Network Analysis of Non-treelike Patterns in Evolution

A thesis submitted to The University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Biology, Medicine and Health

2020

Yaqing Ou

School of Biological Sciences

Contents

Total word count: 32,840

List of Figures

List of Tables

List of Abbreviations

alcohol dehydrogenase gene	Adh
anti-CRISPR	*acr*
antibiotic resistance gene	ARG
antimicrobial resistance	AMR
Clustered Regularly Interspaced Short Palindromic Repeat	CRISPR
connected components	CC
CRISPR RNA	crRNA
CRISPR-associated	Cas
CRISPR-associated complex for antiviral defence	Cascade
CRISPR-associated Rossmann fold	CARF
crossover hotspot instigator	Chi
cyclic oligoadenylate	cOA
double-stranded DNA	dsDNA
double-stranded DNA break	DSB
giant connected component	GCC
higher eukaryotes and prokaryotes nucleotide	HEPN
horizontal gene transfer	HGT
integration host factor	IHF
large subunit	LS
lateral gene transfer	LGT
leader anchoring sequence	LAS
mobile genetic elements	MGE
Mycobacterium tuberculosis complex	MTBC
Odd Ratio	OR
open reading frame	ORF
Order divergence event	ODE
precursor CRISPR RNA	pre-crRNA

protospacer adjacent motifs	PAM
repeat-associated mysterious protein	RAMP
reverse transcriptase	RT
Sequence Similarity Network	SSN
single-stranded	ssDNA
single-stranded RNA	ssRNA
small subunits	SS
trans-activating crRNA	tracrRNA

Abstract

Introgressive descents such as recombination, gene fusion and horizontal gene transfer (HGT) cause reticulate patterns in the evolutionary history of prokaryotes and eukaryotes, which are too complex to show in traditional tree-based models. In this thesis, we introduced network-based approaches such as the sequence similarity network (SSN) and explored its potential to investigating large datasets. Two different genetic features were investigated: (1) composite genes that are generated by the remodelling of two unrelated genetic segments; (2) CRISPR-Cas systems that are widely spread in prokaryotes as adaptive immune systems.

First, we employed a network-based approach to explore gene remodelling. Non-homologous genes can form into a single open reading frame (ORF) through gene fusion. The new gene is called a composite gene while the parental genes are called component genes. To investigate the distribution of composite genes across all of life, we constructed SSNs of a large dataset containing more than 1 million genes from prokaryotes, eukaryotes, viruses and plasmids. In our dataset, 18.57% of genes were identified as composite genes, which were pervasively spread across three domains of life as well as all COG functional categories. We also found eukaryotic genes were more likely to be composites than prokaryotic genes.

Second, we investigated the evolution history of the CRISPR-Cas locus. Prokaryotes are engaged in the constant arms race with foreign mobile genetic elements (MGEs). CRISPR-Cas, an important adaptive immune system in Archaea and Bacteria, is involved in diverse evolutionary processes. While under attack, it is thought that a spacer is directly acquired from the segment of the invader and integrated between the leading sequence and the first spacer, so spacers are ordered chronologically corresponding to the infection time. However, through comparative genome analysis, we found that old spacers were located upstream of new spacers, which indicated either the role of ectopic spacer integration or recombination. Further, we found the distribution of CRISPR-Cas is not uniform across prokaryotic phylogeny. To understand why this is the case, we used a co-occurrence approach to identify the association and disassociation between protein-coding genes and CRISPR-Cas systems. We found that genes that co-occurred with CRISPR-Cas are mainly in metabolic pathways and that the distribution of co-occurred genes in the phylogeny is compatible with the distribution of CRISPR-Cas subtypes, which suggested the influence of genetic background on the distribution of CRISPR-Cas systems.

Collectively, network-based approaches have shown great potential in helping identify non-vertical evolutions.

