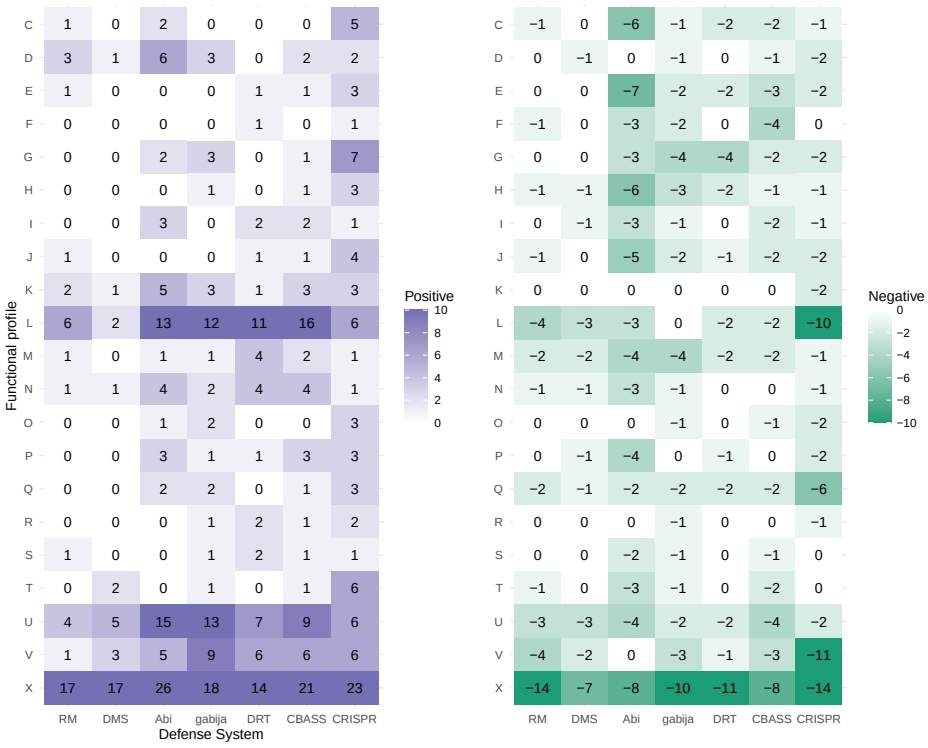

Figure 4.23: Correlation between presence of DFs and the fraction of genes from each functional category (binomial-distributed PGLMM). The color map indicates the number of species that exhibited a significant correlation ($p < 0.05$) between the presence of a defense system and relative gene abundance.

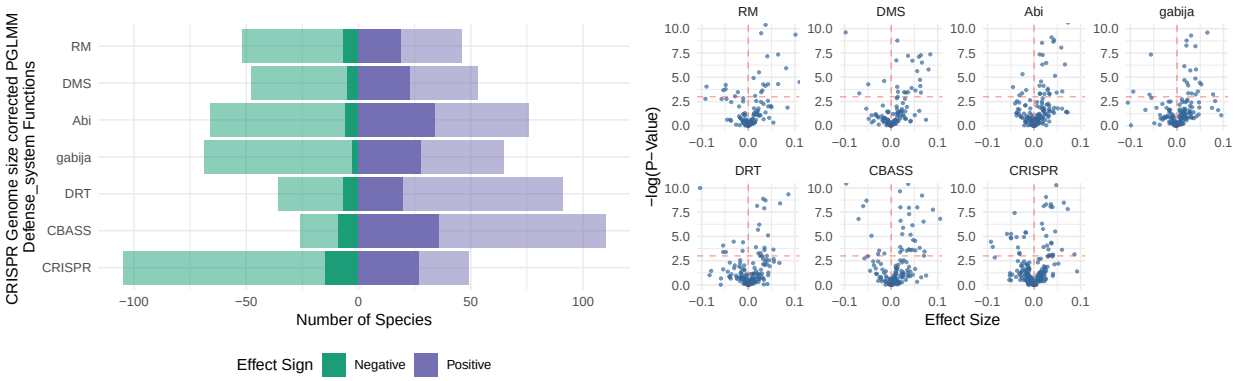

Figure 4.24: Correlation between presence of DFs and genome size (Poisson distributed PGLMM). The bar plot displays the number of species that showed a significant correlation between the presence of the DFs and genome size. Solid bars correspond to a $p < 0.05$ significance cutoff. Semitransparent bars correspond to an effect-size cutoff defined by the smallest absolute effect size that has a significant p-value. The scatter plot displays the distribution of p-values at different effect sizes (each point corresponds to one species).

MGE in the genome, reducing HGT. Because of that, the overall effect of DFs on genome size is not easy to predict in principle. We employed the PGLMM to analyze the correlation between the total number of genes and the presence or absence of the defense system across different species. We discovered 7, 5, 6, 3, 7, 9 NCS and 19, 23, 34, 28, 20, 36 PCS in RM, DMS, Abi, gabija, DRT, CBASS, respectively (Figure 4.24).

Next, we investigated the association between DFs and different types of MGE (prophages, plasmids, and transposons, Figures 4.25 and 4.26). The results confirm that the excess of PCS affects all types of MGE, although they also reveal some MGE-specific trends. For example, CBASS is positively correlated with transposon abundance in 23 species, which could be explained by the fact that Tn7-Like Transposons carry a large number of CBASS (Benler et al., 2021). Abi is widespread in ICEs/IMEs (Botelho, 2023) and showed the highest number of PCS for prophages and plasmids, possibly reflecting the incorrect annotation of some genes from ICEs/IMEs as belonging to prophages and plasmids.

Compared to CRISPR-Cas systems, other DFs display a greater excess of PCS over NCS, both in terms of genome size and MGE content. We investigated if this could be due to a weaker physical association between CRISPR and MGE. To that end, we calculated the percentage of genes from CRISPR-Cas and other defense systems located inside MGEs and tested if the value for CRISPR-Cas was lower than for the rest (Table 4.7). Our analysis confirmed that hypothesis (4.97% vs 8.12%, $p < 10^{-8}$, chi-squared test). The data indicate that the excess of PCS in other DFs compared to CRISPR-Cas could be simply due to their

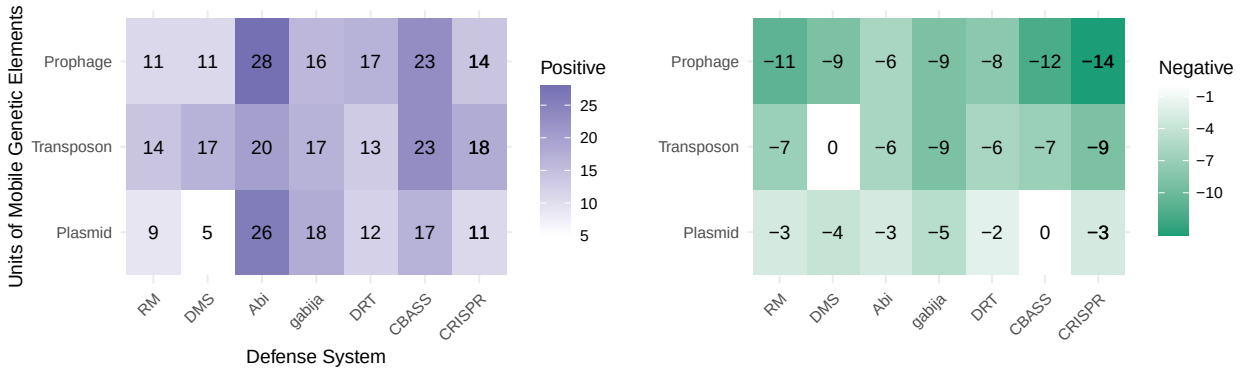

Figure 4.25: Correlation between presence of DFs and the number of genes from different MGE (Poisson-distributed PGLMM). The color map indicates the number of species that exhibited a significant correlation ($p < 0.05$) between the presence of a defense system and the number of genes of each class of MGE.

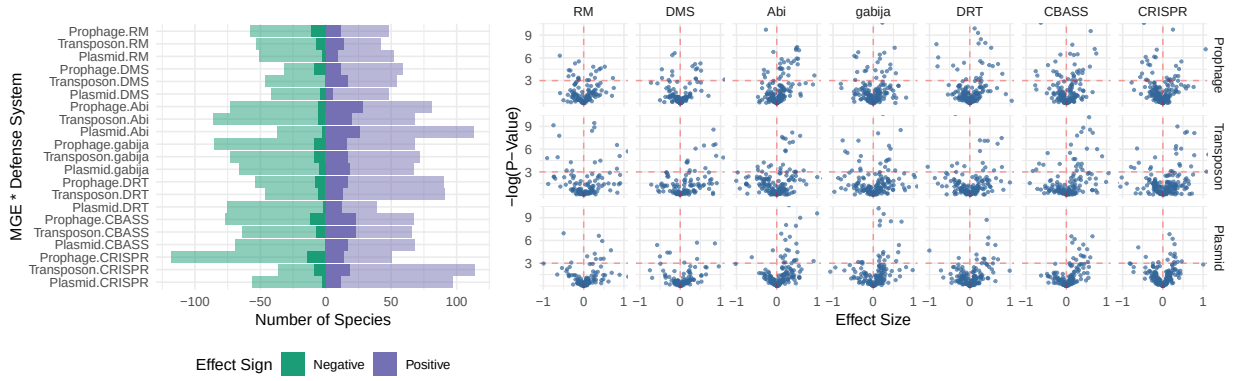

Figure 4.26: Correlation between presence of DFs and the number of genes from different MGE (Poisson-distributed PGLMM). The bar plot displays the number of species that showed a significant correlation between the presence of the DFs and the number of MGE-specific genes. Solid bars correspond to a $p < 0.05$ significance cutoff. Semitransparent bars correspond to an effect-size cutoff defined by the smallest absolute effect size that has a significant p-value. The scatter plot displays the distribution of p-values at different effect sizes (one point per species).

being more frequently located inside MGEs.

Table 4.7: Number of DF proteins inside and outside the MGEs

	In MGE	Not in MGE	Fraction of DF in MGE
CRISPR-Cas	4062	77624	4.97%
Abi	2374	18550	11.35%
CBASS	870	7184	10.80%
DMS	8545	110367	7.19%
DRT	689	5492	11.147%
RM	627	5570	10.12%
gabija	519	7048	6.86

4.3.2 Impact of defense systems on genome content is predominantly on MGEs

Our previous analyses of the CRISPR systems indicated that the correlation of the CRISPR-Cas system with L, U, and K functions are attributed to the fact that these genes are often located within MGEs. Since the same functional annotation profiles were used in this study, we hypothesize that the association of L, U, and K genes with defense systems might also be due to a portion of these genes being located inside the MGEs.

Using complete genomic data, we implemented the binomial PGLMM model for each species, both with and without masking MGE. The results show a decrease in the number of species with significant correlation, despite some correlations still persist (Figure 4.27). This implies that most genes whose abundance correlates with DFs, whether categorized as mobilome or not, are part of MGEs. Taken together, these results indicate that, generally, bacterial defense system affect genome composition through their effect on MGE.

4.3.3 Defense systems affect genome contents mainly through gene gain

We next aimed to investigate whether, similar to the CRISPR-Cas system, the associations between other DFs and gene abundances can be explained by differences in gene gain rates. For each defense system, we defined DF(+) and corresponding DF(-) clades with the same criteria as we did for CRISPR-Cas. We then compared the overall gain rates between sister DF(+) and DF(-) clades. We found that the distributions of the differences have nearly zero median and a heavy positive tail in RM, Abi, DFs, gabija and CBASS but DMS, DRT,

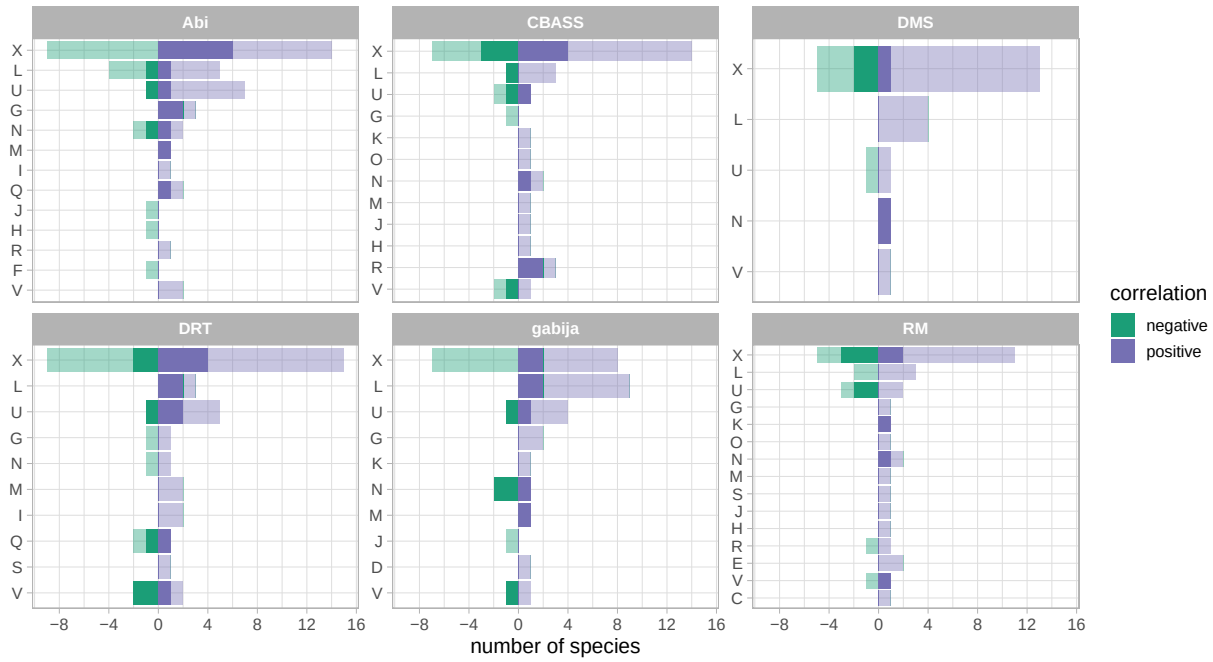

Figure 4.27: Correlation between presence of DFs and gene relative abundances (binomial-distributed PGLMM) with and without masking MGEs in complete genomes. Semi-transparent bars indicate the number of species showing significant correlations ($p < 0.05$) before masking MGEs. Solid bars indicate the number of species showing significant correlations ($p < 0.05$) after masking MGEs.

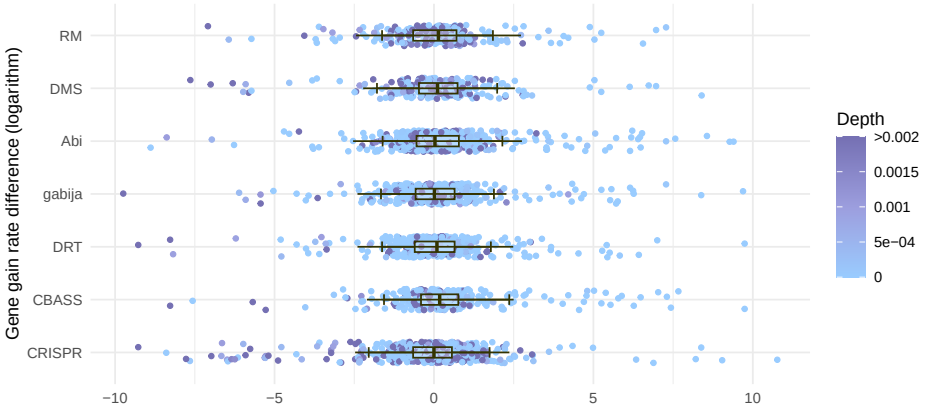

Figure 4.28: The difference of overall gene gain rates between DF(+) clade and sister DF(-) clade. The x-axis is equal to the logarithm of the gene gain rate of the DF(+) clade minus the logarithm of the gain rate of the sister DF(-) clade. The color represents the depth of the DF(+) clade in the strain tree (measured as the average number of substitutions per site in GTDB marker genes). Each point is a pair of DF(+/-) clades.

CRISPR-Cas, indicating higher gain rates in DF(+) clades (Figure 4.28; Table 4.8). The lack of a clear signal in the median is compatible with previous studies that suggest that defense systems have minimal long-term effects on bacterial fitness (Rocha and Bikard, 2022; Hussain et al., 2021; Dedrick et al., 2017; Kunitski et al., 2019; Lu and Henning, 1994; Garcillán-Barcia and de la Cruz, 2008). However, as in the case of CRISPR-Cas, outliers identify a small number of species in which DFs potentially affect HGT rates.

Interestingly, the positive tails (identified as positive outliers in the distribution) typically correspond to recent DF(+) clades, whereas the negative tails include older clades. The former is consistent with the idea that positive correlation of DFs and MGE results from joint acquisition events. The latter suggests that, in order to produce a detectable overall reduction in HGT rates, DFs need to be maintained in a clade for long enough periods of time. According to this interpretation, curtailment of HGT is most often observed for CRISPR-Cas, possibly because it more prevalent in non-mobile regions of the chromosome than other DFs (Table 4.7).

Table 4.8: Skewtest of the Figure 4.28

defense system	rate	z-score	p-val	sample size
RM	gain	1.75	0.081	253

Continued on next page

Table 4.8 – Continued from previous page

defense system	rate	z-score	p-val	sample size
DMS	gain	-1.14	0.25	245
Abi	gain	7.01	2.35e-12	495
gabija	gain	5.39	6.87e-08	424
DRT	gain	-0.78	0.434	416
CBASS	gain	5.49	4.04e-08	304
CRISPR	gain	-3.45	0.0005	549

To further investigate the causes that lead to correlations between DFs and gene abundances, we compared the distributions of gene gain and loss rates in positively and negatively correlated species. Since the effect of defense systems is typically restricted to MGE abundances (see Sections 4.3.1), we further narrowed our focus on the gene gain and loss rates of the mobilome (X category). We defined PCS and NCS based on the Poisson-distributed PGLMM results, setting an effect-size cutoff for PCS and NCS equal to the absolute value of the smallest significant effect size. Our results revealed that DF(+) and DF(-) clades often differ in their gain rates but not in their loss rates (Figure 4.29). Specifically, when looking at PCS, all defense systems are significantly associated with higher gain rates. In the case of NCS, clades with DRT, CBASS and CRISPR-Cas show significantly lower gain rates. In contrast, no significant differences in loss rates were found for any defense system, neither in NCS (Table 4.9).

Table 4.9: P-values of the Figure 4.29

Correlation	Gain vs loss p-val	Rate	DF(+) vs DF(-) p-val	DFs
negative	0.343026	gain	0.276680	RM
negative	0.343026	loss	0.402688	RM
positive	0.000109	gain	0.000055	RM
positive	0.000109	loss	0.016801	RM
negative	0.015423	gain	0.038554	DMS
negative	0.015423	loss	0.394703	DMS
positive	0.000717	gain	0.000217	DMS
positive	0.000717	loss	0.053850	DMS
negative	0.121464	gain	0.947143	Abi

Continued on next page

Table 4.9 – Continued from previous page

Correlation	Gain vs loss p-val	Rate	DF(+) vs DF(-) p-val	DFs
negative	0.121464	loss	0.285214	Abi
positive	0.016447	gain	0.001816	Abi
positive	0.016447	loss	0.033849	Abi
negative	0.614614	gain	0.632705	gabija
negative	0.614614	loss	0.965576	gabija
positive	0.012639	gain	0.006716	gabija
positive	0.012639	loss	0.076570	gabija
negative	0.003386	gain	0.001477	DRT
negative	0.003386	loss	0.277000	DRT
positive	0.000624	gain	0.000165	DRT
positive	0.000624	loss	0.074153	DRT
negative	0.115125	gain	0.005128	CBASS
negative	0.115125	loss	0.124329	CBASS
positive	0.044441	gain	0.000704	CBASS
positive	0.044441	loss	0.046011	CBASS
negative	0.000261	gain	0.003152	CRISPR
negative	0.000261	loss	0.140904	CRISPR
positive	0.002567	gain	0.021936	CRISPR
positive	0.002567	loss	0.948995	CRISPR

Next, we sought to quantify differences in gain and loss rates of prophages, plasmids, and transposons in DF(+) and DF(-) sister clades. As expected, most significant differences correspond to gene gains, with higher gain rates observed in DF(+) clades in PCS and the opposite trend in NCS (Figure 4.30 and Table 4.10). Although the general results are consistent with the joint analysis presented above, we found some differences across defense systems. The biggest reductions in the gain rates of plasmids occur in association with RM, DMS and Abi. Significant drops in gain rates of prophages are observed for DRT. In the case of transposons, significantly reduced gene gain is associated with RM.

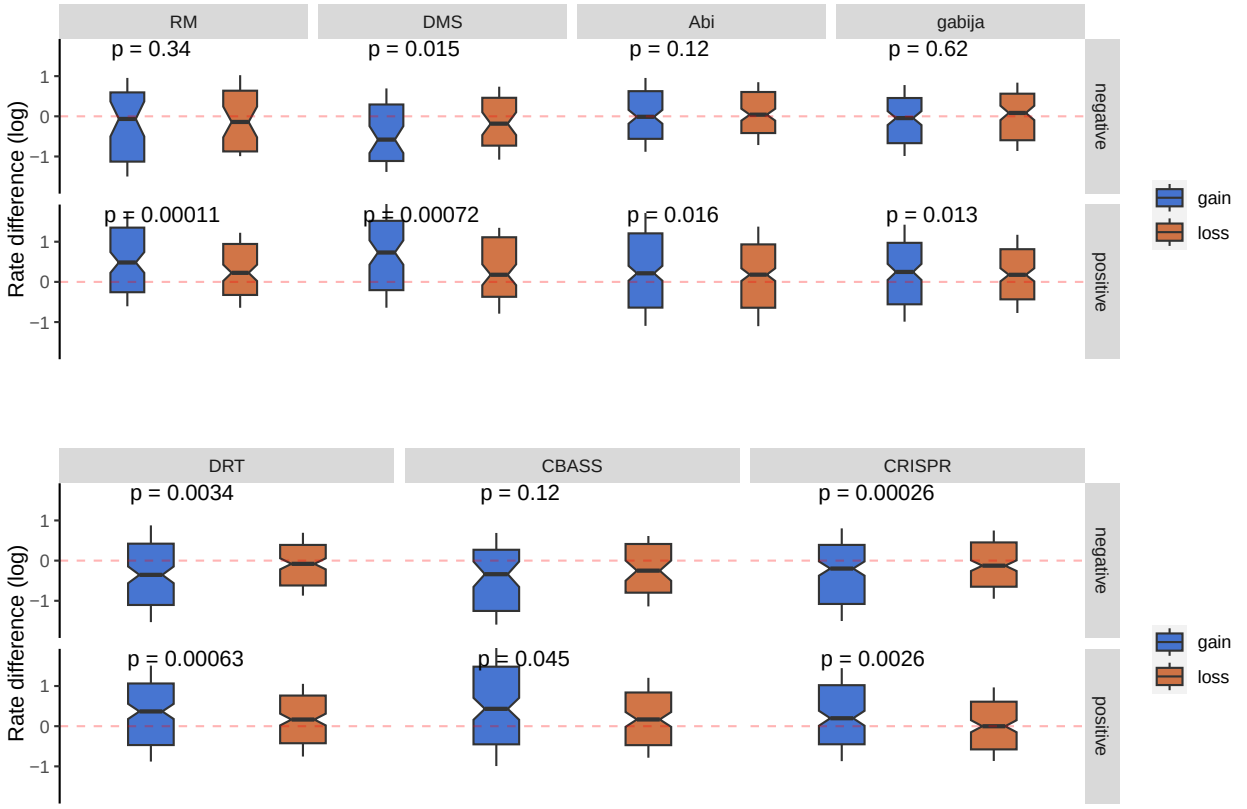

Figure 4.29: Log-difference in gene gain and loss rates between DF(+) and DF(-) sister clades for different defense systems. Gene gain and loss rates correspond to the X functional category (mobilome); "positive" and "negative" refers to the correlation between the presence of a defense system and the number of genes in the mobilome (PGLMM-based PCS and NCS). Notches indicate 95% confidence intervals for the median; the box covers percentiles 25 to 75. P-values correspond to the comparison between differences in gain and loss rates (Wilcoxon test for paired samples). Separate p-values testing the deviation of each median from zero are provided in Table 4.9.

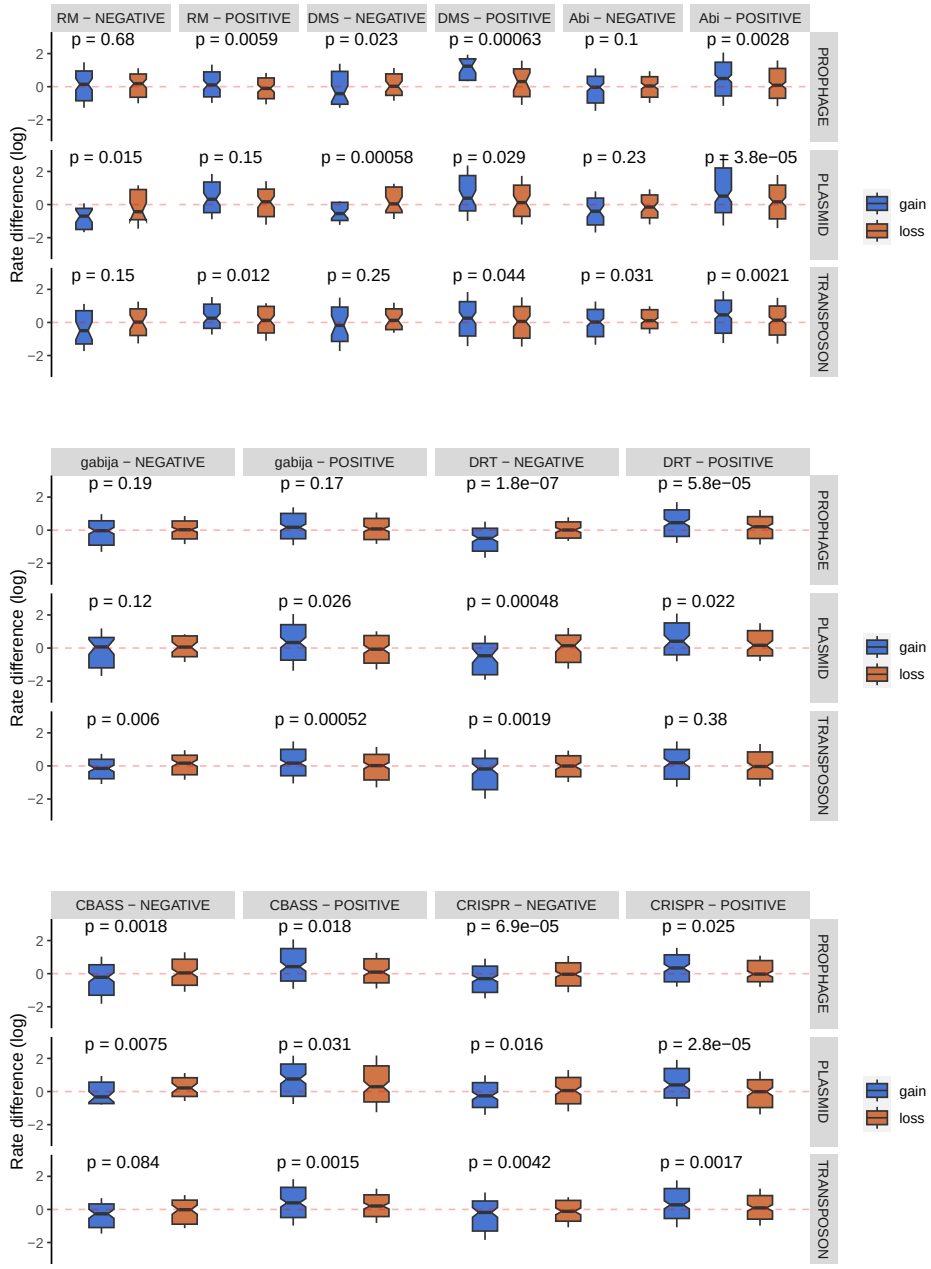

Figure 4.30: Log-difference in gene gain and loss rates between DF(+) and DF(-) sister clades for different defense systems and MGE. Gene gain and loss rates correspond to prophages, plasmids, and transposons, as indicated on the right; "positive" and "negative" refers to the correlation between the presence of a defense system and the number of genes in the MGE of interest (PGLMM-based PCS and NCS). Notches indicate 95% confidence intervals for the median; the box covers percentiles 25 to 75. P-values correspond to the comparison between differences in gain and loss rates (Wilcoxon test for paired samples). Separate p-values testing the deviation of each median from zero are provided in Table 4.10.

Table 4.10: P-values of the Figure 4.30

Category	Correlation	Gain vs Loss P-val	Rate	DF(+) vs DF(-) P-val	DFs
Prophage	negative	0.682130	gain	0.630300	RM
Prophage	negative	0.682130	loss	0.740016	RM
Prophage	positive	0.005846	gain	0.348081	RM
Prophage	positive	0.005846	loss	0.330789	RM
Plasmid	negative	0.014759	gain	0.006452	RM
Plasmid	negative	0.014759	loss	0.608884	RM
Plasmid	positive	0.145183	gain	0.017425	RM
Plasmid	positive	0.145183	loss	0.377175	RM
Transposon	negative	0.147801	gain	0.262857	RM
Transposon	negative	0.147801	loss	0.879293	RM
Transposon	positive	0.011890	gain	0.017961	RM
Transposon	positive	0.011890	loss	0.266315	RM
Prophage	negative	0.023024	gain	0.472078	DMS
Prophage	negative	0.023024	loss	0.334783	DMS
Prophage	positive	0.000631	gain	0.000217	DMS
Prophage	positive	0.000631	loss	0.263476	DMS
Plasmid	negative	0.000583	gain	0.042616	DMS
Plasmid	negative	0.000583	loss	0.124970	DMS
Plasmid	positive	0.028180	gain	0.016615	DMS
Plasmid	positive	0.028180	loss	0.466275	DMS
Transposon	negative	0.253948	gain	0.664522	DMS
Transposon	negative	0.253948	loss	0.442918	DMS
Transposon	positive	0.043481	gain	0.226669	DMS
Transposon	positive	0.043481	loss	0.810840	DMS
Prophage	negative	0.101402	gain	0.516549	Abi
Prophage	negative	0.101402	loss	0.644155	Abi
Prophage	positive	0.002773	gain	0.000013	Abi
Prophage	positive	0.002773	loss	0.043592	Abi
Plasmid	negative	0.227172	gain	0.023090	Abi
Plasmid	negative	0.227172	loss	0.406382	Abi

Continued on next page

Table 4.10 – Continued from previous page

Category	Correlation	Gain vs Loss P-val	Rate	DF(+) vs DF(-) P-val	DFs
Plasmid	positive	0.000038	gain	0.000003	Abi
Plasmid	positive	0.000038	loss	0.141703	Abi
Transposon	negative	0.030966	gain	0.889608	Abi
Transposon	negative	0.030966	loss	0.085430	Abi
Transposon	positive	0.002089	gain	0.002347	Abi
Transposon	positive	0.002089	loss	0.219939	Abi
Prophage	negative	0.188123	gain	0.196730	gabija
Prophage	negative	0.188123	loss	0.943012	gabija
Prophage	positive	0.171675	gain	0.030302	gabija
Prophage	positive	0.171675	loss	0.409595	gabija
Plasmid	negative	0.118818	gain	0.320660	gabija
Plasmid	negative	0.118818	loss	0.833407	gabija
Plasmid	positive	0.025552	gain	0.016420	gabija
Plasmid	positive	0.025552	loss	0.862485	gabija
Transposon	negative	0.005968	gain	0.055235	gabija
Transposon	negative	0.005968	loss	0.282051	gabija
Transposon	positive	0.000523	gain	0.063733	gabija
Transposon	positive	0.000523	loss	0.498651	gabija
Prophage	negative	0.000000	gain	0.000002	DRT
Prophage	negative	0.000000	loss	0.877934	DRT
Prophage	positive	0.000057	gain	0.000108	DRT
Prophage	positive	0.000057	loss	0.099517	DRT
Plasmid	negative	0.000477	gain	0.009022	DRT
Plasmid	negative	0.000477	loss	0.673089	DRT
Plasmid	positive	0.021581	gain	0.004325	DRT
Plasmid	positive	0.021581	loss	0.231098	DRT
Transposon	negative	0.001875	gain	0.048683	DRT
Transposon	negative	0.001875	loss	0.757745	DRT
Transposon	positive	0.379861	gain	0.262163	DRT
Transposon	positive	0.379861	loss	0.945352	DRT
Prophage	negative	0.001751	gain	0.027970	CBASS

Continued on next page

Table 4.10 – Continued from previous page

Category	Correlation	Gain vs Loss P-val	Rate	DF(+) vs DF(-) P-val	DFs
Prophage	negative	0.001751	loss	0.745253	CBASS
Prophage	positive	0.017776	gain	0.000621	CBASS
Prophage	positive	0.017776	loss	0.249967	CBASS
Plasmid	negative	0.007492	gain	0.295618	CBASS
Plasmid	negative	0.007492	loss	0.095683	CBASS
Plasmid	positive	0.031092	gain	0.000036	CBASS
Plasmid	positive	0.031092	loss	0.012000	CBASS
Transposon	negative	0.083768	gain	0.002923	CBASS
Transposon	negative	0.083768	loss	0.509306	CBASS
Transposon	positive	0.001513	gain	0.000576	CBASS
Transposon	positive	0.001513	loss	0.047920	CBASS
Prophage	negative	0.000069	gain	0.001454	CRISPR
Prophage	negative	0.000069	loss	0.568690	CRISPR
Prophage	positive	0.024870	gain	0.007212	CRISPR
Prophage	positive	0.024870	loss	0.697965	CRISPR
Plasmid	negative	0.016295	gain	0.062121	CRISPR
Plasmid	negative	0.016295	loss	0.759844	CRISPR
Plasmid	positive	0.000027	gain	0.005779	CRISPR
Plasmid	positive	0.000027	loss	0.513733	CRISPR
Transposon	negative	0.004163	gain	0.034051	CRISPR
Transposon	negative	0.004163	loss	0.469181	CRISPR
Transposon	positive	0.001671	gain	0.011706	CRISPR
Transposon	positive	0.001671	loss	0.510489	CRISPR

Taken together, our results are compatible with a scenario in which DFs affect genome composition through a curtailment of MGE transfer, with little or no effect on MGE loss.

4.3.4 Future directions: study of anti-defense systems

We compared the prevalence of anti-RM and anti-CBASS proteins in genomes that do and do not harbor each defense system, stratifying by PCS and NCS and by the presence or absence of prophages (since anti-defense proteins are often carried in prophages). We found

that anti-RM are significantly more prevalent in genomes with RM systems, regardless of the sign of the association between RM and MGE abundance. This behavior is different than what we observed in the case of anti-CRISPR, in which the relation between CRISPR-Cas presence/absence and anti-CRISPR prevalence qualitatively changes for PCS and NCS (Figure 4.31).

We did not find any significant trend for anti-CBASS. However, the characterization of anti-defense proteins is still an ongoing area of research. The lack of comprehensive data is currently a limitation for our analyses. Therefore, these results should be reassessed in the future as more data become available.

4.4 Co-occurrence and mutual exclulsivity of defense systems across bacteria

Previous research has shown that maintaining multiple defense systems can be costly (Wu et al., 2023; Iranzo et al., 2013; Rocha and Bikard, 2022). To study the co-occurrence or mutual exclusivity of different DFs, we fitted a set pf pairwise binomial PGLMM, with the presence (1) or absence (0) of one DF as predictor variable and the presence (1) or absence (0) of another DF as response variable. Our analysis of 197 species revealed no significant association between different defense systems in the majority of the cases (Figure 4.32). This finding suggests that the presence of one defense system does not typically influence the presence or absence of another, reflecting a considerable degree of independence in the selection and functioning of these systems within bacterial genomes. Our findings are in line with recent research, which indicates that co-occurring and mutually exclusive DFs are not conserved across bacterial orders, possibly due to environmental and genetic factors (Wu et al., 2023).

As a counterexample, a recent study of *Pseudomonas aeruginosa* suggests that the overall phage resistance scales with the number of defense systems in the bacterial genome, with some clinical isolates having over 19 defense systems (Costa et al., 2023). Our results also show a significant coexistence of RM, DMS, and Abi in *Pseudomonas aeruginosa* (Table 4.11). These observations suggest that maintaining multiple defense systems could be beneficial in *Pseudomonas aeruginosa*, perhaps due to synergistic effects against MGE or by providing resistance against diverse MGE.

Approximately 27 species show a positive correlation between RM and DMS. However, DMS does not represent a unique and well-defined defense system. Indeed, it is a collection of

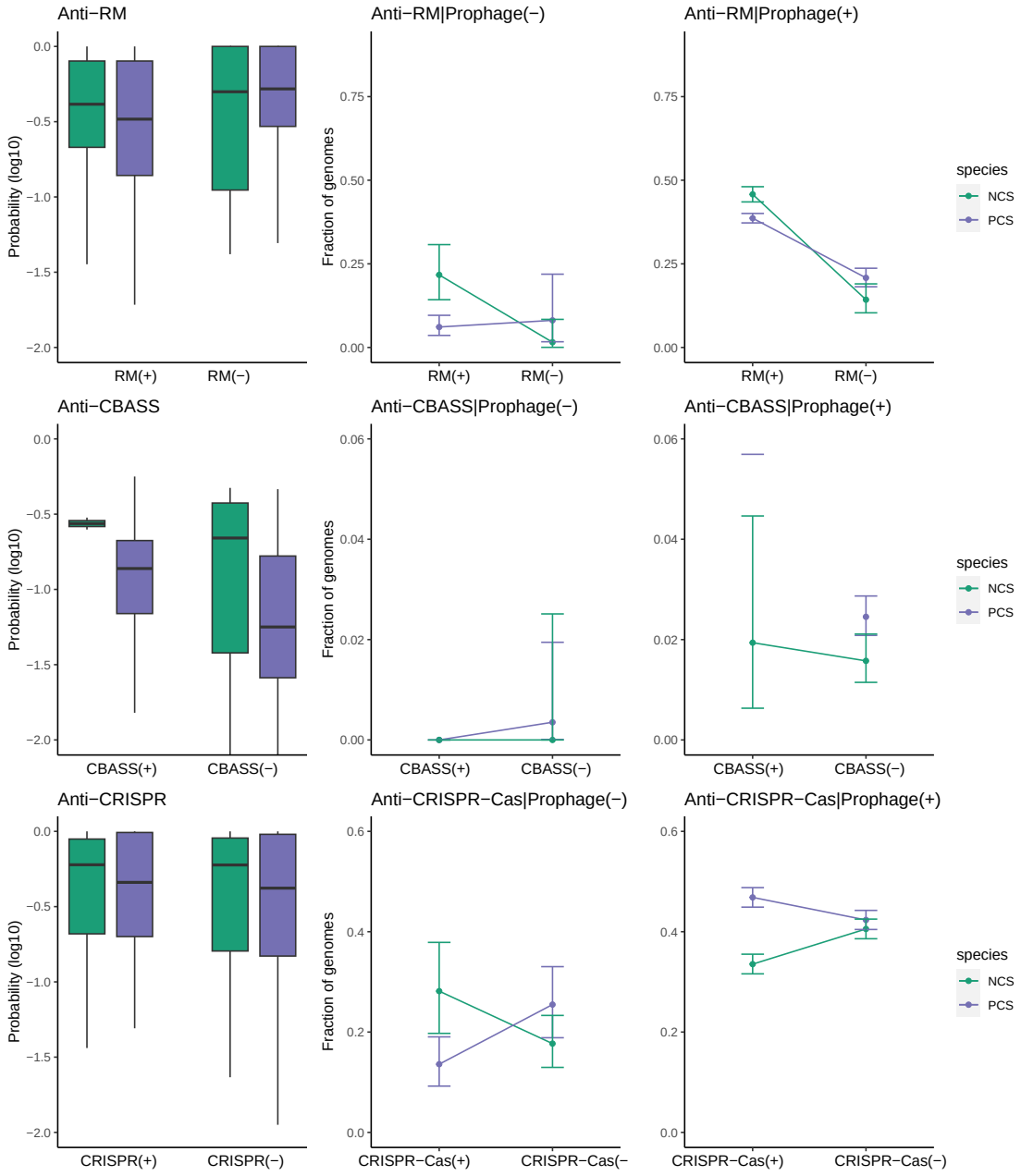

Figure 4.31: Anti-DF proteins in PCS and NCS. Left: box plots indicate the fraction of genomes with anti-DF per species (species with no anti-DF were excluded). The central line corresponds to the median; boxes cover the data between percentiles 25 and 75. Right: whisker plots represent the overall fraction of genomes harboring anti-DF (pooled over all species). Whiskers represent the 95% confidence intervals. Data for genomes with prophages and without prophages are separately shown. All figures compare the prevalence of anti-DF in DF(+) and DF(-) genomes, and in PCS and NCS (defined by setting an effect size threshold for each DF based on the minimum significant effect size for the number of MGE genes).