Declaration

I declare that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

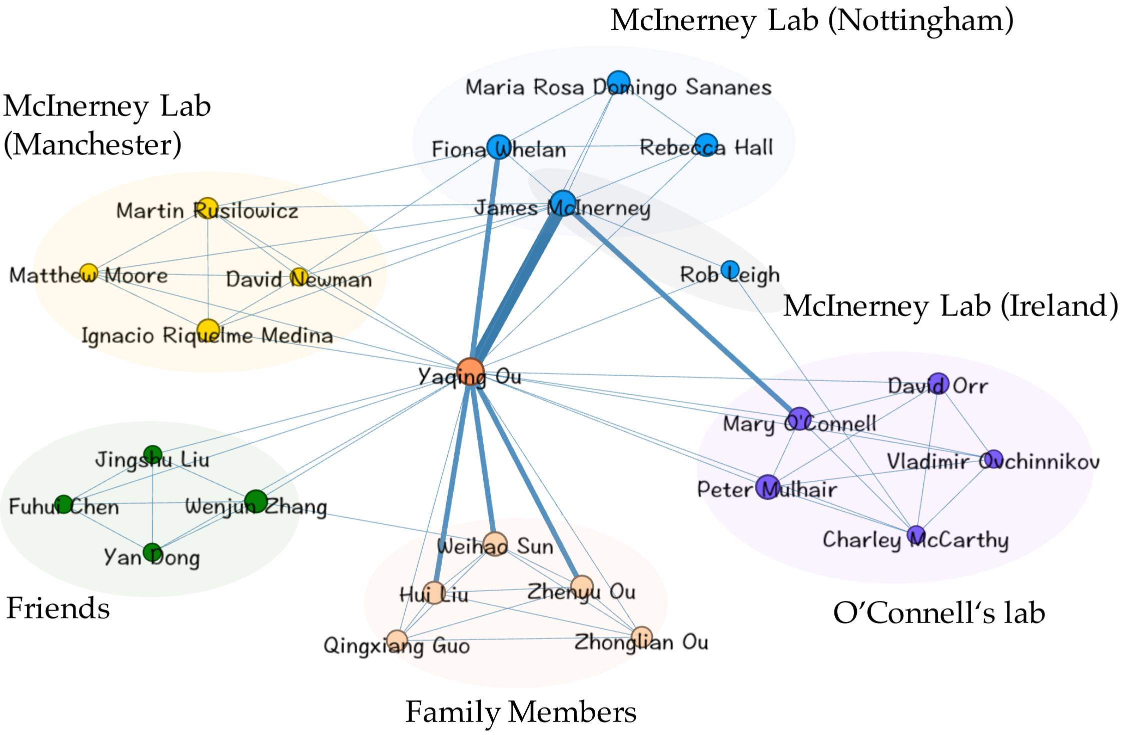
Acknowledgement

I would like to express my deepest and sincerest gratitude towards my supervisor Professor James McInerney for being the supervisor and mentor that any student could ever wish for. Thank you for your guidance, inspiration and support throughout my whole PhD. I could never finish my PhD without your encouragement. Specially, I would like to thank members of McInerney group in Nottingham, Fiona Whelan, Maria Rosa Domingo Sananes and Rebecca Hall for countless ideas inspired in scientific discussions, precious supports during the lockdown and gracious helps in revising my thesis. Along with Peter Mulhair and David Orr from O'Connell group, I would like to thank you all for creating such a positive and pleasant research atmosphere in the lab. I really appreciate all great time we shared together.

In addition, I would like to thank the former members of McInerney group, Martin Rusilowicz, Ignacio Riquelme Medina and Rob Leigh for their patient and generous help at the beginning of my PhD, the time that I did not even know what does *pwd* mean in Unix Command. In addition, I would like to thank The University of Manchester and Chinese Scholarship Council for sponsoring my PhD.

My special and warmest thankfulness goes to my family. Thank you Mum (Hui Liu), for being my best friend and my role model. I could never imagine how my life would be without you. Thank you, Dad (Zhenyu Ou) and Grandma, for being my rocks and supporting me with your love. A special gratitude towards my Grandpa, who passed away on the second year of my PhD, thank you for being my Grandpa. Last but not least, I am extremely grateful for the massive mental (and a little financial) supports from my husband Weihao Sun during the past four years. Thank you for being in my life, completing my life and always appreciating me for just who I am.