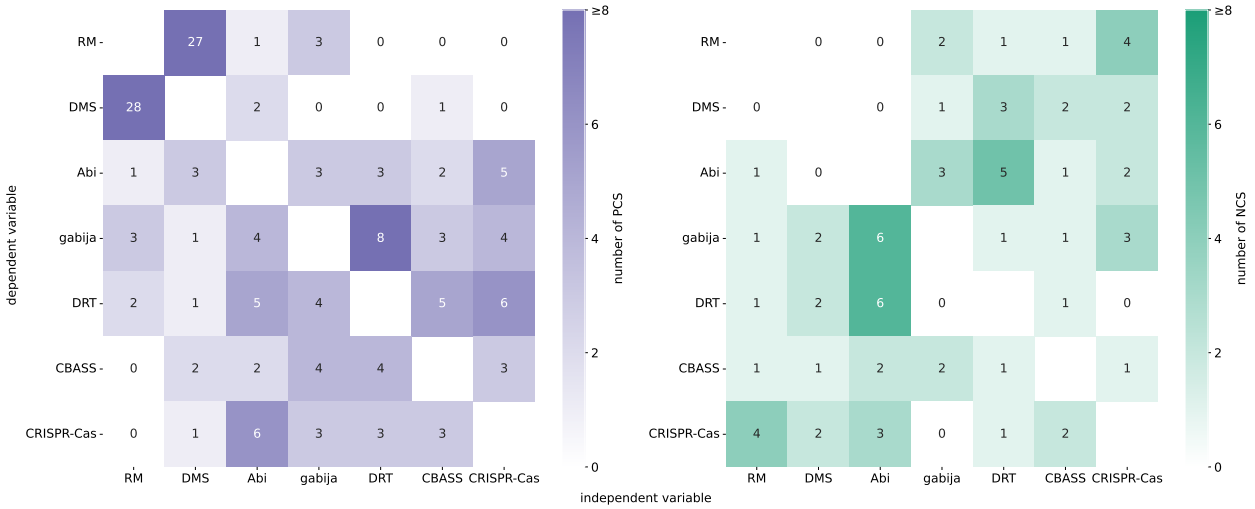

Figure 4.32: Co-occurrence (left) and mutual exclusivity (right) of different defense systems in prokaryotic genomes. The number of species with significant association was determined based on a binomial PGLMM, with a $p < 0.05$ threshold.

potential new DNA-Modification based Systems (e.g., BREX, DISARM), which could be considered as new types of RM systems. The frequent co-occurrence of RM and DMS suggests that the latter may function as synergistic ancillary or complementary components to RM systems. For example, it has been demonstrated that *E. coli* plasmids that encode both BREX and type IV RM systems offer complementary protection against phages (Picton et al., 2021). Further supporting the tight connection between RM and DMS, we found that both systems often display the same quantitative association with the gene abundances in the mobilome (Figure A.5).

Table 4.11: RM, DMS, and Abi systems significantly coexist within *Pseudomonas aeruginosa*.

Species	predictor variable	response variable	effect size	Std.Error	Zscore	p-val
P. aeruginosa	RM	DMS	3.48538	0.55495	6.2806	3.373e-10
P. aeruginosa	RM	Abi	1.28621	0.35186	3.6555	0.0002567
P. aeruginosa	DMS	RM	3.68168	0.62468	5.8937	3.777e-09
P. aeruginosa	DMS	Abi	1.35832	0.55522	2.4464	0.01443
P. aeruginosa	Abi	RM	1.34184	0.36952	3.6313	0.000282
P. aeruginosa	Abi	DMS	1.4109	0.5284	2.6702	0.00758

4.5 Do biases in genomic databases require phylogenetically corrected analyses?

Sampling bias is a known limitation of public sequencing databases. Many closely related strains are often sequenced, leading to non-random representation of genomes across the phylogeny. This problem has been widely reported in the case of ESKAPE pathogens (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter spp.* The sequenced ESKAPE strains are usually from clinical isolates where a specific Multi-Locus Sequence Typing (MLST) profile is sequenced more frequently than others. This may lead to significant differences in the presence or absence of defense systems within a given species. For instance, a study by (Botelho et al., 2023) revealed that the most prevalent MLST profile in *A. baumannii* only comprises strains without CRISPR-Cas.

We confirmed the existence of phylogenetic correlations in the presence of defense systems for some selected ESKAPE pathogens in our dataset (see Figure A.6, Figure A.8, and Figure A.7). Such phylogenetic correlations could lead artifactual results when combined with biased sampling. Thus, our intention in using the PGLMM model is to eliminate phylogenetic effects on statistical tests involving CRISPR-Cas as well as other defense systems.

4.6 Synteny-Based Clustering of Orthologous CRISPR Arrays

CRISPR-Cas systems are capable of recognizing viruses and other MGEs, avoiding their proliferation in bacteria and archaea. A key property of CRISPR-Cas is its adaptive nature. That is, the system integrates spacers derived from MGEs into CRISPR arrays, serving as a memory to combat future invasions (Shmakov et al., 2017). The rate at which CRISPR spacers are updated determines the ability of CRISPR-Cas systems to adapt to changes in the composition of the local virome and counteract MGE escape mutations. Metagenomic studies of free-living microbial communities have shown that the rates of spacer acquisition and loss are highly variable across species and environments (Meaden et al., 2022). Therefore, our original goal was to conduct a large-scale, systematic study of spacer turnover rates in CRISPR arrays using public data from high-quality genomes. The central idea of our method was to compare the divergence in the spacer composition of CRISPR arrays with the divergence time of the carrier strains. Because CRISPR-Cas systems experience high rates of HGT, the first step was to identify groups of CRISPR arrays that were related by vertical descent,

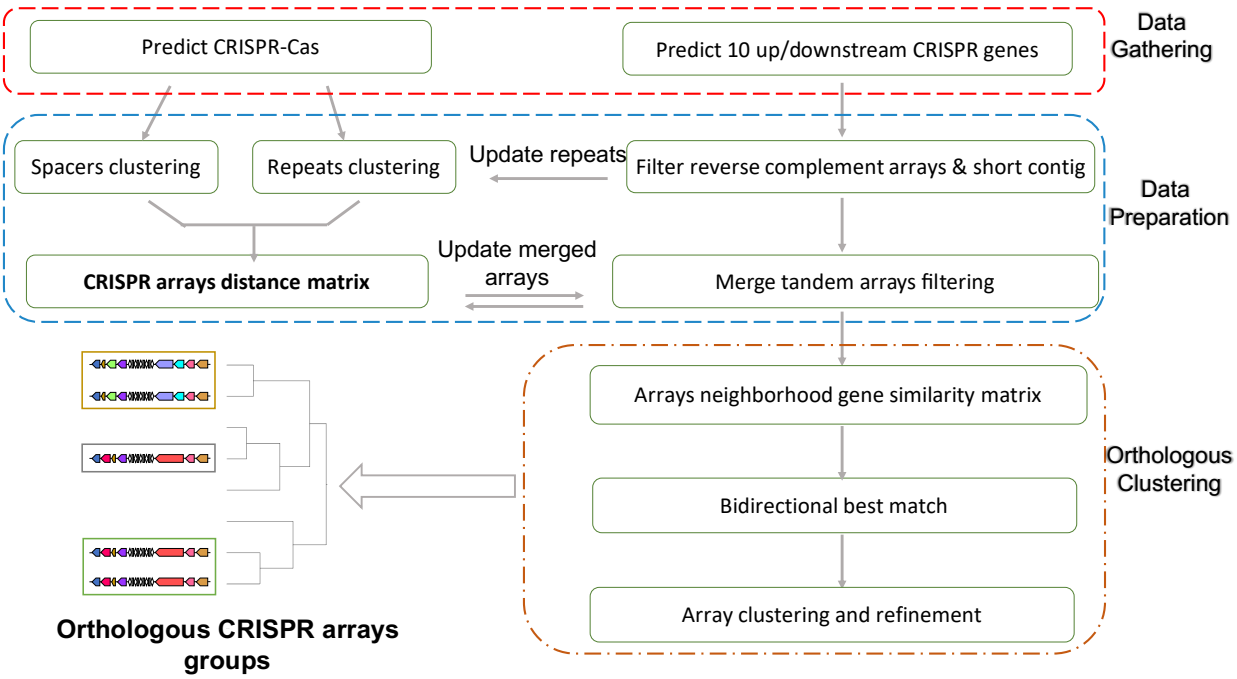

Figure 4.33: Workflow for clustering orthologous CRISPR arrays in prokaryotic genomes.

that is, orthologous CRISPR arrays. Identifying orthologous CRISPR arrays is a challenging task because of their fast divergence (Shmakov et al., 2018). We employed a synteny-based method, which effectively minimizes HGT contamination. The rationale of this method was to identify clusters with conserved genomic locations, and therefore unaffected by recent HGT events. We developed a semi-automatic workflow comprising the following steps (Figure 4.33): (1) classification of the spacers and repeats from CRISPR arrays, (2) curation of reverse complement repeats and tandem arrays, (3) calculation of CRISPR array distances based on the fraction of shared spacers, (4) identification of candidate species-level orthologous CRISPR arrays based on shared repeats, (5) collection of CRISPR array neighborhoods (6) obtention of array distance matrices based on neighborhood synteny conservation, (7) construction of bidirectional best match network, (8) clustering of orthologous CRISPR arrays using algorithms for network community detection, (9) ancestral reconstruction of CRISPR-Cas presence/absence and subsequent refinement of orthologous CRISPR arrays..

Following this procedure, we extracted 24,446 arrays from 11,620 CRISPR(+) genomes across 197 species, collected 2,836,048 pairwise distances and clustered 12,832 CRISPR arrays into non-singleton groups (see Methods for details). Then, for each pair of arrays from the same cluster, we compared the divergence in spacer content with the phylogenetic distance between the genomes (Figure 4.34). As expected, for any fixed spacer content divergence, the phylogenetic distances between orthologous CRISPR arrays are generally smaller than

those involving non-orthologus pairs. This validates our approach to orthology detection based on genomic context. However, due to the small size of most orthologous groups the amount of informative data was insufficient to ascertain the spacer turnover rates of CRISPR arrays. Even in clusters with enough sample size, the relation between phylogenetic and spacer content distances is often noisy, with low R^2 values, while estimation of turnover rates would require good linear fits. Two possible causes for this lack of a clean linear trend are (a) the existence of frequent recombination between CRISPR arrays from closely related strains, and (b) a high variability of spacer turnover rates across strains.

As future directions, we plan to focus on species with a dense sampling of closely related genomes, in which sample size is not limiting, and use alternative methods such as those developed by (Kupczok et al., 2015; Ou and McInerney, 2022) to assess cross-strain recombination and variability in spacer turnover rates.

There are three critical criteria to infer the evolutionary history of CRISPR based our database: 1) Phylogenetic tree distance, which determines the CRISPR evolution along bacterial homologous recombination; 2) CRISPR array distance, which assesses the history of CRISPR encounters with mobile genetic elements (MGEs); and 3) CRISPR array neighborhood distance, which identifies CRISPR arrays from the same ancestry. Theoretically, CRISPR arrays of the same origin are expected to encounter similar MGEs (have similar spacers) or to stem from phylogenetically close bacteria. A divergence from this pattern suggests possible horizontal transfer, as opposed to vertical inheritance from parent to offspring. In addition, our orthologous databases could facilitate the prediction of CRISPR function by enabling comparative analysis between arrays with known functions and those with uncharacterized roles providing insights into the potential roles and activities of unannotated arrays through neighborhood analyses.

4.7 Comprehensive functional and evolutionary analysis of a large collection of Phage-inducible chromosomal islands and P4-like satellites

4.7.1 Identity new protein family in phage satellites and PICIs

In the preceding section, we found that the effects of defense systems on the transfer of mobile genetic elements is species-dependent. The effect sizes vary across different species. To further investigate the defense systems harbored within MGEs, we focused our attention on phage

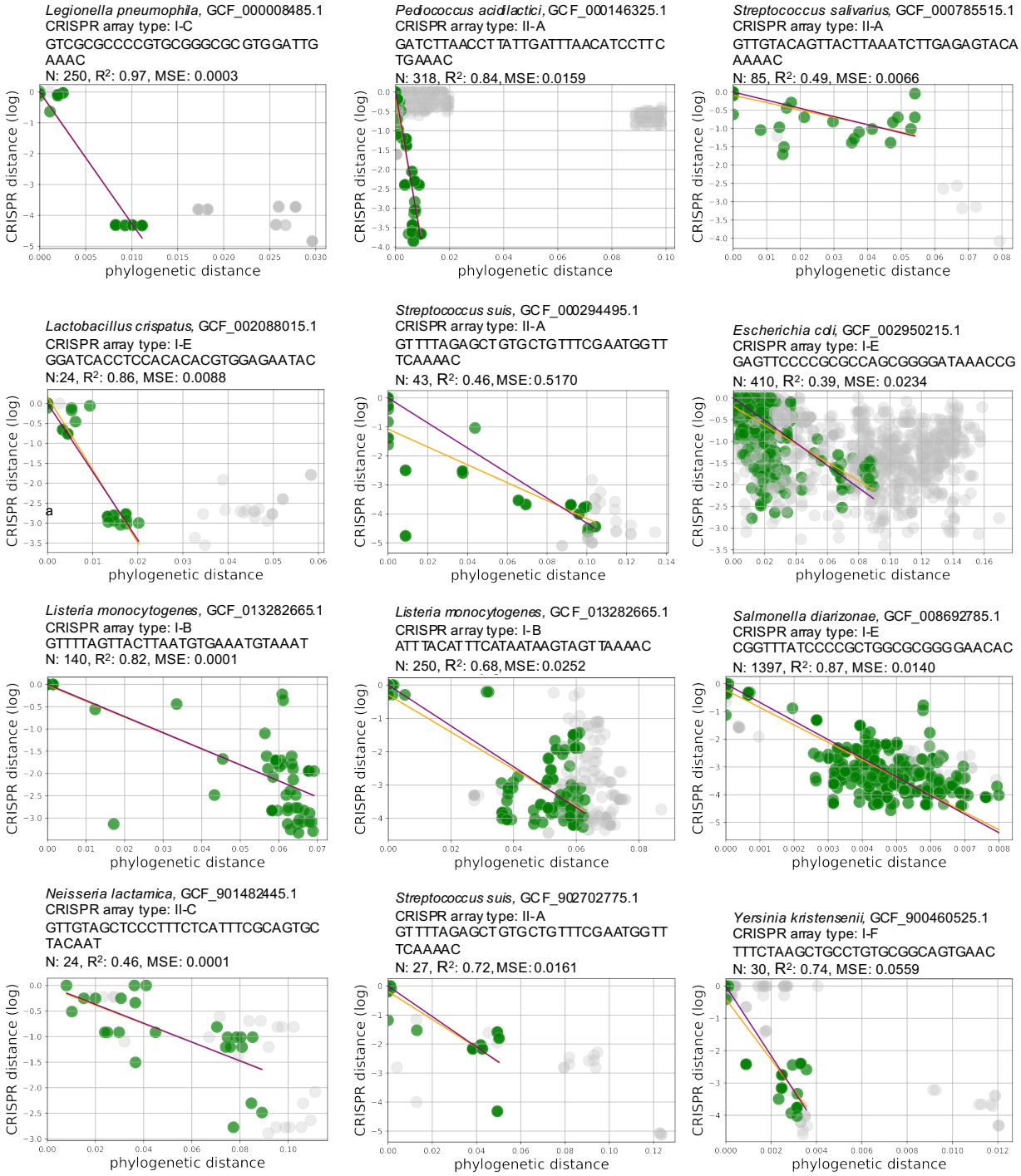

Figure 4.34: Orthologous CRISPR turnover rates. Each green dot is a comparison of a pair of orthologous arrays; each grey dot is a comparison of a pair of non-orthologous arrays. The yellow line is the linear regression of the green dot; and the purple line is regression the with 0 intercept. The head text of each sub-figure indicate the GTDB speices name, representative genome id, type and repeats of the CRISPR array, sample size, coefficient of determination and mean squared error.

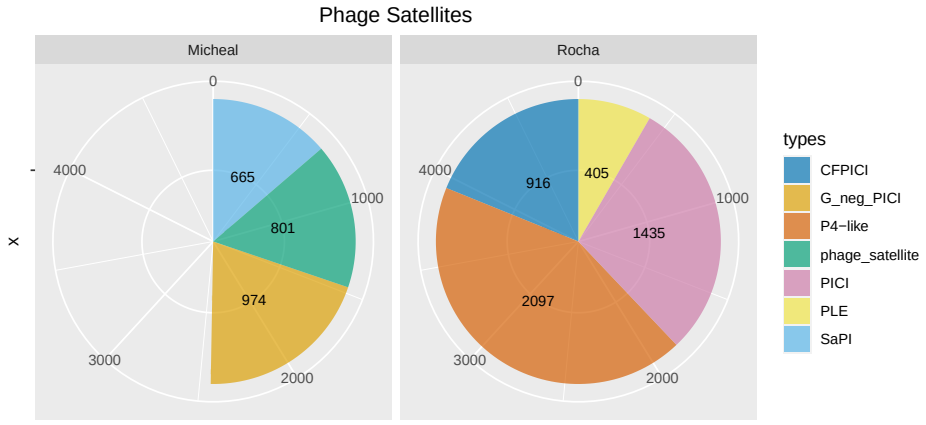

Figure 4.35: Composition of the dataset studied this section. Phage satellites are organized according to their original classification. Abbreviations, PICI: Phage-inducible chromosomal islands; CFPICI: capsid-forming PICI; G_neg_PICI: Gram negative PICI; P4-like: P4-like phage satellites; PLE: PICI-like elements; SaPI: Staphylococcus aureus pathogenicity islands.

satellites. This subclass of hyperparasitic and hypermobile MGEs often encodes mobility genes of phage origin that enable them to adapt and disseminate horizontally across bacterial hosts. The commonality that brings the diverse group of phage satellites together is that they rely on helper phages, which they hijack to package their own DNA into phage capsids instead of the phage DNA. Due to the intrinsic characteristics of phage satellites and their complex way of propagation, they have developed a close relationship with defense systems, which provide them with the ability to navigate the intricate interactions with bacteria and other phages (Rocha and Bikard, 2022).

The analysis of phage satellites in this section was primarily based on two sources: a dataset published by Moura de Sousa (de Sousa et al., 2023) and an in-house dataset produced by Michael Widdowson at the University of Copenhagen.

Our initial investigation only included PICIs and cfPICIs, as the marker genes that we used to fish them out were only specific to these elements and not P4-like satellites or PLEs. In the pipelines, the gathering and categorization of the various phage satellites were conducted based on different features. Primarily, phage satellites were pinpointed by the co-localization

of primase-replicase genes, coupled with the absence of *alpA* and *sis* genes. Specifically, SaPI-type PICIs were identified through the proximity of colocalized *alpA* or *sis* genes to the integrase. Sequences were categorized as gram-negative PICIs under certain conditions, including the absence of *alpA* and *sis*, alongside additional distinctive features of gram-negative PICIs (see https://github.com/lyonliuyang/phd_thesis_supp).

Following these criteria, we successfully identified a diverse array of phage-related elements: 685 SaPIs, 801 phage satellites, and 974 Gram-negative PICIs, that we combined with 2097 P4-like elements, 1435 PICIs, 916 CFPICIs, and 405 PLEs shared by Eduardo Rocha's lab, as illustrated in Figure 4.35. In total, these elements encompass 125,437 phage satellite proteins, which we categorized as follows: 19,176 G-neg-PICIs, 18,486 phage-satellite, 15,135 SaPIs, 20,594 PICIs, 17,862 CFPICIs, 24,363 P4-like elements, and 9,821 PLEs. Ultimately, we clustered these proteins into distinct families. A network of satellite phages was constructed based on the proportion of shared protein families, and clusters of similar phage satellites were identified using the weighted Louvain community detection method (Figure 4.36). Based on this clustering, we reclassified the original sequences into 414 PLEs, 1140 SaPI-like elements, 2131 P4-like elements, 3209 PICIs, and 429 undefined satellites. We expect that the reclassified dataset will provide resources for future research on phage satellites and the discovery of new protein families.

4.7.2 Accessory functions enriched in different satellite families

We selected satellite genes along with 10 adjacent upstream and downstream genes, annotating these genes based on several databases, including defense system, virulence factors, and antibiotic resistance genes. Our decision to include neighboring genes was motivated by two reasons: 1) Co-located proteins often display functional synergy (Wu et al., 2023), and 2)existing software tools have difficulties in precisely delineating phage satellite boundaries using HMM marker genes. By applying the network-based clustering described above, we categorized the annotation profiles into the four groups: PLEs, SaPIs, P4-like elements, and PICIs (Figure 4.37).

The annotation of defense systems (Figure 4.38) indicates that P4-like elements typically possess more diverse defense systems than other elements. Both SaPI and PICI harbor Abi2/AbiD, which mediate parasitism on helper phages and competition with other phages. Additionally, PICI carries AbiJ, while AbiU is found on P4-like elements, and AbiE in their vicinity. The abortive infection (Abi) system is involved in different mechanisms but they share the common goal of protecting bacteria and aborting other phages. AbiJ typically functions by inducing cell death upon phage infection (Anantharaman et al., 2013), thereby

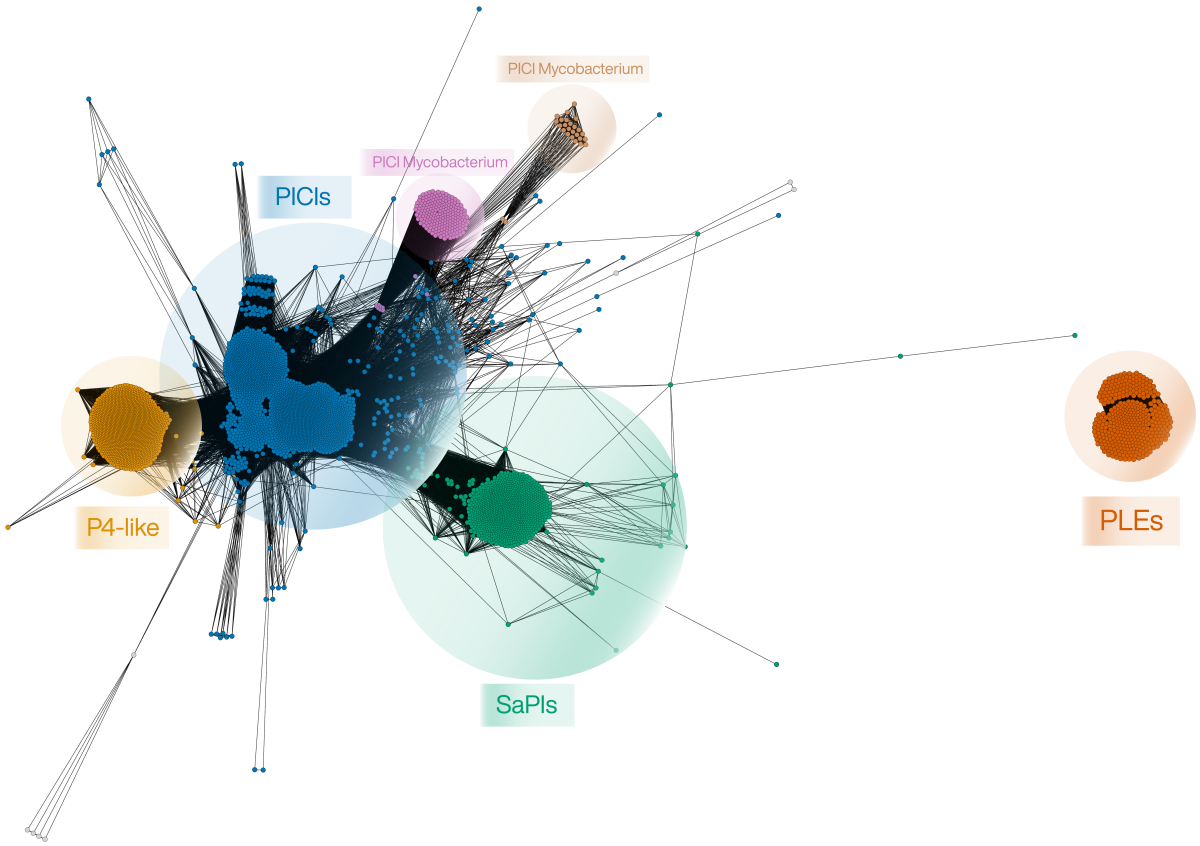

Figure 4.36: Phage satellites gene sharing network. Each node represents a phage satellite, and each edge represents the similarity score between a pair of nodes, weighted by the protein families they share. The protein families are defined by embedding-based protein clustering. The color represents reclassified phage satellites.

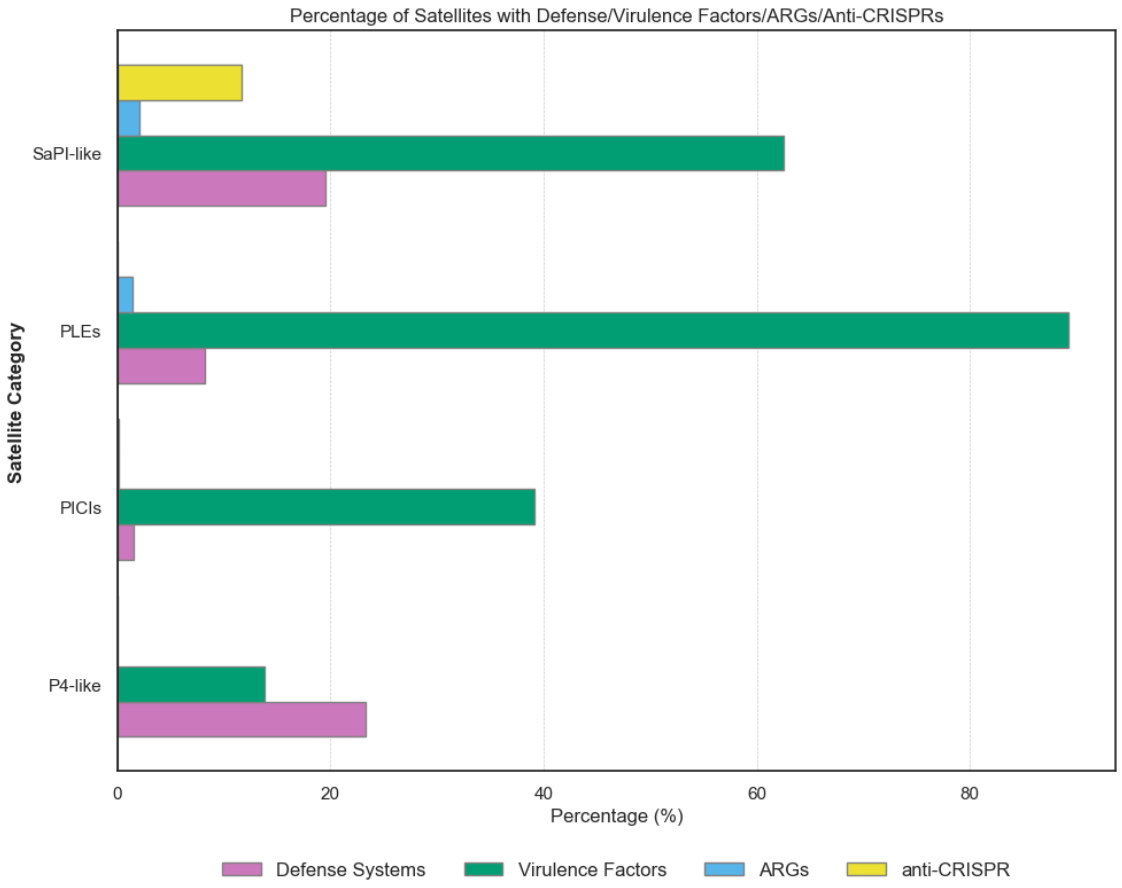

Figure 4.37: Summary of four types of phage satellites annotation. The bar displays the fraction of satellites that carried those element compared to total number of corresponding satellites. The y-axis represents reclassified phage satellites. The color denotes the elements on satellites or in their vicinity.

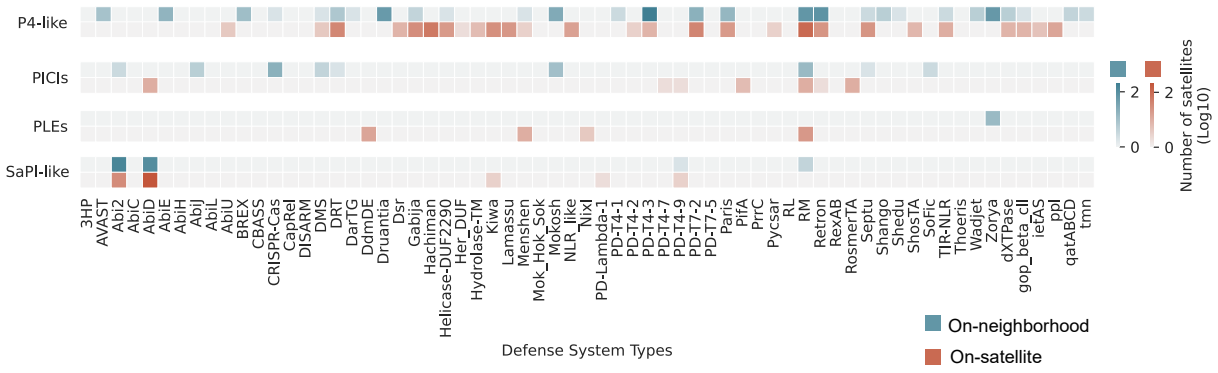

Figure 4.38: Defense systems in phage satellites. The number was determined by counting the phage satellites that carried at least one corresponding defense system gene or had it present in their upstream or downstream 10-gene neighborhoods. The link to the list of systems is available at https://github.com/lyonliuyang/phd_thesis_supp.

preventing the replication and spread of the phage (Dy et al., 2014). On the other hand, AbiU might interfere with the nucleic acid replication process within the bacterium (Dy et al., 2014). AbiD, or Abi2, is known to inhibit phage replication, possibly by interacting with phage replication proteins (Dy et al., 2014; Anantharaman et al., 2013). Restriction Modification (RM) systems are present on P4, PICI, PLE, and in the vicinity of SaPI. A number of different defense systems have been identified in the vicinity of PICI and P4. Diverse DNA-Modification based Systems (DMS) are observed the vicinity of both PICI and P4. In addition, the gabija system is specifically detected adjacent to P4, while the DRT systems are predominantly found near PICI and the CBASS system is absent in PICI. CRISPR-Cas systems, predominantly Class 1 Subtype I-F, were identified in the neighboring genes of PICIs and P4-like elements. Type III-A CRISPR system was found in *Pectobacterium carotovorum* (GCF_000023605.1). However, its prevalence cannot be confirmed based on a single instance.

Further analysis of the defense systems across various genera (Figure 4.39) shows that P4-like elements are the satellites with the highest prevalence of defense systems across genera. PICIs are also major carriers of defense systems in multiple genera, with a notable presence in *Bacillus_A* and *Xenorhabdus*. Defense systems previously identified in P4-like elements were acquired by phylogenetically closely related genera, such as *Citrobacter*, *Escherichia*, *Enterobacter*, *Klebsiella*. Additionally, similar PICIs were found to carry different defense systems (Figure A.9). For instance, RM systems in PICIs from *Shewanella* have different protein compositions in phylogenetically close species.

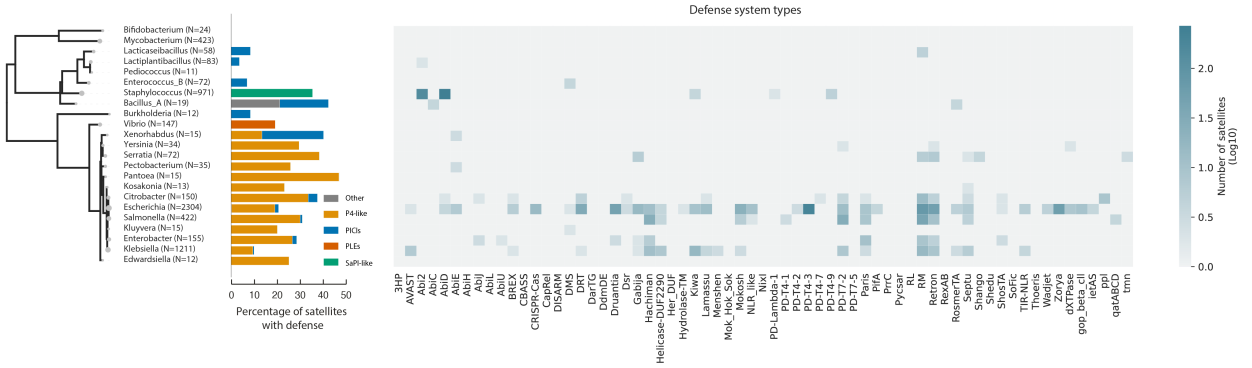

Figure 4.39: Defense systems in phage satellites organized by genus. The number was determined by counting the phage satellites that carried at least one corresponding defense system gene or had it present in their peripheral 10 genes. The link to the list of systems is available at https://github.com/lyonliuyang/phd_thesis_supp.

These results highlight the rapid genomic dynamics not only of MGEs but also of defense systems, underscoring the complex, nested genomic structures that reflect the intricate competition between viruses and bacteria.

As both defense systems and anti-defense mechanisms can be carried within MGEs, it is not difficult to assume that they could also recruit anti-defense mechanisms to facilitate their selfish nature. Our findings indicate that PICI and SaPI-like elements contain anti-CRISPR and anti-defense proteins within the phage satellite genome and adjacent regions (Figure 4.40). These findings suggest the possibility of co-transferring phage satellites, defense systems, and anti-defense systems. This could potentially associate with the effect of defense systems on HGT and genomic plasticity.

Virulence factors (VFs) were identified within the clustered phage satellites and on their adjacent sequences (see Figure 4.41). For example, the Type IV and Type VI secretion systems were detected in PLEs. Additionally, the Per and PhoPQ two-component systems were present in PICIs and their adjacent regions. The presence of the iron-related phenotype, pyoverdine, was observed in both PICIs and PLEs. A genus-specific analysis of virulence factor (VF) genes indicated that PICIs generally harbor a greater number of virulence genes across species (Figure 4.42). Additionally, unclassified phage satellites in *Bifidobacterium* and *Mycobacterium* were found to exhibit a substantial number of VFs. These phage satellites might enhance the pathogenicity of bacteria by living in symbiosis with them. Besides, since they lack defense systems (Figure 4.39), it is less likely to frequently interact with other phages other than helper phage or the defense system on them remain unidentified. SaPIs and PLEs are typically found within a single genus and result in a distinct range of VF types.

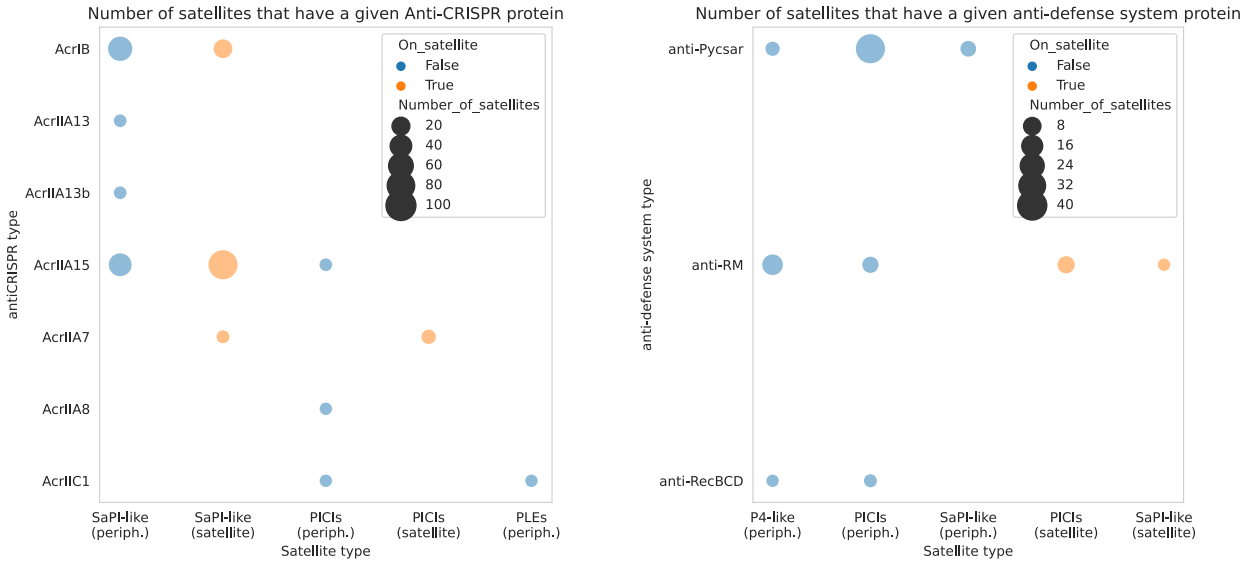

Figure 4.40: Anti-defense proteins on phage satellites. The numbers were determined by the phage satellites that have at least one corresponding defense protein on themselves or located on their periphery.

In contrast, PICIs and P4-like satellites are present across various genera.

Correspondingly, Antibiotic resistance genes (ARGs) were identified (see Figure 4.43; Figure 4.44). SaPIs contain high numbers of ARGs within their regions, while PICIs have a higher concentration of ARGs in their neighboring areas (Figure 4.43). Specifically, SaPIs include four types of ARGs: fusidane, penam, phosphonic, and tetracycline. In contrast, PICIs predominantly contain elfamycin ARGs. PLEs, on the other hand, demonstrate unique encompass glycopeptide ARG.

In summary, we have identified various defense systems, including VF, ARG, and anti-defense genes, in and in proximity to phage satellites. Our findings show:

- P4-like and PICI satellites exhibit a higher number and greater variety of defense system types compared to other satellites.
- Phage satellites usually exhibit a lower abundance of ARGs on satellites than in their neighborhood.
- PICIs, PLEs, and SaPI-like satellites are rich in virulence factors.
- SaPI-like elements and PICIs often possess anti-CRISPR mechanisms to circumvent host immunity.

Looking ahead, we anticipate identifying new protein families of phage satellites from clustered

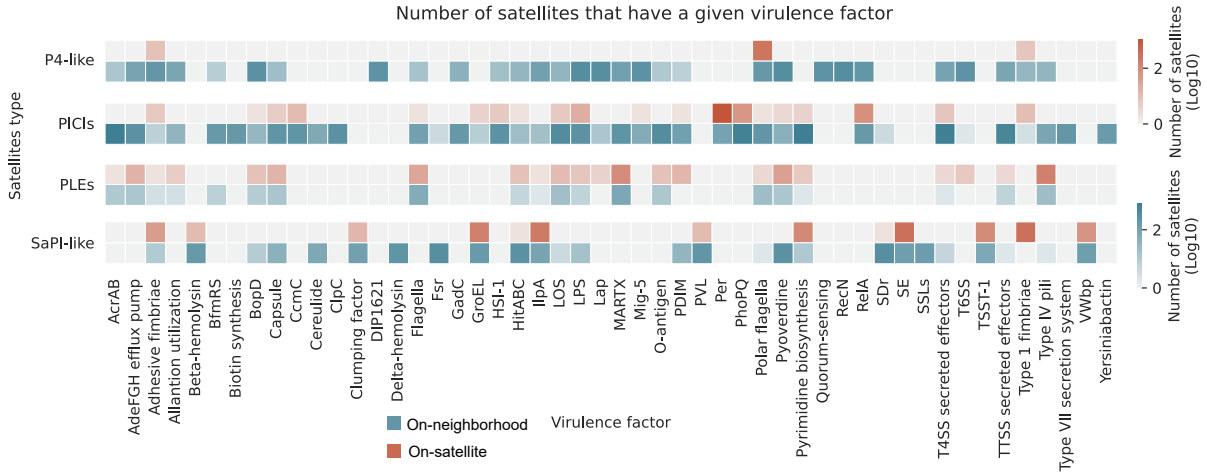

Figure 4.41: The virulence genes on phage satellites. The number was determined by counting the phage satellites that carried at least one corresponding virulence gene or had it present in their peripheral 10 genes.

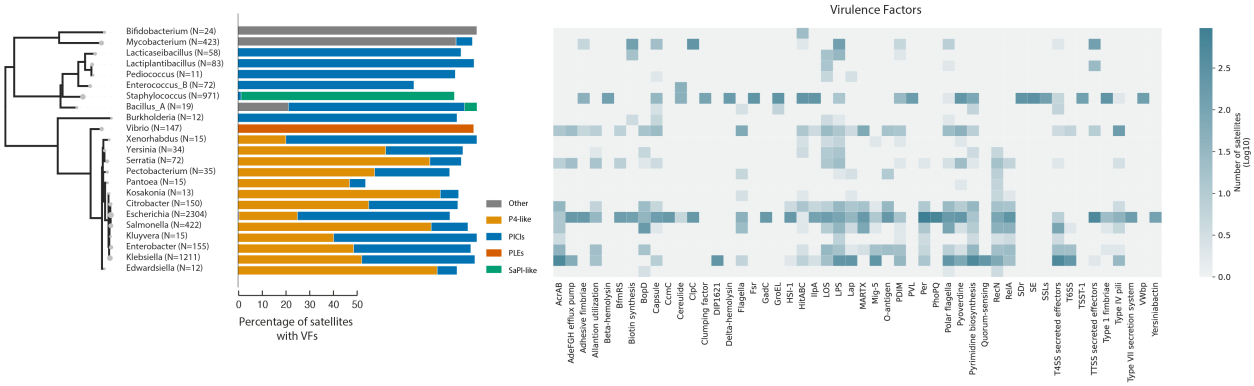

Figure 4.42: The virulence genes on phage satellites organized by genus. The number was determined by counting the phage satellites that carried at least one corresponding virulence gene or had it present in their peripheral 10 genes.