I would like to thank my dearest friends Wenjun Zhang, Jingshu Liu and Jiayun Wang for the uncountable happy and meaningful moments we shared. Thank you for bringing sunshine into my life when I am depressed, being happy for my happiness and motivating me when I did not believe in myself.



The network of people that I would like to thank throughout my PhD. This idea is inspired by Jananan Pathmanathan, who is also the developer of programme CompositeSearch.

Rational for submitting the thesis in a journal format

This thesis is approved to be presented in a journal format. The results (Chapters 2, 3 and 4) are in the style of peer-reviewed journal articles. Chapter 2 is in the format of *Genes* journal and has been accepted for publication (https://www.mdpi.com/2073-4425/10/9/648/htm). Chapters 3 and 4 are written in the format of *Genome Biology and Evolution*. References of all chapters are unified for consistency and listed together at the end of this thesis.

Chapter 1.

General Introduction

1.1 Network Thinking

1.1.1 Evolution Is Not Always Tree-like

Evolutionary biologists are keen on depicting the processes that shape diverse lives. Ever since Charles Darwin proposed the use of tree diagrams in order to portray phylogenies (Darwin, 1859), they have become the dominant mechanisms for recognizing and illuminating species relationships. When trying to interpret analysis of vertical genealogical relationships, such as offspring divergent from one common ancestor, the phylogenetic tree approach is critical approach (O'Hara 1997; Bapteste et al. 2012). Scientists such as Darwin believed that species evolved gradually in a slow rhythm. However, tree-like models might not be appropriate in order to fully describe saltational evolutionary events such as gene fusion, hybridization, and horizontal gene transfer (HGT, also known lateral gene transfer (LGT)) or symbiosis (Bapteste et al., 2013; Coleman et al., 2015).

Genetic material that transits between more than one lineage and propagates is called "introgressive descent" (Bapteste et al., 2012), which is an important evolutionary process aside from vertical descent. Gene introgression impacts across all levels of biological organization (Bapteste et al., 2012; Corel et al., 2016). For example, at the level of molecules, composite genes can arise from introgression of genes from different families (Figure 1.1a) (Jachiet et al., 2013). At the level of genomes, composite genome formation can occur through HGT from individual sequences, or indeed entire genomes can merge with other genomes (Figure 1.1b) (Alvarez-Ponce et al., 2013; Hotopp et al., 2007; Nelson-Sathi et al., 2012). At the level of organisms (Figure 1.1c), composite organisms can be formed between endosymbionts and hosts (Andam et al., 2011; Bapteste, 2014). A very important hypothesis presented by Rivera & Lake (2004) called "The ring of life", postulated that the origin of eukaryote genomes came about through a fusion of archaeal and bacterial genomes. This hypothesis broke the traditional bifurcation model and inspired network thinking (Alvarez-Ponce et al., 2013; McInerney et al., 2014).

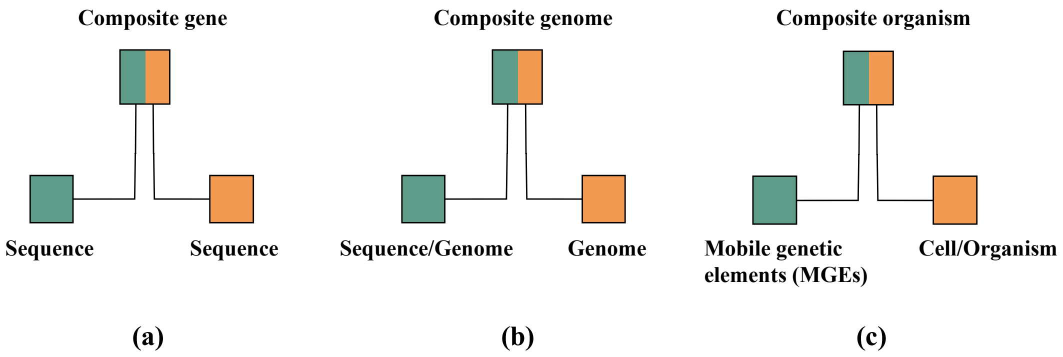


Figure 1.1 Patterns of introgression infiltrating in different levels of biological organisms. (a) Composite gene constructs from fusion of sequences at the molecular level. **(b)** Composite genome arises by introgression of a gene or genome into a genome at the genomic level. **(c)** Composite organism forms from the mosaicism of mobile elements, such as plasmid, and a host cell or organism at the organismal level. Figure adapted from Corel et al. 2016.

1.1.2 Horizontal Gene Transfer (HGT) in Prokaryotes and Eukaryotes

One of the main approaches of interchanging genetic materials is through HGT, which is essential and widespread in the evolution of the three domains of life (Keeling and Palmer, 2008; Koonin et al., 2001; Polz et al., 2013). HGT describes the movement of genetic materials between distant or closely related organisms (Keeling and Palmer, 2008) and firstly been described by Griffith in an infectious experiment of *Streptococcus pneumoniae* and mice (Griffith, 1928).

HGT predominantly proceeds through three mechanisms: transformation, the ability to take in naked DNA from the surrounding environment; transduction, which acquires or moves DNA fragments between bacteria by bacteriophage; conjugation, the transmission of genetic information between species by physical interaction between donor and recipient (Ochman et al., 2000).

Among prokaryotes, HGT exists universally and occupies an important position in gene neo-functionalization (Koonin et al., 2001; Ochman et al., 2000). Through comparison analysis of three *Escherichia coli* genomes, Welch et al. (2002) identified multiple HGT events in discrete gene islands. Moreover, HGT facilitates the acquisition of antibiotic resistance (Kay et al., 2002), virulence attributes (Gemski et al., 1980; Hacker et al., 1997), and metabolic properties as principal traits of Bacteria (Ochman et al., 2000). HGT also plays a vital role in the evolution of Archaea (Williams et al., 2017). A study from Nelson-Sathi et al. (2012) has shown that the origin of Haloarchaea was driven by the acquisition of many functional genes from eubacteria through HGT. This includes genes encoding catabolic and heterotrophic carbon metabolism, membrane transporters, respiratory chain components, and additional cofactor biosynthesis genes. Furthermore, as for the genetic relationships between bacteria and archaea, the transfers are quite unsymmetrical, in that the transfer of DNA from archaea to bacteria is more than five times more frequent as in the other direction (Nelson-Sathi et al., 2015).

The importance of HGT in eukaryotes is often overshadowed by its pervasive impact on prokaryotes (Van Etten and Bhattacharya, 2020; Keeling and Palmer, 2008). Many recent studies have revealed that genes with adaptive or important metabolic functions are derived through HGT. For example, the ice-binding proteins that can mitigate freezing damage in Antarctic algae *Chlamydomonas* sp. are acquired from bacteria through HGT (Raymond and Morgan-Kiss, 2017). HGT was also observed in the antifreeze genes of fish, which allows fish to live in icy environments (Graham et al., 2008). To live in an extreme hot spring environment, the red alga *Galdieria sulphuraria* encodes a series of functional genes, such as genes involved in detoxifying heavy metals to glycerol as the carbon source. These genes originated in bacteria, while genes encoding the ability to resist high salinity are from archaea (Schönknecht et al., 2013). Additionally, approximately 2.5% of genes in the human gut parasite *Blastocystis* spp. have been acquired by HGT from both eukaryotic and prokaryotic donors. These genes participate in diverse metabolic pathways and promote better adaptation of the host in an anaerobic environment (Eme et al., 2017).

1.1.3 Sequence Similarity Network (SSNs) in Evolutionary Studies

Introgressive descent like HGT results in reticulate evolution and has created a complex evolutionary pattern, which is the hobgoblin of tree-based models (DeSalle and Riley, 2020). With the rapid development of high-throughput sequencing technology, enormous biological datasets have become available. To comprehensively investigate reticulate evolutionary processes, a novel approach of using network-based models has been employed in evolutionary studies (Corel et al., 2016; Proulx et al., 2005).