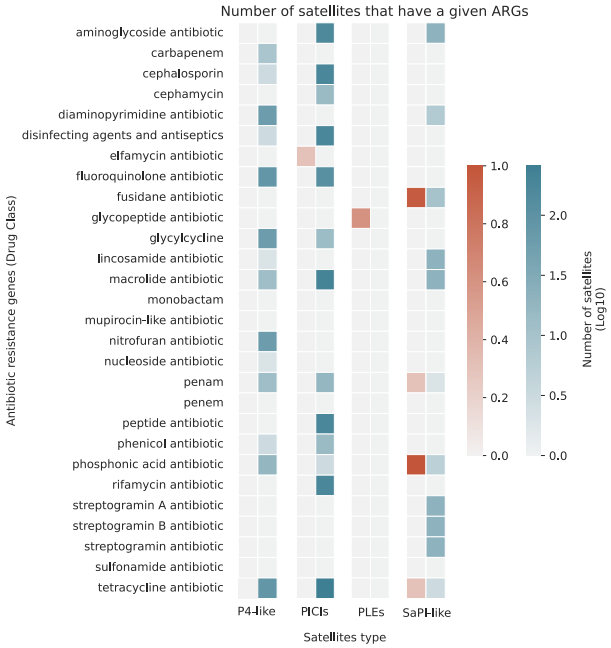

Figure 4.43: The ARGs on phage satellites. The number was determined by counting the phage satellites that carried at least one corresponding antimicrobial resistance gene or had it present in their peripheral 10 genes.

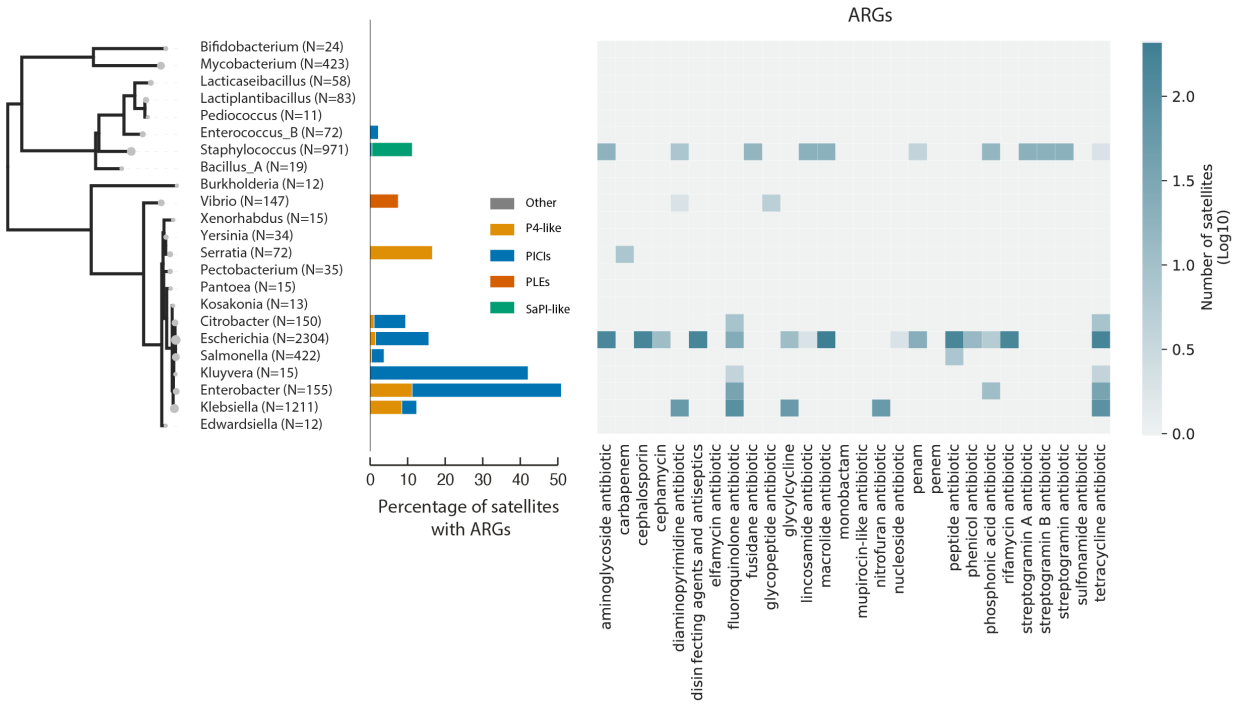

Figure 4.44: The ARGs on phage satellites organized by genus. The number was determined by counting the phage satellites that carried at least one corresponding antimicrobial resistance gene or had it present in their peripheral 10 genes.

protein groups. Additionally, we plan to continue studying the interplay between MGEs and defense systems in prokaryotic genomes.

Chapter 5

Discussion

5.1 Gene clustering criteria for pangenome analyses

Orthology is often considered the best criterion for clustering gene sequences in comparative genomics (Gabaldón and Koonin, 2013). This is because orthologs tend to maintain their function and evolve consistently with speciation patterns, making them the most appropriate choice for functional and phylogenomic studies. However, in practice, using orthology as a gold standard for pangenome analysis presents technical and conceptual challenges. On the technical side, distinguishing orthologs from paralogs can be computationally expensive. Therefore, to handle large (meta)genomic datasets, orthology prediction tools often use heuristic algorithms. However, these algorithms may miss true orthologs or include out-paralogs (paralogs that duplicated before the last common ancestor of the clade of interest). On the conceptual side, gene duplication within a species and horizontal gene transfer can create gene lineages with multiple copies. This can complicate downstream analyses that assume vertical transmission, even if the orthology criterion at the species level is not violated. In such cases, it may be preferable to use a stricter criterion that only groups together members of a gene family that have been vertically transmitted.

In light of these considerations, we evaluated the impact of three primary clustering criteria (homology, orthology, and synteny) as implemented by five commonly used pangenome analysis tools on pangenome reconstructions and subsequent phylogenomic analyses. Although we tested only a limited number of tools and parameter settings, our results indicate that the underlying formal criterion for paralog discrimination –shared ancestry at speciation for orthology, conserved gene neighborhood for synteny– is what drives qualitative differences across methods.

Previous studies have shown that estimates of pangenome size and diversity for a single species are highly dependent on the gene clustering method used (Tonkin-Hill et al., 2020; Sitto and Battistuzzi, 2020; Bayliss et al., 2019). We have confirmed these observations and expanded on their implications for comparative pangenome analyses. In comparative studies, inconsistencies in relative differences and cross-species trends are of greater concern than absolute differences in single-species estimates. Regarding genome size trends, pangenome and core genome sizes are generally consistent across methods. However, cross-species comparisons of genome plasticity and pangenome diversity are highly sensitive to the clustering criterion, which cannot be explained by linear or nonlinear data transformation. Inconsistencies in pangenome diversity, which can account for up to 50% of the total between-species variance in Proteobacteria, are a significant concern for studies investigating the factors that shape microbial pangenomes. For comparison, it has been estimated that habitat and phylogeny explain approximately 20% of the between-species variance in pangenome diversity (Maistrenko et al., 2020). Regardless of the method used to discriminate paralogs, genes associated with central cell functions consistently exhibit the lowest rates of gain and loss, while mobile genetic elements and defense systems exhibit the highest rates. Therefore, the inverse association between gene flux and essentiality, as described by previous studies (Iranzo et al., 2019; Puigbò et al., 2014; O. Cohen and Pupko, 2010; Sela et al., 2019), is a robust feature of genome plasticity.

Assessing the contribution of a specific gene clustering method to pangenome variability can only be achieved by comparing it with other methods. However, this is often unfeasible in large datasets, resulting in unnoticed methodological 'noise'. This noise can potentially contribute to unexplained variance or, in the worst-case scenario, act as a confounding factor if methodological biases correlate with the biological variables of interest. To minimize such risks, the selection of tools for gene clustering should be primarily guided by the nature of the research goals, rather than by computational considerations such as runtime and memory usage. For applications that involve tracking vertically transmitted genes, the best gene clusters are those that effectively discriminate among in-paralogs, which is generally achieved by applying synteny criteria. However, those same gene clusters are not optimal to study within-species expansions and contractions of accessory gene families. (For that purpose, gene clusters should keep in-paralogs together, as dictated by classical orthology.)

5.2 Effect of CRISPR and other defense systems on genome fluidity

CRISPR-Cas and other defense systems could potentially have a significant impact on genome evolution by effectively blocking the transfer of MGE, reducing gene flow and limiting the spread of accessory genes. However, because defense systems are often carried by MGE, a positive association between gene flow and defense systems cannot be ruled out *a priori*. We assessed the relative weight of these two opposite scenarios by quantifying the association between defense systems and gene flow in a large number of bacterial species.

Our results build upon previous research that investigated the effects of CRISPR-Cas on microbial evolution and diversification (Wheatley and MacLean, 2021; Shehreen et al., 2019). A large-scale study conducted in 2015 found no evidence to support an overall association between CRISPR-Cas activity and gene acquisition (via HGT) at evolutionary time scales (Gophna et al., 2015). They suggested several reasons for that lack of association:

- CRISPR-Cas systems and CRISPR arrays are mobile, so their presence or absence in a genome does not necessarily indicate their long-term impact.
- Exposure of microorganisms to a large number of MGE can overwhelm the capacity of CRISPR-Cas to control them.
- CRISPR-Cas systems selectively target commonly encountered MGE, such as highly infectious viruses. Therefore, horizontal gene transfer could still be mediated by other non-targeted MGE.

Our analyses support the general conclusion that CRISPR-Cas has no significant effect on overall gene acquisition and loss rates in most bacterial species. Specifically, the difference in gene gain and loss between clades that do and do not harbor CRISPR-Cas is negligible. That said, we identified opposite trends in clades spanning short and medium evolutionary time scales. In particular, positive associations between CRISPR-Cas and HGT rates are more frequent at short time scales (of the order of 5×10^{-4} substitutions per bp in nearly universal core genes), whereas negative associations tend to occur at longer time scales (of the order of 0.002 substitutions per bp). These opposite trends suggest that the actual effects of CRISPR-Cas systems on gene exchange may be obscured by recent co-transfer events involving MGE. As a result, negative effects only become detectable if the CRISPR-Cas system is maintained for a long enough period of time (Figure 4.19).

Beyond that general view, our findings are also consistent with recent research showing that CRISPR-Cas can significantly reduce gene acquisition via MGE in some species, including

Pseudomonas aeruginosa and *Klebsiella pneumoniae* (Wheatley and MacLean, 2021; Botelho et al., 2023). And yet, our study reveals that those species represent special cases rather than the rule, even in the context of host-associated bacteria. In fact, the association between CRISPR-Cas and gene flow is strongly species-dependent and its sign cannot be easily explained by simple ecological, environmental, or genomic variables.

We hypothesize that, in most species, any negative effect of CRISPR-Cas on HGT is masked by the fact that defense systems often travel together with MGE and are more likely retained in scenarios of high exposure to MGE. Indeed, because CRISPR-Cas systems can be quickly gained and lost along a lineage (van Houte et al., 2016; Koonin, Makarova, and Wolf, 2017), it may be difficult to disentangle their effect on gene flow from the causes that lead to their presence or absence, especially at short evolutionary time scales. Moreover, CRISPR-Cas systems play a central role in conflicts between MGE, serving as effective weapons in competitions among viruses and plasmids (Iranzo et al., 2020; Rocha and Bikard, 2022; Haudiquet et al., 2022; Pinilla-Redondo et al., 2022). As a result, it is conceivable that the impact of CRISPR-Cas systems on gene flow varies depending on whether the system is carried by MGE or located in the chromosome.

Another factor that could modulate the effect of CRISPR-Cas on gene exchange is the presence of anti-CRISPR proteins. Anti-CRISPR proteins are often carried by MGE and have the ability to inhibit the CRISPR-Cas system (Davidson et al., 2020; Camara-Wilpert et al., 2023). Indeed, we observed that the prevalence of anti-CRISPR proteins is different for species that display positive association of CRISPR-Cas and MGE than those that display a negative association. It has been hypothesized that anti-CRISPR proteins could facilitate symbiotic interactions between bacteria and relatively benign MGE by down-regulating CRISPR-Cas when it is not needed. Although our results cannot confirm that hypothesis, they suggest that anti-CRISPR proteins could explain high rates of HGT in bacterial genomes that contain CRISPR-Cas systems.

When reproducing the same analyses for other widespread defense systems, we found that their presence is most often positively associated with higher rates of gene gain and higher numbers of MGE and accessory genes. Compared to CRISPR-Cas, other defense systems are more frequently located within or next to MGEs and may have alternative functions related with MGE propagation, which could explain such overall positive association. For example, Abi systems, that were originally described as anti-phage defense mechanisms in prokaryotic genomes, have also been identified in PICIs as accessory genes that facilitate their parasitic lifecycle (Ibarra-Chávez et al., 2021; Figure 4.38). The selective targeting of MGE by some defense systems may be another reason why those systems do not significantly interfere with

HGT. For instance, the GmrSD type IV RM system only targets phages with glucosylated hydroxymethylcytosine (Bair and Black, 2007); Cas9 cannot cleave the DNA of phage T4 due to its heavily modified cytosine residues (Bryson et al., 2015); and the Thoeris defense system provides protection exclusively against phages from the Myoviridae family (Doron et al., 2018). Such specificity of action complicates the identification of general trends in the overall effects of defense systems.

Genomic analyses have shown that defense systems are frequently located in chromosomal regions called defense islands. Defense islands are also enriched with MGE that facilitate their mobilization across genomes (Makarova, Wolf, et al., 2011). It is typical for many bacteria and archaea to encode multiple defense systems of the same type, often within the same defense island. As an extreme example, *Helicobacter pylori* F30 encodes 3 type I RM systems, 11 type II RM systems, 1 type III RM system, and 1 type IV RM system (Oliveira et al., 2014). The coexistence of multiple defense systems in the same genome raises the possibility that they cooperate. Indeed, synergistic anti-phage activity has been demonstrated for defense systems containing sensory switch ATPase domains, such as *tmn* (Wu et al., 2023). Motivated by those observations, we conducted phylogenetically-corrected association tests to investigate the coexistence of multiple defense systems. We found that associations between different classes of defense systems are restricted to a small number of species (Figure 4.32; Figure A.5). In contrast, similar defense systems, such as RM and DMSs, broadly coexist, suggesting that similar systems may act synergistically.

5.3 Phage satellites as vehicles for the transfer of defense systems

Phage satellites can use different strategies to hijack the life cycle of a helper phage (Boyd et al., 2024). Many satellites have developed mechanisms for capsid remodeling to fit their compact genomes while excluding the larger genomes of their helper phages (Penadés and Christie, 2015; Moura de Sousa and Rocha, 2022; O'Hara et al., 2017). Phage-inducible chromosomal islands (PICIs) are the most common type of satellites (Fillol-Salom et al., 2018; Martínez-Rubio et al., 2017). In this study, we identified and categorized four primary types of phage satellites using protein language alignment and community detection methods. We identified genes associated with defense systems, virulence factors, and antibiotic resistance that are carried and disseminated by phage satellites.

The study identified several defense systems located within PICIs or on their adjacent regions. These systems include Abi, CRISPR-Cas, RM, DMS, and DRT (Figure 4.38). Recent research

on marine *Vibrionaceae*, also known as *vibrios*, has shown that most phage defense genes are encoded on MGEs (Hussain et al., 2021). These MGEs are rapidly gained and lost, which distinguishes isolates that are otherwise genomically identical (Hussain et al., 2021). Previous research had found that *E. coli* P2-like phages and their parasitic P4-like satellites harbor various anti-phage systems (Rousset et al., 2022). By analyzing a larger dataset, we confirmed those results and singled out P4-like satellites as the satellites harboring the greatest number of defense systems (Figure 4.38). According to (Rousset et al., 2022), anti-phage defense systems could transform the parasitic satellite-phage relationship into a mutualistic one in the presence of a shared competitor.

Furthermore, we identified anti-defense proteins that are closely associated with PICIs and SaPIs. These proteins include anti-Pycsar (pyrimidine cyclase system for antiphage resistance), anti-RecBCD, anti-RM, and anti-CRISPR proteins (Figure 4.40). Among anti-CRISPR proteins, anti-AcrIIA15 is specific to *Staphylococcus* (Watters et al., 2020), while anti-AcrIIA7 is one of the most widespread proteins in bacteria (Uribe et al., 2019). However, the role of these anti-defense proteins in phage-satellite-host dynamics remains poorly understood. Additional research is necessary to evaluate the effect of different anti-defense proteins on gene flow and their impact on microbial evolvability.

Several studies have shown that marine bacteria can transfer various chromosomal islands, some of which may function as satellites. Recently, new families of satellites has been identified, including Phage-Inducible Chromosomal Minimalist Islands (PICMIs) (Barcia-Cruz et al., 2024), Tycheposons (a family of DNA transposons) (Hackl et al., 2023) and Virion Encapsidated Integrative Mobile Element (VEIME) (Eppley et al., 2022). Unlike typical satellites, PICMIs do not contain genes for capsid remodeling, package their DNA in a concatemeric form (Barcia-Cruz et al., 2024). Also, PICMIs depend on virulent phage particles for transmission to other bacteria and offer protection to their hosts against competitive phages without interfering with their helper phage (Barcia-Cruz et al., 2024). Despite their structural and functional interest, we excluded these entities from our initial analyses to simplify the computational workflow for satellite detection and classification. Expanding our study to these elements constitutes a major goal for future work.

A recent review (Horne et al., 2023) suggests that PICIs ultimately disrupt phage replication and promote HGT. In agreement with that, (Ibarra-Chávez et al., 2022) observed an increase in phage-mediated transduction associated with PICIs. However, this enhancement was not directly attributed to PICIs increasing HGT rates. Rather, the acquisition of PICIs provides protection to gene recipients against phage lysis, promoting the survival of bacterial strains and the maintenance of genetic diversity. These findings indicate that PICIs may positively

contribute to anti-phage defense and illustrate how that could lead to a positive correlation between defense systems and gene acquisition, as we empirically observed for a large number of species (Figure 4.22; Figure 4.28).

Chapter 6

Conclusions

1) The reusability and meta-analysis of pangenome datasets are currently hindered by the incompatibility of operational gene clusters obtained by different methods. To address this limitation and foster future research, a consensus should be reached on a set of methods that cover the most relevant criteria for paralog discrimination.

2) The selection of appropriate gene clustering criteria is key for an unbiased analysis of pangenomes, particularly in light of the role of HGT in microbial evolution.

3) The impact of CRISPR-Cas systems on genome composition varies among species, with major effects observed on genes belonging to mobile genetic elements.

4) CRISPR affects genomic plasticity by curtailing gene gain on relatively short evolutionary timescales.

5) Defense systems display different forms of association with genomic plasticity, with the sign and magnitude of the association varying across species. Compared to CRISPR, other defense systems tend to have a stronger positive correlation with the size of the mobilome, underscoring the frequent joint transfer of defense systems and mobile genetic elements.

6) Phage satellites act as hotspots and possible vehicles for the transfer of defense systems.

References

Ahrens, J. B., Teufel, A. I., & Siltberg-Liberles, J. (2020). A phylogenetic rate parameter indicates different sequence divergence patterns in orthologs and paralogs. *Journal of Molecular Evolution*, *88*, 720–730.

Akhter, S., Aziz, R. K., & Edwards, R. A. (2012). Phispy: A novel algorithm for finding prophages in bacterial genomes that combines similarity-and composition-based strategies. *Nucleic acids research*, *40*(16), e126–e126.

Alcock, B. P., Huynh, W., Chalil, R., Smith, K. W., Raphenya, A. R., Wlodarski, M. A., Edalatmand, A., Petkau, A., Syed, S. A., Tsang, K. K., et al. (2023). Card 2023: Expanded curation, support for machine learning, and resistome prediction at the comprehensive antibiotic resistance database. *Nucleic acids research*, *51*(D1), D690–D699.

Altenhoff, A. M., & Dessimoz, C. (2012). Inferring orthology and paralogy. *Evolutionary Genomics: Statistical and Computational Methods, Volume 1*, 259–279.

Anantharaman, V., Makarova, K. S., Burroughs, A. M., Koonin, E. V., & Aravind, L. (2013). Comprehensive analysis of the hepn superfamily: Identification of novel roles in intragenomic conflicts, defense, pathogenesis and rna processing. *Biology direct*, *8*(1), 1–28.

Bair, C. L., & Black, L. W. (2007). A type iv modification dependent restriction nuclease that targets glucosylated hydroxymethyl cytosine modified dnas. *Journal of molecular biology*, *366*(3), 768–778.