Traditionally, networks are extensively used in the analysis of traffic systems, such as railroads mapping (Jordan and Turnquist, 1983). In the modern society, the usage has expanded to Internet searching, web building (Chau, 2011) and analysing relationships on social media such as Facebook (Viswanath et al., 2009). In the epidemiological field, the importance of network-based models is first reported in tracing AIDS “Patient Zero” and revealed the transmission pathway of sexual contacts (Klovdahl, 1985). This network thinking was also used in the study of the SARS-COV-2 pandemic from the perspective of tracing origins (Forster et al., 2020) and tracking the spread route in population (Gudbjartsson et al., 2020).

In the field of genetic analyses, sequence similarity networks (SSNs) are widely used in presenting reticulate evolutionary relationships since its first proposal by Tatusov et al. (1997). In SSNs, entities (such as genes, genomes or species) are represented as nodes whilst the evolutionary connections (such as significant similarity between genes, common features among genomes) are represented by edges (Figure 1.2) (Bapteste et al., 2012; Coleman et al., 2015). Normally, SSNs consist of at least one connected component(s) (CCs) (Figure 1.2) which are made up of related sequences (Alvarez-Ponce et al., 2013). Those entities in any given clique do not have to be genealogically related to all others, but usually have comparable functions (Pradhan et al., 2012).

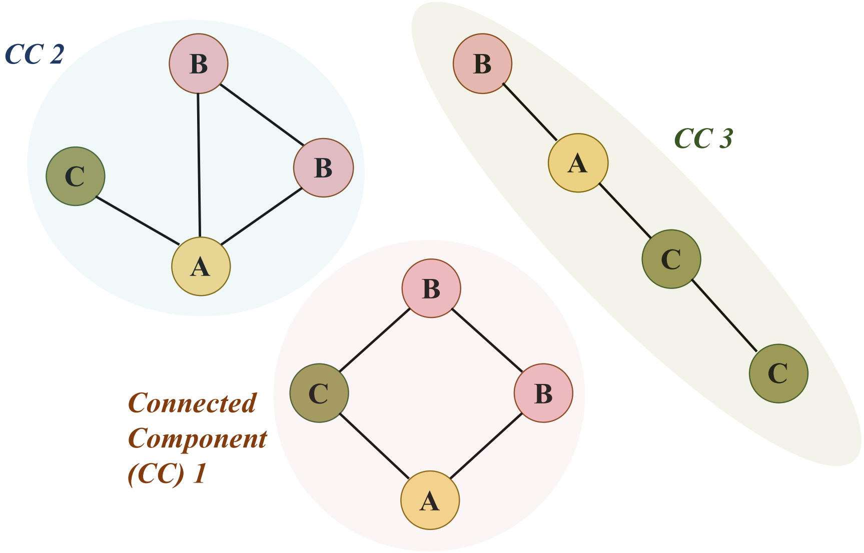


Figure 1.2 Sequence Similarity Networks (SSNs) are mostly composed of more than one connected component (CCs). Encoding sequences are represented by nodes (coloured circles A/B/C), while the evolutionary relationships are illustrated by edges (short lines) between nodes. Several CCs samples are highlighted as large colour aureolas which contains high accordance nodes. Adapted from Corel et al. (2016).

SSNs have been employed in many research projects analyzing evolutionary relationships (Alvarez-Ponce et al., 2013), detecting gene sharing and recombination, and performing functional annotation. In 2010, Fondi and Fani constructed SSNs to reveal that through multiple HGT events, distantly located or related bacteria can exceed the distance both geographically and genealogically. They described a global influence of this type of "horizontal flow" on whole bacterial community regarding antibiotic resistance determinants (Fondi and Fani, 2010). With the help of SSNs, Halary et al. (2013) assembled a database including over half million protein sequences from prokaryotes, eukaryotes, and mobile genetic elements (MGEs) and investigated the plasmids and their relevant genes. They found that although most genes were widely shared, in *Borrelia* genes were more like private genetic goods and were not so widely shared (McInerney et al. 2011). It is likely that this restriction in gene sharing contributed to the survival of *Borrelia* against the host immune environment (Barbour et al., 2006; Chaconas and Kobryn, 2010). Later, Cheng et al. (2014) identified two major diversification events in the evolution of prokaryotic pathways through the SSN approach. SSNs can also be used in investigating community diversity. Arroyo et al. (2020) discovered a putative novel holozoan group in *Tara* Ocean through building and exploring a unicellular holozoa network which contains collected environmental sequences and reference sequences.

Many tools already have been developed to visualise and modularize networks such as Gephi (Bastian et al., 2009), Cytoscape (Smoot et al., 2011), and a recent tool Graphia that can present networks in 3D space (Freeman et al., 2020). Although network-based models have better performance in presenting introgressive descent, treelike histories and reticulate histories are not necessarily in conflict with one another. Composite entities that result from introgression can subsequently carry out vertical descent. Fusion between entities from introgressive descent generate a network between bifurcating trees (Corel et al., 2016). Therefore, network methods have gained attention owing to their capacity to depict both vertical and lateral evolutionary history at the same time (Corel et al., 2016; McInerney et al., 2011). My PhD focussed on applying network thinking in investigating reticulate evolution on composite genes and CRISPR-Cas loci. In this section, the mechanisms and recent findings about these two systems were briefly introduced.

1.2 Origin and Pervasive Existence of New Composite Gene Structures

Composite genes are produced by gene remodelling and play a key role in the evolution of many biological entities including eukaryotes, prokaryotes, plasmids and viruses (Chothia et al., 2003; Jachiet et al., 2013, 2014; Méheust et al., 2018). Composite genes are composed of component genes which are genetic fragments derived from other gene families (Corel et al., 2016). As shown in Figure 1.3a, sequence A and B, C and B show similarity to each other whereas sequence B and C has no overlap. Therefore, A is regarded as composite gene whilst B and C are regarded as component genes and the clique A, B and C is called a “non-transitive triplet” (Haggerty et al., 2014).

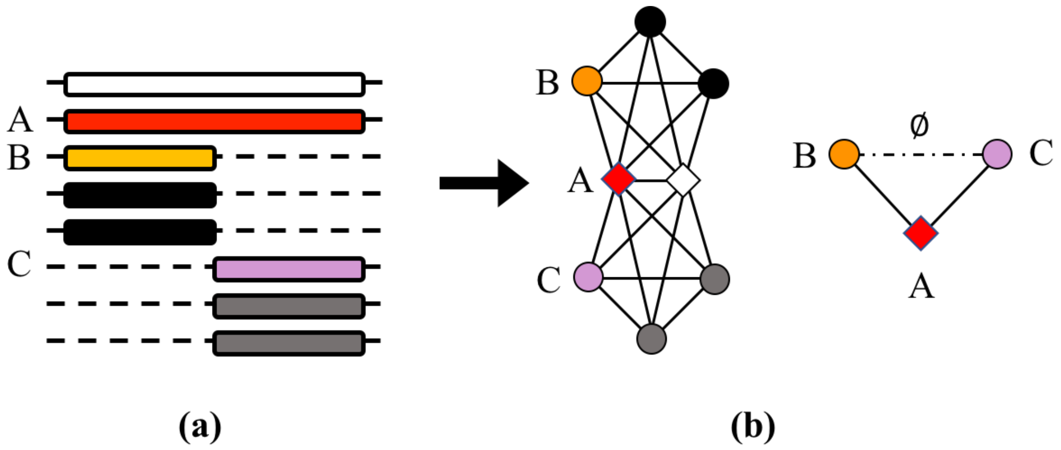


Figure 1.3 Patterns of non-transitive triplet, composite and component genes in sequence similarity network (SSN). (a) Sequence A is known as composite gene while sequences B and C is known as component genes. When component genes show no similarity to each other, in other words, there is no overlap between two component genes, sequence A, B and C are in the nontransitive relationship. **(b)** Non-transitive triplets in SSN. Adapted from Jachiet et al. (2013) and Haggerty et al. (2014).

Many factors can contribute to these non-transitive relationships, including gene fusion, exon or domain shuffling and non-homologous recombination (Bapteste et al., 2012; Corel et al., 2016; Haggerty et al., 2014).