Barcia-Cruz, R., Goudenège, D., Moura de Sousa, J. A., Piel, D., Marbouty, M., Rocha, E. P. C., & Le Roux, F. (2024). Phage-inducible chromosomal minimalist islands (picmis), a novel family of small marine satellites of virulent phages. *Nature Communications*, *15*(1), 664.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., & Horvath, P. (2007). Crispr provides acquired resistance against viruses in prokaryotes. *Science*, *315*(5819), 1709–1712.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*.

Bayliss, S. C., Thorpe, H. A., Coyle, N. M., Sheppard, S. K., & Feil, E. J. (2019). Pirate: A fast and scalable pangenomics toolbox for clustering diverged orthologues in bacteria. *Gigascience*, *8*(10), giz119.

Benler, S., Faure, G., Altae-Tran, H., Shmakov, S., Zhang, F., & Koonin, E. (2021). Cargo genes of tn 7-like transposons comprise an enormous diversity of defense systems, mobile genetic elements, and antibiotic resistance genes. *Mbio*, *12*(6), e02938–21.

Bennett, P. M. (2008). Plasmid encoded antibiotic resistance: Acquisition and transfer of antibiotic resistance genes in bacteria. *British journal of pharmacology*, *153*(1), 347–357.

Bernheim, A., & Sorek, R. (2020). The pan-immune system of bacteria: Antiviral defence as a community resource. *Nature Reviews Microbiology*, *18*(2), 113–119.

Bikard, D., Hatoum-Aslan, A., Mucida, D., & Marraffini, L. A. (2012). Crispr interference can prevent natural transformation and virulence acquisition during in vivo bacterial infection. *Cell host & microbe*, *12*(2), 177–186.

Bin Jang, H., Bolduc, B., Zablocki, O., Kuhn, J. H., Roux, S., Adriaenssens, E. M., Brister, J. R., Kropinski, A. M., Krupovic, M., Lavigne, R., et al. (2019). Taxonomic assignment of uncultivated prokaryotic virus genomes is enabled by gene-sharing networks. *Nature biotechnology*, *37*(6), 632–639.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Bobay, L.-M., & Ochman, H. (2018). Factors driving effective population size and pan-genome evolution in bacteria. *BMC evolutionary biology*, *18*, 1–12.

Bolotin, A., Quinquis, B., Sorokin, A., & Ehrlich, S. D. (2005). Clustered regularly interspaced short palindrome repeats (crisprs) have spacers of extrachromosomal origin. *Microbiology*, *151*(8), 2551–2561.

Bondy-Denomy, J., Pawluk, A., Maxwell, K. L., & Davidson, A. R. (2013). Bacteriophage genes that inactivate the crispr/cas bacterial immune system. *Nature*, *493*(7432), 429–432.

Botelho, J. (2023). Defense systems are pervasive across chromosomally integrated mobile genetic elements and are inversely correlated to virulence and antimicrobial resistance. *Nucleic Acids Research*, *51*(9), 4385–4397.

Botelho, J., Cazares, A., & Schulenburg, H. (2023). The eskape mobilome contributes to the spread of antimicrobial resistance and crispr-mediated conflict between mobile genetic elements. *Nucleic Acids Research*, *51*(1), 236–252.

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T., Schulz, F., Jarett, J., Rivers, A. R., Eloe-Fadrosh, E. A., et al. (2017). Minimum information about a single amplified genome (misag) and a metagenome-assembled genome (mimag) of bacteria and archaea. *Nature biotechnology*, *35*(8), 725–731.

Boyd, C. M., Subramanian, S., Dunham, D. T., Parent, K. N., & Seed, K. D. (2024). A vibrio cholerae viral satellite maximizes its spread and inhibits phage by remodeling hijacked phage coat proteins into small capsids. *Elife*, *12*, RP87611.

Brockhurst, M. A., Harrison, E., Hall, J. P., Richards, T., McNally, A., & MacLean, C. (2019). The ecology and evolution of pangenomes. *Current Biology*, *29*(20), R1094–R1103.

Brouns, S. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J., Snijders, A. P., Dickman, M. J., Makarova, K. S., Koonin, E. V., & Van Der Oost, J. (2008). Small crispr rnas guide antiviral defense in prokaryotes. *Science*, *321*(5891), 960–964.

Bryson, A. L., Hwang, Y., Sherrill-Mix, S., Wu, G. D., Lewis, J. D., Black, L., Clark, T. A., & Bushman, F. D. (2015). Covalent modification of bacteriophage t4 dna inhibits crispr-cas9. *MBio*, *6*(3), 10–1128.

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using diamond. *Nature methods*, *18*(4), 366–368.

Bujnicki, J. M. (2001). Understanding the evolution of restriction-modification systems: Clues from sequence and structure comparisons. *Acta Biochimica Polonica*, *48*(4), 935–967.

Burrus, V., & Waldor, M. K. (2004). Shaping bacterial genomes with integrative and conjugative elements. *Research in microbiology*, *155*(5), 376–386.

Camara-Wilpert, S., Mayo-Muñoz, D., Russel, J., Fagerlund, R. D., Madsen, J. S., Fineran, P. C., Sørensen, S. J., & Pinilla-Redondo, R. (2023). Bacteriophages suppress crispr–cas immunity using rna-based anti-crisprs. *Nature*, *623*(7987), 601–607.

Camargo, A. P., Roux, S., Schulz, F., Babinski, M., Xu, Y., Hu, B., Chain, P. S., Nayfach, S., & Kyrpides, N. C. (2023). Identification of mobile genetic elements with genomad. *Nature Biotechnology*, 1–10.

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). Eggnog-mapper v2: Functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular biology and evolution*, *38*(12), 5825–5829.

Casjens, S. R., Mongodin, E. F., Qiu, W.-G., Luft, B. J., Schutzer, S. E., Gilcrease, E. B., Huang, W. M., Vujadinovic, M., Aron, J. K., Vargas, L. C., et al. (2012). Genome

stability of lyme disease spirochetes: Comparative genomics of borrelia burgdorferi plasmids. *PloS one*, *7*(3), e33280.

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 783–791.

Chopin, M.-C., Chopin, A., & Bidnenko, E. (2005). Phage abortive infection in lactococci: Variations on a theme. *Current opinion in microbiology*, *8*(4), 473–479.

Cohen, D., Melamed, S., Millman, A., Shulman, G., Oppenheimer-Shaanan, Y., Kacen, A., Doron, S., Amitai, G., & Sorek, R. (2019). Cyclic gmp–amp signalling protects bacteria against viral infection. *Nature*, *574*(7780), 691–695.

Cohen, O., Ashkenazy, H., Belinky, F., Huchon, D., & Pupko, T. (2010). Gloome: Gain loss mapping engine. *Bioinformatics*, *26*(22), 2914–2915.

Cohen, O., & Pupko, T. (2010). Inference and characterization of horizontally transferred gene families using stochastic mapping. *Molecular biology and evolution*, *27*(3), 703–713.

Collins, R. E., & Higgs, P. G. (2012). Testing the infinitely many genes model for the evolution of the bacterial core genome and pangenome. *Molecular biology and evolution*, *29*(11), 3413–3425.

Costa, A. R., van den Berg, D. F., Esser, J. Q., Muralidharan, A., van den Bossche, H., Bonilla, B. E., van der Steen, B. A., Haagsma, A. C., Fluit, A. C., Nobrega, F. L., Haas, P.-J., & Brouns, S. J. (2023). Accumulation of defense systems in phage resistant strains of pseudomonas aeruginosa. *bioRxiv*.

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E. P., Vergnaud, G., Gautheret, D., & Pourcel, C. (2018). Crisprcasfinder, an update of crisrfinder, includes a portable version, enhanced performance and integrates search for cas proteins. *Nucleic acids research*, *46*(W1), W246–W251.

Cummins, E. A., Hall, R. J., McInerney, J. O., & McNally, A. (2022). Prokaryote pangenomes are dynamic entities. *Current Opinion in Microbiology*, *66*, 73–78.

D'agostino, R. B., Belanger, A., & D'Agostino Jr, R. B. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician*, *44*(4), 316–321.

Daubin, V., & Ochman, H. (2004). Start-up entities in the origin of new genes. *Current opinion in genetics & development*, *14*(6), 616–619.

Davidson, A. R., Lu, W.-T., Stanley, S. Y., Wang, J., Mejdani, M., Trost, C. N., Hicks, B. T., Lee, J., & Sontheimer, E. J. (2020). Anti-crisprs: Protein inhibitors of crispr-cas systems. *Annual review of biochemistry*, *89*, 309–332.

Dedrick, R. M., Jacobs-Sera, D., Bustamante, C. A. G., Garlena, R. A., Mavrich, T. N., Pope, W. H., Reyes, J. C. C., Russell, D. A., Adair, T., Alvey, R., et al. (2017).

Prophage-mediated defence against viral attack and viral counter-defence. *Nature microbiology*, *2*(3), 1–13.

de Sousa, J. A. M., Fillol-Salom, A., Penadés, J. R., & Rocha, E. P. (2023). Identification and characterization of thousands of bacteriophage satellites across bacteria. *Nucleic Acids Research*, *51*(6), 2759–2777.

Ding, W., Baumdicker, F., & Neher, R. A. (2018). Panx: Pan-genome analysis and exploration. *Nucleic acids research*, *46*(1), e5–e5.

Dong, C., Hao, G.-F., Hua, H.-L., Liu, S., Labena, A. A., Chai, G., Huang, J., Rao, N., & Guo, F.-B. (2018). Anti-crisprdb: A comprehensive online resource for anti-crispr proteins. *Nucleic acids research*, *46*(D1), D393–D398.

Doron, S., Melamed, S., Ofir, G., Leavitt, A., Lopatina, A., Keren, M., Amitai, G., & Sorek, R. (2018). Systematic discovery of antiphage defense systems in the microbial pangenome. *Science*, *359*(6379), eaar4120.

Dy, R. L., Przybilski, R., Semeijn, K., Salmond, G. P., & Fineran, P. C. (2014). A widespread bacteriophage abortive infection system functions through a type iv toxin–antitoxin mechanism. *Nucleic acids research*, *42*(7), 4590–4605.

Eddy, S. R. (2011). Accelerated profile hmm searches. *PLoS computational biology*, *7*(10), e1002195.

Emms, D. M., & Kelly, S. (2019). Orthofinder: Phylogenetic orthology inference for comparative genomics. *Genome biology*, *20*, 1–14.

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, *30*(7), 1575–1584.

Eppley, J. M., Biller, S. J., Luo, E., Burger, A., & DeLong, E. F. (2022). Marine viral particles reveal an expansive repertoire of phage-parasitizing mobile elements. *Proceedings of the National Academy of Sciences*, *119*(43), e2212722119.

Faure, G., Shmakov, S. A., Yan, W. X., Cheng, D. R., Scott, D. A., Peters, J. E., Makarova, K. S., & Koonin, E. V. (2019). Crispr–cas in mobile genetic elements: Counter-defence and beyond. *Nature Reviews Microbiology*, *17*(8), 513–525.

Fillol-Salom, A., Martínez-Rubio, R., Abdulrahman, R. F., Chen, J., Davies, R., & Penadés, J. R. (2018). Phage-inducible chromosomal islands are ubiquitous within the bacterial universe. *The ISME journal*, *12*(9), 2114–2128.

Fitch, W. M. (2000). Homology: A personal view on some of the problems. *Trends in genetics*, *16*(5), 227–231.

Forterre, P., & Prangishvili, D. (2009). The great billion-year war between ribosome-and capsid-encoding organisms (cells and viruses) as the major source of evolutionary novelties. *Annals of the New York Academy of Sciences*, *1178*(1), 65–77.

Fouts, D. E., Brinkac, L., Beck, E., Inman, J., & Sutton, G. (2012). Panoct: Automated clustering of orthologs using conserved gene neighborhood for pan-genomic analysis of bacterial strains and closely related species. *Nucleic acids research*, *40*(22), e172–e172.

Fraser, C. M., Casjens, S., Huang, W. M., Sutton, G. G., Clayton, R., Lathigra, R., White, O., Ketchum, K. A., Dodson, R., Hickey, E. K., et al. (1997). Genomic sequence of a lyme disease spirochaete, borrelia burgdorferi. *Nature*, *390*(6660), 580–586.

Frost, L. S., Leplae, R., Summers, A. O., & Toussaint, A. (2005). Mobile genetic elements: The agents of open source evolution. *Nature Reviews Microbiology*, *3*(9), 722–732.

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). Cd-hit: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152.

Gabaldón, T., & Koonin, E. V. (2013). Functional and evolutionary implications of gene orthology. *Nature Reviews Genetics*, *14*(5), 360–366.

Galperin, M. Y., Kristensen, D. M., Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2019). Microbial genome analysis: The cog approach. *Briefings in bioinformatics*, *20*(4), 1063–1070.

Galperin, M. Y., Wolf, Y. I., Makarova, K. S., Vera Alvarez, R., Landsman, D., & Koonin, E. V. (2021). Cog database update: Focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic acids research*, *49*(D1), D274–D281.

Gao, L., Altae-Tran, H., Böhning, F., Makarova, K. S., Segel, M., Schmid-Burgk, J. L., Koob, J., Wolf, Y. I., Koonin, E. V., & Zhang, F. (2020). Diverse enzymatic activities mediate antiviral immunity in prokaryotes. *Science*, *369*(6507), 1077–1084.

Gao, N. L., Zhang, C., Zhang, Z., Hu, S., Lercher, M. J., Zhao, X.-M., Bork, P., Liu, Z., & Chen, W.-H. (2018). Mvp: A microbe–phage interaction database. *Nucleic acids research*, *46*(D1), D700–D707.

Garcillán-Barcia, M. P., & de la Cruz, F. (2008). Why is entry exclusion an essential feature of conjugative plasmids? *Plasmid*, *60*(1), 1–18.

Gautreau, G., Bazin, A., Gachet, M., Planel, R., Burlot, L., Dubois, M., Perrin, A., Médigue, C., Calteau, A., Cruveiller, S., et al. (2020). Ppanggolin: Depicting microbial diversity via a partitioned pangenome graph. *PLoS computational biology*, *16*(3), e1007732.

Gogarten, J. P., & Townsend, J. P. (2005). Horizontal gene transfer, genome innovation and evolution. *Nature Reviews Microbiology*, *3*(9), 679–687.

Goldfarb, T., Sberro, H., Weinstock, E., Cohen, O., Doron, S., Charpak-Amikam, Y., Afik, S., Ofir, G., & Sorek, R. (2015). Brex is a novel phage resistance system widespread in microbial genomes. *The EMBO journal*, *34*(2), 169–183.

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

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., & Tiedje, J. M. (2007). Dna–dna hybridization values and their relationship to whole-genome sequence similarities. *International journal of systematic and evolutionary microbiology*, *57*(1), 81–91.

Guédon, G., Libante, V., Coluzzi, C., Payot, S., & Leblond-Bourget, N. (2017). The obscure world of integrative and mobilizable elements, highly widespread elements that pirate bacterial conjugative systems. *Genes*, *8*(11), 337.

Hackl, T., Laurenceau, R., Ankenbrand, M. J., Bliem, C., Cariani, Z., Thomas, E., Dooley, K. D., Arellano, A. A., Hogle, S. L., Berube, P., et al. (2023). Novel integrative elements and genomic plasticity in ocean ecosystems. *Cell*, *186*(1), 47–62.

Haft, D. H., Selengut, J., Mongodin, E. F., & Nelson, K. E. (2005). A guild of 45 crispr-associated (cas) protein families and multiple crispr/cas subtypes exist in prokaryotic genomes. *PLoS computational biology*, *1*(6), e60.

Hao, W., & Golding, G. B. (2006). The fate of laterally transferred genes: Life in the fast lane to adaptation or death. *Genome research*, *16*(5), 636–643.

Haudiquet, M., de Sousa, J. M., Touchon, M., & Rocha, E. P. (2022). Selfish, promiscuous and sometimes useful: How mobile genetic elements drive horizontal gene transfer in microbial populations. *Philosophical Transactions of the Royal Society B*, *377*(1861), 20210234.

Horne, T., Orr, V. T., & Hall, J. P. (2023). How do interactions between mobile genetic elements affect horizontal gene transfer? *Current opinion in microbiology*, *73*, 102282.

Huerta-Cepas, J., Serra, F., & Bork, P. (2016). Ete 3: Reconstruction, analysis, and visualization of phylogenomic data. *Molecular biology and evolution*, *33*(6), 1635–1638.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., et al. (2019). Eggnog 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic acids research*, *47*(D1), D309–D314.

Hussain, F. A., Dubert, J., Elsherbini, J., Murphy, M., VanInsberghe, D., Arevalo, P., Kauffman, K., Rodino-Janeiro, B. K., Gavin, H., Gomez, A., et al. (2021). Rapid evolutionary turnover of mobile genetic elements drives bacterial resistance to phages. *Science*, *374*(6566), 488–492.

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*, 1–11.

Hyman, P., & Abedon, S. T. (2010). Bacteriophage host range and bacterial resistance. *Advances in applied microbiology*, *70*, 217–248.

Ibarra-Chávez, R., Brady, A., Chen, J., Penadés, J. R., & Haag, A. F. (2022). Phage-inducible chromosomal islands promote genetic variability by blocking phage reproduction and protecting transductants from phage lysis. *PLoS Genetics*, *18*(3), e1010146.

Ibarra-Chávez, R., Hansen, M. F., Pinilla-Redondo, R., Seed, K. D., & Trivedi, U. (2021). Phage satellites and their emerging applications in biotechnology. *FEMS Microbiology Reviews*, *45*(6), fuab031.

Iranzo, J., Wolf, Y., Koonin, E., & Sela, I. (2019). Gene gain and loss push prokaryotes beyond the homologous recombination barrier and accelerate genome sequence divergence. nat commun 10: 5376.

Iranzo, J., Cuesta, J. A., Manrubia, S., Katsnelson, M. I., & Koonin, E. V. (2017). Disentangling the effects of selection and loss bias on gene dynamics. *Proceedings of the National Academy of Sciences*, *114*(28), E5616–E5624.

Iranzo, J., Faure, G., Wolf, Y. I., & Koonin, E. V. (2020). Game-theoretical modeling of interviral conflicts mediated by mini-crispr arrays. *Frontiers in Microbiology*, *11*, 381.

Iranzo, J., & Koonin, E. V. (2018). How genetic parasites persist despite the purge of natural selection (a). *Europhysics Letters*, *122*(5), 58001.

Iranzo, J., Krupovic, M., & Koonin, E. V. (2016). The double-stranded dna virosphere as a modular hierarchical network of gene sharing. *MBio*, *7*(4), 10–1128.

Iranzo, J., Lobkovsky, A. E., Wolf, Y. I., & Koonin, E. V. (2013). Evolutionary dynamics of the prokaryotic adaptive immunity system crispr-cas in an explicit ecological context. *Journal of bacteriology*, *195*(17), 3834–3844.

Iranzo, J., Puigbò, P., Lobkovsky, A. E., Wolf, Y. I., & Koonin, E. V. (2016). Inevitability of genetic parasites. *Genome biology and evolution*, *8*(9), 2856–2869.

Ishikawa, S. A., Zhukova, A., Iwasaki, W., & Gascuel, O. (2019). A fast likelihood method to reconstruct and visualize ancestral scenarios. *Molecular biology and evolution*, *36*(9), 2069–2085.

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. *Journal of bacteriology*, *169*(12), 5429–5433.

Ives, A. R., & Garland Jr, T. (2010). Phylogenetic logistic regression for binary dependent variables. *Systematic biology*, *59*(1), 9–26.

Ives, A. R., & Helmus, M. R. (2011). Generalized linear mixed models for phylogenetic analyses of community structure. *Ecological Monographs*, *81*(3), 511–525.

Jain, A., & Srivastava, P. (2013). Broad host range plasmids. *FEMS microbiology letters*, *348*(2), 87–96.

Jansen, R., Embden, J. D. v., Gaastra, W., & Schouls, L. M. (2002). Identification of genes that are associated with dna repeats in prokaryotes. *Molecular microbiology*, *43*(6), 1565–1575.

Katoh, K., & Standley, D. M. (2013). Mafft multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular biology and evolution*, *30*(4), 772–780.

Kislyuk, A. O., Haegeman, B., Bergman, N. H., & Weitz, J. S. (2011). Genomic fluidity: An integrative view of gene diversity within microbial populations. *BMC genomics*, *12*, 1–10.

Konstantinidis, K. T., & Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences*, *102*(7), 2567–2572.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, *39*, 309–338.

Koonin, E. V. (2016). Viruses and mobile elements as drivers of evolutionary transitions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1701), 20150442.

Koonin, E. V. (2019). Crispr: A new principle of genome engineering linked to conceptual shifts in evolutionary biology. *Biology & Philosophy*, *34*(1), 9.

Koonin, E. V. (2023). Antitoxins within toxins: A new theme in bacterial antivirus defense. *Proceedings of the National Academy of Sciences*, *120*(31), e2311001120.

Koonin, E. V., & Dolja, V. V. (2013). A virocentric perspective on the evolution of life. *Current opinion in virology*, *3*(5), 546–557.

Koonin, E. V., & Krupovic, M. (2015). Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nature Reviews Genetics*, *16*(3), 184–192.

Koonin, E. V., & Makarova, K. S. (2022). Evolutionary plasticity and functional versatility of crispr systems. *PLoS Biology*, *20*(1), e3001481.

Koonin, E. V., Makarova, K. S., & Wolf, Y. I. (2017). Evolutionary genomics of defense systems in archaea and bacteria. *Annual review of microbiology*, *71*, 233–261.

Koonin, E. V., Makarova, K. S., Wolf, Y. I., & Krupovic, M. (2020). Evolutionary entanglement of mobile genetic elements and host defence systems: Guns for hire. *Nature Reviews Genetics*, *21*(2), 119–131.

Koonin, E. V., Makarova, K. S., & Zhang, F. (2017). Diversity, classification and evolution of crispr-cas systems. *Current opinion in microbiology*, *37*, 67–78.

Kunitski, M., Eicke, N., Huber, P., Köhler, J., Zeller, S., Voigtsberger, J., Schlott, N., Henrichs, K., Sann, H., Trinter, F., et al. (2019). Double-slit photoelectron interference in strong-field ionization of the neon dimer. *Nature communications*, *10*(1), 1.

Kupczok, A., Landan, G., & Dagan, T. (2015). The contribution of genetic recombination to crispr array evolution. *Genome biology and evolution*, *7*(7), 1925–1939.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). Lmertest package: Tests in linear mixed effects models. *Journal of statistical software*, *82*(13).

Lancichinetti, A., Radicchi, F., Ramasco, J. J., & Fortunato, S. (2011). Finding statistically significant communities in networks. *PloS one*, *6*(4), e18961.

Li, D., Dinnage, R., Nell, L. A., Helmus, M. R., & Ives, A. R. (2020). Phyr: An r package for phylogenetic species-distribution modelling in ecological communities. *Methods in Ecology and Evolution*, *11*(11), 1455–1463.

Liao, J., Guo, X., Weller, D. L., Pollak, S., Buckley, D. H., Wiedmann, M., & Cordero, O. X. (2021). Nationwide genomic atlas of soil-dwelling listeria reveals effects of selection and population ecology on pangenome evolution. *Nature Microbiology*, *6*(8), 1021–1030.

Liu, B., Zheng, D., Zhou, S., Chen, L., & Yang, J. (2022). Vfdb 2022: A general classification scheme for bacterial virulence factors. *Nucleic acids research*, *50*(D1), D912–D917.

Liu, M., Li, X., Xie, Y., Bi, D., Sun, J., Li, J., Tai, C., Deng, Z., & Ou, H.-Y. (2019). Iceberg 2.0: An updated database of bacterial integrative and conjugative elements. *Nucleic acids research*, *47*(D1), D660–D665.

Liu, Y., Ahator, S. D., Wang, H., Feng, Q., Xu, Y., Li, C., Zhou, X., & Zhang, L.-H. (2022). Microevolution of the mext and lasr reinforces the bias of quorum sensing system in laboratory strains of pseudomonas aeruginosa pao1. *Frontiers in microbiology*, *13*, 821895.

Lopatina, A., Tal, N., & Sorek, R. (2020). Abortive infection: Bacterial suicide as an antiviral immune strategy. *Annual review of virology*, *7*, 371–384.

Lopatkin, A. J., Huang, S., Smith, R. P., Srimani, J. K., Sysoeva, T. A., Bewick, S., Karig, D. K., & You, L. (2016). Antibiotics as a selective driver for conjugation dynamics. *Nature microbiology*, *1*(6), 1–8.

Lu, M.-J., & Henning, U. (1994). Superinfection exclusion by t-even-type coliphages. *Trends in microbiology*, *2*(4), 137–139.

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R., Potter, S. C., Finn, R. D., et al. (2019). The embl-ebi search and sequence analysis tools apis in 2019. *Nucleic acids research*, *47*(W1), W636–W641.

Mahendra, C., Christie, K. A., Osuna, B. A., Pinilla-Redondo, R., Kleinstiver, B. P., & Bondy-Denomy, J. (2020). Broad-spectrum anti-crispr proteins facilitate horizontal gene transfer. *Nature microbiology*, *5*(4), 620–629.

Maistrenko, O. M., Mende, D. R., Luetge, M., Hildebrand, F., Schmidt, T. S., Li, S. S., Rodrigues, J. F. M., von Mering, C., Pedro Coelho, L., Huerta-Cepas, J., et al. (2020). Disentangling the impact of environmental and phylogenetic constraints on prokaryotic within-species diversity. *The ISME journal*, *14*(5), 1247–1259.

Makarova, K. S., Aravind, L., Grishin, N. V., Rogozin, I. B., & Koonin, E. V. (2002). A dna repair system specific for thermophilic archaea and bacteria predicted by genomic context analysis. *Nucleic acids research*, *30*(2), 482–496.

Makarova, K. S., Aravind, L., Wolf, Y. I., & Koonin, E. V. (2011). Unification of cas protein families and a simple scenario for the origin and evolution of crispr-cas systems. *Biology direct*, *6*, 1–27.

Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I., & Koonin, E. V. (2006). A putative rna-interference-based immune system in prokaryotes: Computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic rnai, and hypothetical mechanisms of action. *Biology direct*, *1*(1), 1–26.

Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F. J., Wolf, Y. I., Yakunin, A. F., et al. (2011). Evolution and classification of the crispr–cas systems. *Nature Reviews Microbiology*, *9*(6), 467–477.

Makarova, K. S., & Koonin, E. V. (2015). Annotation and classification of crispr-cas systems. *CRISPR: methods and protocols*, 47–75.

Makarova, K. S., Wolf, Y. I., Alkhnbashi, O. S., Costa, F., Shah, S. A., Saunders, S. J., Barrangou, R., Brouns, S. J., Charpentier, E., Haft, D. H., et al. (2015). An updated evolutionary classification of crispr–cas systems. *Nature Reviews Microbiology*, *13*(11), 722–736.

Makarova, K. S., Wolf, Y. I., Iranzo, J., Shmakov, S. A., Alkhnbashi, O. S., Brouns, S. J., Charpentier, E., Cheng, D., Haft, D. H., Horvath, P., et al. (2020). Evolutionary classification of crispr–cas systems: A burst of class 2 and derived variants. *Nature Reviews Microbiology*, *18*(2), 67–83.

Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2009). Comprehensive comparative-genomic analysis of type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes. *Biology direct*, *4*, 1–38.

Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2013). Comparative genomics of defense systems in archaea and bacteria. *Nucleic acids research*, *41*(8), 4360–4377.

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. *Journal of bacteriology*, *193*(21), 6039–6056.

Makarova, K. S., Wolf, Y. I., van der Oost, J., & Koonin, E. V. (2009). Prokaryotic homologs of argonaute proteins are predicted to function as key components of a novel system of defense against mobile genetic elements. *Biology direct*, *4*(1), 1–15.

Manzano-Morales, S., Liu, Y., González-Bodí, S., Huerta-Cepas, J., & Iranzo, J. (2023). Comparison of gene clustering criteria reveals intrinsic uncertainty in pangenome analyses. *Genome Biology*, *24*(1), 250.

Marino, N. D., Pinilla-Redondo, R., Csörgő, B., & Bondy-Denomy, J. (2020). Anti-crispr protein applications: Natural brakes for crispr-cas technologies. *Nature methods*, *17*(5), 471–479.

Marraffini, L. A., & Sontheimer, E. J. (2008). Crispr interference limits horizontal gene transfer in staphylococci by targeting dna. *science*, *322*(5909), 1843–1845.

Martínez-Rubio, R., Quiles-Puchalt, N., Martí, M., Humphrey, S., Ram, G., Smyth, D., Chen, J., Novick, R. P., & Penadés, J. R. (2017). Phage-inducible islands in the gram-positive cocci. *The ISME journal*, *11*(4), 1029–1042.

Mayo-Muñoz, D., Pinilla-Redondo, R., Birkholz, N., & Fineran, P. C. (2023). A host of armor: Prokaryotic immune strategies against mobile genetic elements. *Cell Reports*, *42*(7).

McInerney, J. O., McNally, A., & O'connell, M. J. (2017). Why prokaryotes have pangenomes. *Nature microbiology*, *2*(4), 1–5.

Meaden, S., Biswas, A., Arkhipova, K., Morales, S. E., Dutilh, B. E., Westra, E. R., & Fineran, P. C. (2022). High viral abundance and low diversity are associated with increased crispr-cas prevalence across microbial ecosystems. *Current Biology*, *32*(1), 220–227.

Millman, A., Bernheim, A., Stokar-Avihail, A., Fedorenko, T., Voichek, M., Leavitt, A., Oppenheimer-Shaanan, Y., & Sorek, R. (2020). Bacterial retrons function in anti-phage defense. *Cell*, *183*(6), 1551–1561.

Millman, A., Melamed, S., Amitai, G., & Sorek, R. (2020). Diversity and classification of cyclic-oligonucleotide-based anti-phage signalling systems. *Nature microbiology*, *5*(12), 1608–1615.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L., Tosatto, S. C., Paladin, L., Raj, S., Richardson, L. J., et al. (2021). Pfam: The protein families database in 2021. *Nucleic acids research*, *49*(D1), D412–D419.

Mitrofanov, A., Alkhnbashi, O. S., Shmakov, S. A., Makarova, K. S., Koonin, E. V., & Backofen, R. (2021). Crispridentify: Identification of crispr arrays using machine learning approach. *Nucleic acids research*, *49*(4), e20–e20.

Mojica, F. J., Díez-Villaseñor, C., Soria, E., & Juez, G. (2000). Biological significance of a family of regularly spaced repeats in the genomes of archaea, bacteria and mitochondria. *Molecular microbiology*, *36*(1), 244–246.

Mojica, F. J., Díez-Villaseñor, C., García-Martínez, J., & Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of molecular evolution*, *60*, 174–182.

Mojica, F. J., Juez, G., & Rodriguez-Valera, F. (1993). Transcription at different salinities of haloferax mediterranei sequences adjacent to partially modified psti sites. *Molecular microbiology*, *9*(3), 613–621.

Morris, J. H., Apeltsin, L., Newman, A. M., Baumbach, J., Wittkop, T., Su, G., Bader, G. D., & Ferrin, T. E. (2011). Clustermaker: A multi-algorithm clustering plugin for cytoscape. *BMC bioinformatics*, *12*(1), 1–14.

Moura de Sousa, J. A., & Rocha, E. P. (2022). To catch a hijacker: Abundance, evolution and genetic diversity of p4-like bacteriophage satellites. *Philosophical Transactions of the Royal Society B*, *377*(1842), 20200475.

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). Iq-tree: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, *32*(1), 268–274.

Novick, R. P., Christie, G. E., & Penadés, J. R. (2010). The phage-related chromosomal islands of gram-positive bacteria. *Nature Reviews Microbiology*, *8*(8), 541–551.

Ofir, G., Melamed, S., Sberro, H., Mukamel, Z., Silverman, S., Yaakov, G., Doron, S., & Sorek, R. (2018). Disarm is a widespread bacterial defence system with broad anti-phage activities. *Nature microbiology*, *3*(1), 90–98.

O'Hara, B. J., Barth, Z. K., McKitterick, A. C., & Seed, K. D. (2017). A highly specific phage defense system is a conserved feature of the vibrio cholerae mobilome. *PLoS genetics*, *13*(6), e1006838.

Oliveira, P. H., Touchon, M., Cury, J., & Rocha, E. P. (2017). The chromosomal organization of horizontal gene transfer in bacteria. *Nature communications*, *8*(1), 841.

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 research*, *42*(16), 10618–10631.

Ou, Y., & McInerney, J. O. (2022). High frequency of dynamic rearrangements in crispr loci. *bioRxiv*. https://doi.org/10.1101/2022.05.19.492656

Padilha, V. A., Alkhnbashi, O. S., Shah, S. A., de Carvalho, A. C., & Backofen, R. (2020). Crisprcasidentifier: Machine learning for accurate identification and classification of crispr-cas systems. *GigaScience*, *9*(6), giaa062.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, *31*(22), 3691–3693.

Pantolini, L., Studer, G., Pereira, J., Durairaj, J., & Schwede, T. (2022). Embedding-based alignment: Combining protein language models and alignment approaches to detect structural similarities in the twilight-zone. *bioRxiv*, 2022–12.

Parks, D. H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A. J., & Hugenholtz, P. (2020). A complete domain-to-species taxonomy for bacteria and archaea. *Nature biotechnology*, *38*(9), 1079–1086.

Parks, D. H., Chuvochina, M., Rinke, C., Mussig, A. J., Chaumeil, P.-A., & Hugenholtz, P. (2022). Gtdb: An ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic acids research*, *50*(D1), D785–D794.

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature biotechnology*, *36*(10), 996–1004.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). Checkm: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome research*, *25*(7), 1043–1055.

Partridge, S. R., Kwong, S. M., Firth, N., & Jensen, S. O. (2018). Mobile genetic elements associated with antimicrobial resistance. *Clinical microbiology reviews*, *31*(4), 10–1128.

Payne, L. J., Meaden, S., Mestre, M. R., Palmer, C., Toro, N., Fineran, P. C., & Jackson, S. A. (2022). Padloc: A web server for the identification of antiviral defence systems in microbial genomes. *Nucleic acids research*, *50*(W1), W541–W550.

Payne, L. J., Todeschini, T. C., Wu, Y., Perry, B. J., Ronson, C. W., Fineran, P. C., Nobrega, F. L., & Jackson, S. A. (2021). Identification and classification of antiviral defence systems in bacteria and archaea with padloc reveals new system types. *Nucleic Acids Research*, *49*(19), 10868–10878.

Penadés, J. R., & Christie, G. E. (2015). The phage-inducible chromosomal islands: A family of highly evolved molecular parasites. *Annual review of virology*, *2*, 181–201.

Perrin, A., & Rocha, E. P. (2021). Panacota: A modular tool for massive microbial comparative genomics. *NAR genomics and bioinformatics*, *3*(1), lqaa106.

Peters, J. E., Makarova, K. S., Shmakov, S., & Koonin, E. V. (2017). Recruitment of crispr-cas systems by tn7-like transposons. *Proceedings of the National Academy of Sciences*, *114*(35), E7358–E7366.

Petitjean, C., Makarova, K. S., Wolf, Y. I., & Koonin, E. V. (2017). Extreme deviations from expected evolutionary rates in archaeal protein families. *Genome biology and evolution*, *9*(10), 2791–2811.

Picton, D. M., Luyten, Y. A., Morgan, R. D., Nelson, A., Smith, D. L., Dryden, D. T., Hinton, J. C., & Blower, T. R. (2021). The phage defence island of a multidrug resistant plasmid uses both brex and type iv restriction for complementary protection from viruses. *Nucleic Acids Research*, *49*(19), 11257–11273.

Pinilla-Redondo, R., Russel, J., Mayo-Muñoz, D., Shah, S. A., Garrett, R. A., Nesme, J., Madsen, J. S., Fineran, P. C., & Sørensen, S. J. (2022). Crispr-cas systems are widespread accessory elements across bacterial and archaeal plasmids. *Nucleic Acids Research*, *50*(8), 4315–4328.

Pinilla-Redondo, R., Shehreen, S., Marino, N. D., Fagerlund, R. D., Brown, C. M., Sørensen, S. J., Fineran, P. C., & Bondy-Denomy, J. (2020). Discovery of multiple anti-crisprs highlights anti-defense gene clustering in mobile genetic elements. *Nature communications*, *11*(1), 5652.

Pourcel, C., Salvignol, G., & Vergnaud, G. (2005). Crispr elements in yersinia pestis acquire new repeats by preferential uptake of bacteriophage dna, and provide additional tools for evolutionary studies. *Microbiology*, *151*(3), 653–663.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). Fasttree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, *5*(3), e9490.

Puigbò, P., Lobkovsky, A. E., Kristensen, D. M., Wolf, Y. I., & Koonin, E. V. (2014). Genomes in turmoil: Quantification of genome dynamics in prokaryote supergenomes. *BMC biology*, *12*, 1–19.

Pursey, E., Dimitriu, T., Paganelli, F. L., Westra, E. R., & van Houte, S. (2022). Crispr-cas is associated with fewer antibiotic resistance genes in bacterial pathogens. *Philosophical Transactions of the Royal Society B*, *377*(1842), 20200464.

Rocha, E. P., & Bikard, D. (2022). Microbial defenses against mobile genetic elements and viruses: Who defends whom from what? *PLoS biology*, *20*(1), e3001514.

Rodriguez-Valera, F., Martin-Cuadrado, A.-B., Rodriguez-Brito, B., Pasic, L., Thingstad, T. F., Rohwer, F., & Mira, A. (2009). Explaining microbial population genomics through phage predation. *Nature Precedings*, 1–1.

Rousset, F., Depardieu, F., Miele, S., Dowding, J., Laval, A.-L., Lieberman, E., Garry, D., Rocha, E. P., Bernheim, A., & Bikard, D. (2022). Phages and their satellites encode hotspots of antiviral systems. *Cell host & microbe*, *30*(5), 740–753.

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. *The CRISPR journal*, *3*(6), 462–469.

Samson, J. E., Magadán, A. H., Sabri, M., & Moineau, S. (2013). Revenge of the phages: Defeating bacterial defences. *Nature Reviews Microbiology*, *11*(10), 675–687.

Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., & Siksnys, V. (2011). The streptococcus thermophilus crispr/cas system provides immunity in escherichia coli. *Nucleic acids research*, *39*(21), 9275–9282.

Seed, K. D., Lazinski, D. W., Calderwood, S. B., & Camilli, A. (2013). A bacteriophage encodes its own crispr/cas adaptive response to evade host innate immunity. *Nature*, *494*(7438), 489–491.

Sela, I., Wolf, Y. I., & Koonin, E. V. (2016). Theory of prokaryotic genome evolution. *Proceedings of the National Academy of Sciences*, *113*(41), 11399–11407.

Sela, I., Wolf, Y. I., & Koonin, E. V. (2019). Selection and genome plasticity as the key factors in the evolution of bacteria. *Physical Review X*, *9*(3), 031018.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., & Ideker, T. (2003). Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome research*, *13*(11), 2498–2504.

Shapiro, B. J. (2017). The population genetics of pangenomes. *Nature microbiology*, *2*(12), 1574–1574.

Shaw, J., & Yu, Y. W. (2023). Fast and robust metagenomic sequence comparison through sparse chaining with skani. *bioRxiv*, 2023–01.

Shehreen, S., Chyou, T.-y., Fineran, P. C., & Brown, C. M. (2019). Genome-wide correlation analysis suggests different roles of crispr-cas systems in the acquisition of antibiotic resistance genes in diverse species. *Philosophical Transactions of the Royal Society B*, *374*(1772), 20180384.

Shmakov, S. A., Faure, G., Makarova, K. S., Wolf, Y. I., Severinov, K. V., & Koonin, E. V. (2019). Systematic prediction of functionally linked genes in bacterial and archaeal genomes. *Nature protocols*, *14*(10), 3013–3031.

Shmakov, S. A., Makarova, K. S., Wolf, Y. I., Severinov, K. V., & Koonin, E. V. (2018). Systematic prediction of genes functionally linked to crispr-cas systems by gene neighborhood analysis. *Proceedings of the National Academy of Sciences*, *115*(23), E5307–E5316.

Shmakov, S. A., Sitnik, V., Makarova, K. S., Wolf, Y. I., Severinov, K. V., & Koonin, E. V. (2017). The crispr spacer space is dominated by sequences from species-specific mobilomes. *MBio*, *8*(5), 10–1128.

Sitto, F., & Battistuzzi, F. U. (2020). Estimating pangenomes with roary. *Molecular biology and evolution*, *37*(3), 933–939.

Smillie, C., Garcillán-Barcia, M. P., Francia, M. V., Rocha, E. P., & de la Cruz, F. (2010). Mobility of plasmids. *Microbiology and Molecular Biology Reviews*, *74*(3), 434–452.

Snipen, L., & Liland, K. H. (2015). Micropan: An r-package for microbial pan-genomics. *BMC bioinformatics*, *16*, 1–8.

Soucy, S. M., Huang, J., & Gogarten, J. P. (2015). Horizontal gene transfer: Building the web of life. *Nature Reviews Genetics*, *16*(8), 472–482.

Stamatakis, A. (2014). Raxml version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313.

Steinegger, M., & Söding, J. (2017). Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, *35*(11), 1026–1028.

Stevenson, C., Hall, J. P., Harrison, E., Wood, A., & Brockhurst, M. A. (2017). Gene mobility promotes the spread of resistance in bacterial populations. *The ISME journal*, *11*(8), 1930–1932.

Suyama, M., Torrents, D., & Bork, P. (2006). Pal2nal: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research*, *34*(suppl_2), W609–W612.

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

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., et al. (2005). Genome analysis of multiple pathogenic isolates of streptococcus agalactiae: Implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, *102*(39), 13950–13955.

Thomas, C. M., & Nielsen, K. M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nature reviews microbiology*, *3*(9), 711–721.

Thompson, C. C., Chimetto, L., Edwards, R. A., Swings, J., Stackebrandt, E., & Thompson, F. L. (2013). Microbial genomic taxonomy. *BMC genomics*, *14*(1), 1–8.

Tonkin-Hill, G., MacAlasdair, N., Ruis, C., Weimann, A., Horesh, G., Lees, J. A., Gladstone, R. A., Lo, S., Beaudoin, C., Floto, R. A., et al. (2020). Producing polished prokaryotic pangenomes with the panaroo pipeline. *Genome biology*, *21*, 1–21.

Touchon, M., Hoede, C., Tenaillon, O., Barbe, V., Baeriswyl, S., Bidet, P., Bingen, E., Bonacorsi, S., Bouchier, C., Bouvet, O., et al. (2009). Organised genome dynamics in the escherichia coli species results in highly diverse adaptive paths. *PLoS genetics*, *5*(1), e1000344.

Treangen, T. J., & Rocha, E. P. (2011). Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS genetics*, *7*(1), e1001284.

Uribe, R. V., van der Helm, E., Misiakou, M.-A., Lee, S.-W., Kol, S., & Sommer, M. O. (2019). Discovery and characterization of cas9 inhibitors disseminated across seven bacterial phyla. *Cell host & microbe*, *25*(2), 233–241.

van Houte, S., Buckling, A., & Westra, E. R. (2016). Evolutionary ecology of prokaryotic immune mechanisms. *Microbiology and Molecular Biology Reviews*, *80*(3), 745–763.

Wang, M., Liu, G., Liu, M., Tai, C., Deng, Z., Song, J., & Ou, H.-Y. (2023). Iceberg 3.0: Functional categorization and analysis of the integrative and conjugative elements in bacteria. *Nucleic Acids Research*, gkad935.

Washburne, A. D., Silverman, J. D., Leff, J. W., Bennett, D. J., Darcy, J. L., Mukherjee, S., Fierer, N., & David, L. A. (2017). Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ*, *5*, e2969.

Watson, B. N., Staals, R. H., & Fineran, P. C. (2018). Crispr-cas-mediated phage resistance enhances horizontal gene transfer by transduction. *MBio*, *9*(1), 10–1128.

Watters, K. E., Shivram, H., Fellmann, C., Lew, R. J., McMahon, B., & Doudna, J. A. (2020). Potent crispr-cas9 inhibitors from staphylococcus genomes. *Proceedings of the National Academy of Sciences*, *117*(12), 6531–6539.

Wayne, L., Brenner, D., Colwell, R., Grimont, P., Kandler, O., Krichevsky, M., Moore, L., Moore, W., Murray, R., Stackebrandt, E., et al. (1987). Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *International Journal of Systematic and Evolutionary Microbiology*, *37*(4), 463–464.

Westra, E. R., & Levin, B. R. (2020). It is unclear how important crispr-cas systems are for protecting natural populations of bacteria against infections by mobile genetic elements. *Proceedings of the National Academy of Sciences*, *117*(45), 27777–27785.

Wheatley, R. M., & MacLean, R. C. (2021). Crispr-cas systems restrict horizontal gene transfer in pseudomonas aeruginosa. *The ISME Journal*, *15*(5), 1420–1433.

Whelan, F. J., Hall, R. J., & McInerney, J. O. (2021). Evidence for selection in the abundant accessory gene content of a prokaryote pangenome. *Molecular biology and evolution*, *38*(9), 3697–3708.

Williams, J. E. (1983). Warning on a new potential for laboratory-acquired infections as a result of the new nomenclature for the plague bacillus. *Japanese Journal of Medical Science and Biology*, *36*(5), 295–297.

Wolf, Y. I., Makarova, K. S., Lobkovsky, A. E., & Koonin, E. V. (2016). Two fundamentally different classes of microbial genes. *Nature microbiology*, *2*(3), 1–6.

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., Millard, A. D., Gençay, Y. E., Koonin, E. V., & Nobrega, F. L. (2023). Synergistic anti-phage activity of bacterial defence systems. *bioRxiv*.

Yan, Y., Zheng, J., Zhang, X., & Yin, Y. (2024). Dbapis: A database of a nti-p rokaryotic i mmune s ystem genes. *Nucleic Acids Research*, *52*(D1), D419–D425.

Zhou, Z., Charlesworth, J., & Achtman, M. (2020). Accurate reconstruction of bacterial pan-and core genomes with peppan. *Genome research*, *30*(11), 1667–1679.

Annexes

Table A.1: Number of species with specific CRISPR types.

CRISPR_type	Negative	No correlation	Positive
II-B	1	1	0
I-C	10	20	16
II-C	8	9	4
I-B	7	10	8
I-F	15	14	17
I-E	11	23	19
II-A	19	16	18
I-G	1	0	0
VI-B	2	0	0
V-A	1	0	0
III-A	2	6	8
IV-A3	1	3	1
III-B	0	2	1
III-D	0	0	1

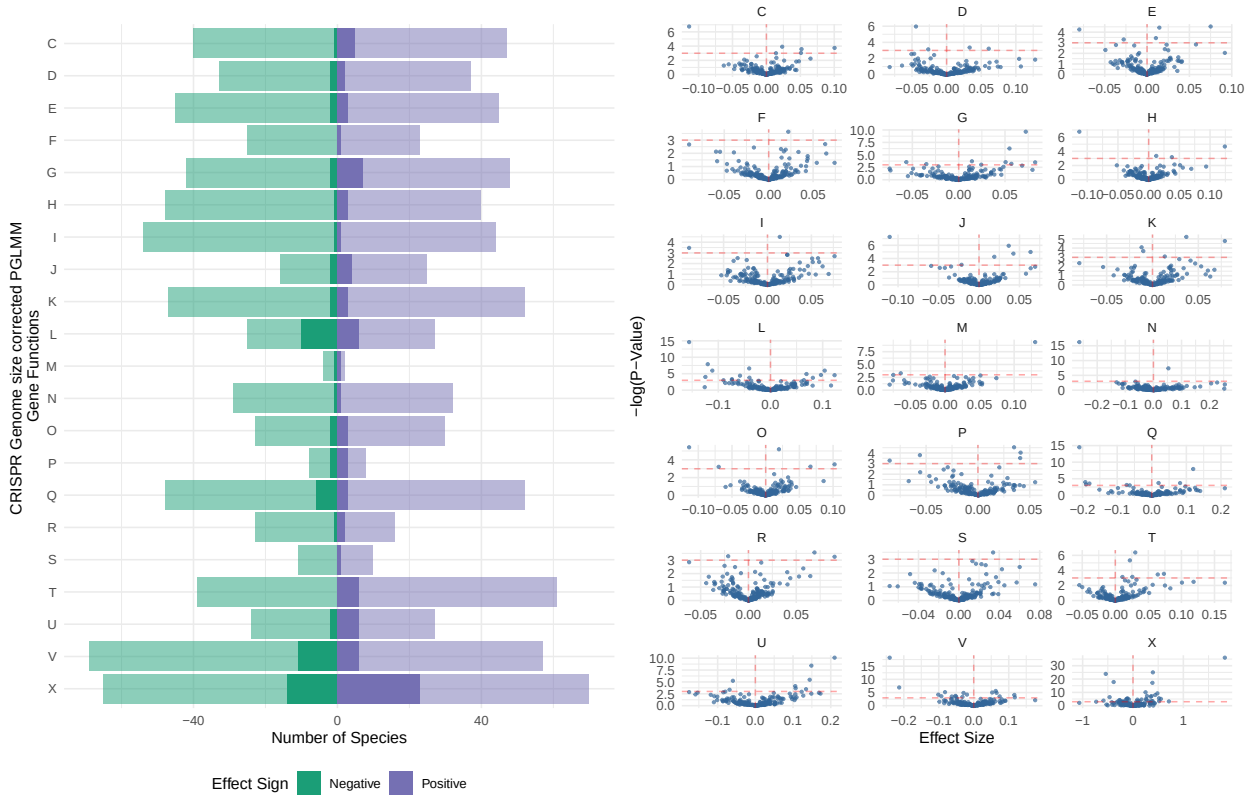

Figure A.1: Correlation between presence of CRISPR-Cas and the number of genes from different functional categories (binomial-distributed PGLMM). The bar plot displays the number of species that showed a significant correlation between the presence of the CRISPR-Cas and the number of functional-specific genes. Solid bars correspond to a $p < 0.05$ significance cutoff. Semitransparent bars correspond to an effect-size cutoff defined by the smallest absolute effect size that has a significant p-value. The scatter plot displays the distribution of p-values at different effect sizes (one point per species).

Figure A.2: The number of MGE gene annotations that remain after masking MGEs in the complete genomes.

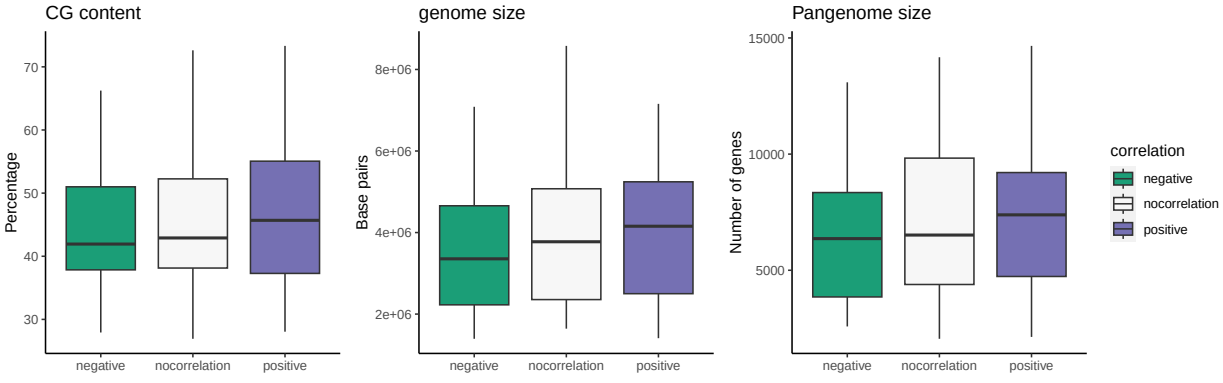

Figure A.3: Difference of GC content, genome size and pangenome size between negative correlated species, positive correlated species, and no correlation species.

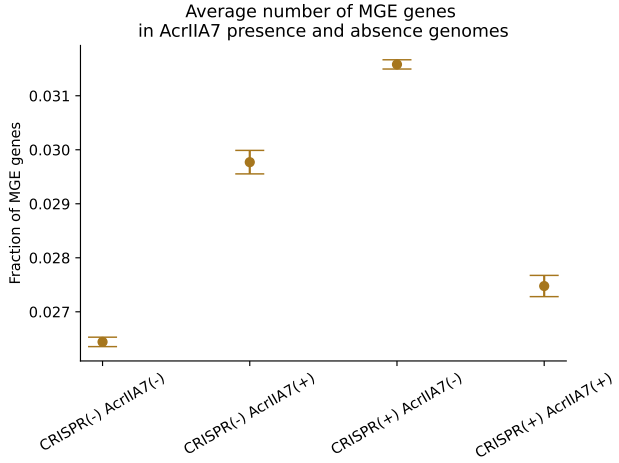

Figure A.4: Comparison of MGE gene abundance on the genomes with and without AcrIIA7 protein in the condition of have or not have a CRISPR-Cas system. Whiskers represent the 95% confidence intervals.

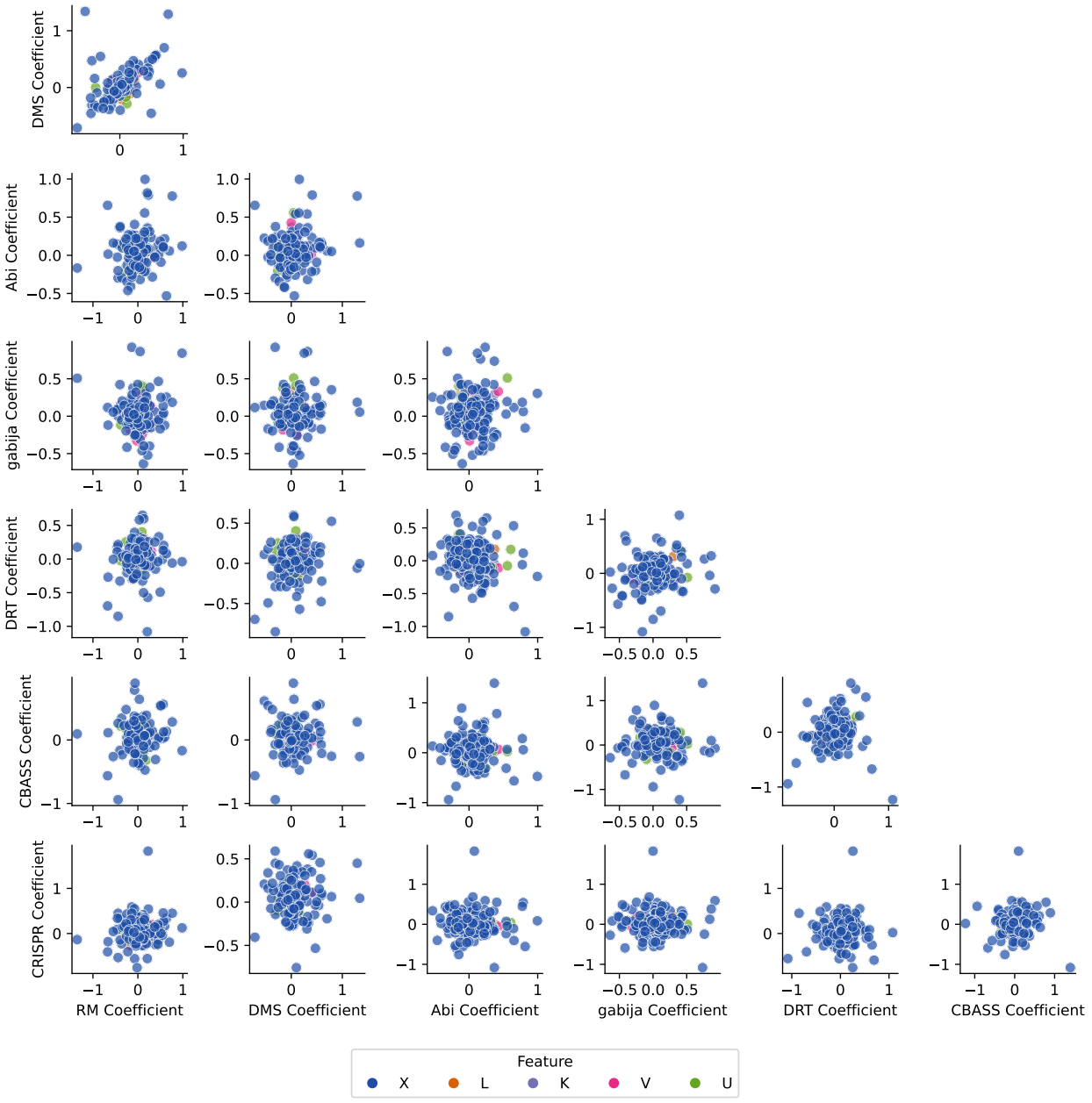

Figure A.5: Comparison of effect size between defense systems. For each defense systems, the effect size of gene functional categories obtained from Poisson-distributed PGLMM.

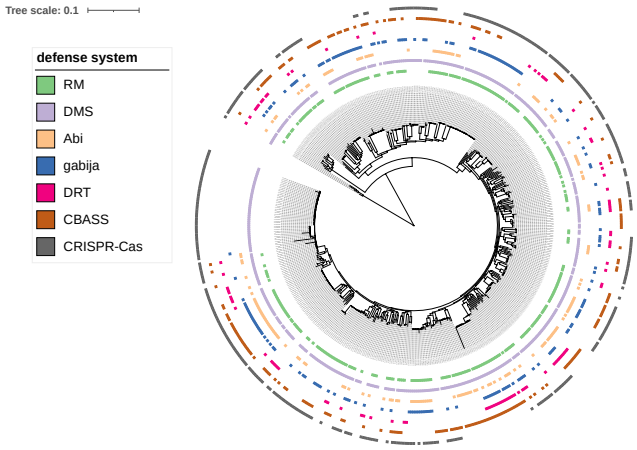

Figure A.6: Presence and absence of defense systems in each genome of *Pseudomonas aeruginosa* along the high quality species tree.

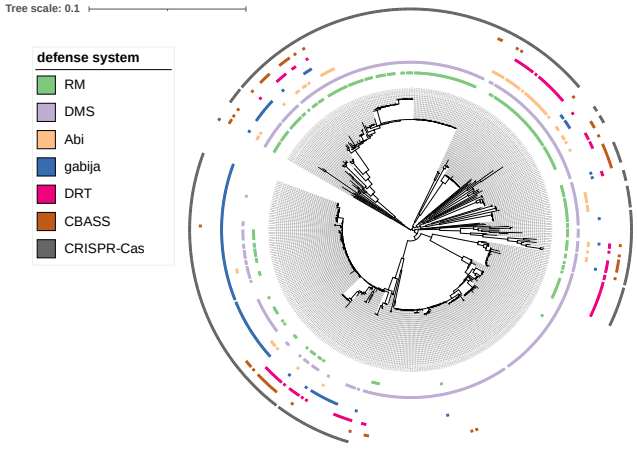

Figure A.7: Presence and absence of defense systems in each genome of *Acinetobacter baumannii* along the high quality species tree.

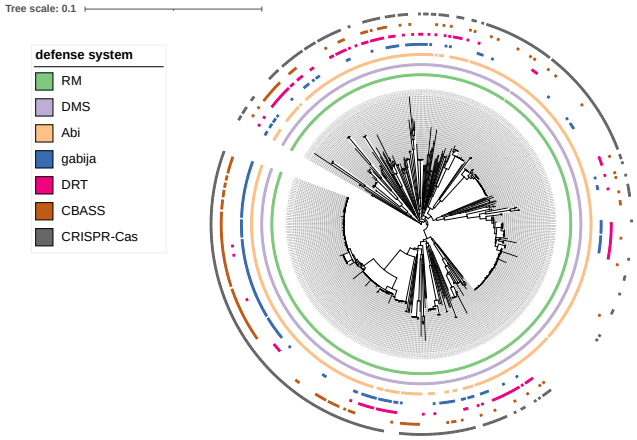

Figure A.8: Presence and absence of defense systems in each genome of *Klebsiella pneumoniae* along the high quality species tree.

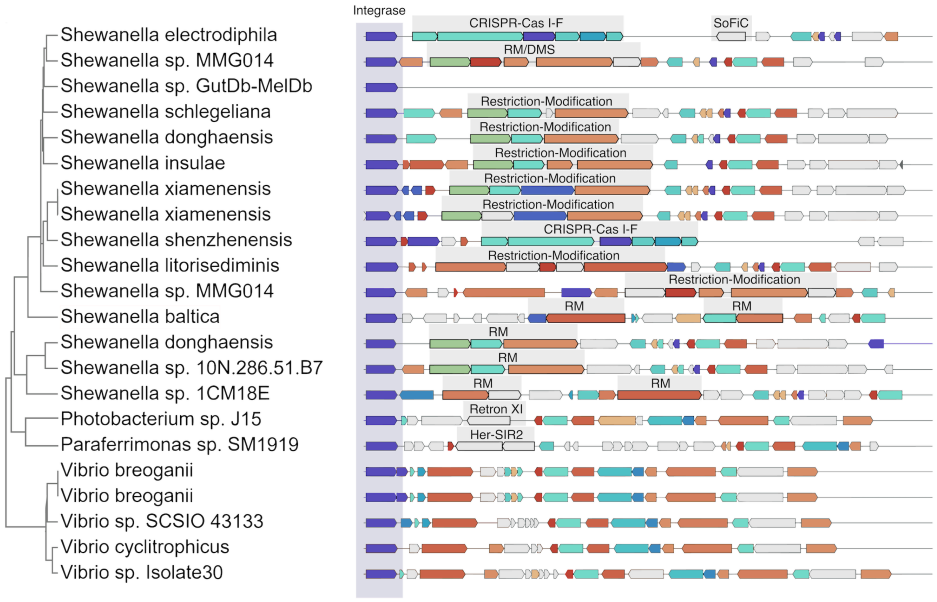

Figure A.9: Highly similar PICIs, found to carry different defense systems, were identified in collaboration with Mario Rodríguez Mestre. The color represents that the proteins belong to the same families.