1.2.1 Mechanisms of Gene Fusion

Gene fusion occurs when open reading frames (ORFs) from previously separate sources combine and form a new transcription unit, which is an important molecular mechanism of generating new genes (Long et al., 2003), especially multi-domain genes (Pasek et al., 2006). There are two main mechanisms promote the formation of fusion genes: genetic rearrangement such as gene duplication (Figure 1.4a), and transcription read-through (Figure 1.4b) (Kaessmann, 2010).

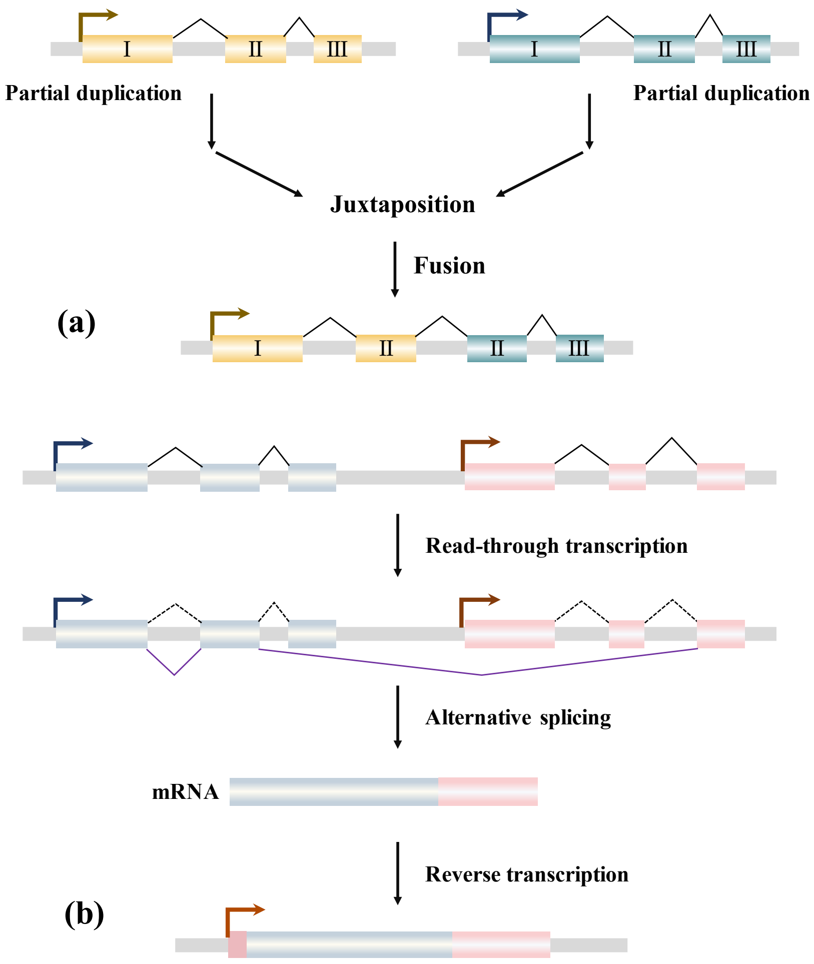


Figure 1.4 The mechanisms gene fusion. (a) Partial duplicated gene can be juxtaposed and fused to a new fusion gene. **(b)** Fusion genes can also form through transcription read-through with intergenic splicing. Chimeric mRNA might be integrated into genome as fusion genes through reverse transcription. Adapted from Kaessmann (2010).

Rearrangement-mediated gene fusions

Gene fusion could occur through rearrangements of duplicated gene copies at the DNA level (Figure 1.4a) (Kaessmann, 2010; Rogers et al., 2009). One of the most famous examples is origin of the *jingwei* gene in *Drosophila* (Long, 2000). *Jingwei* (*jgw*) is formed through a duplication event of *Yellow-emperor* (called *yande*), a retroposition event of the alcohol dehydrogenase gene (*Adh*) into the middle of *yande* and a recombination of *Adh* and partial exons of *yande* (Long et al., 2003, 2013). Expressed JGW proteins showed a novel specificity of detoxifying long-chain alcohols whereas its ancestral ADH preferred short-chain alcohols such as ethanol

(Zhang et al., 2004). Additionally, for fusion genes from duplication, the original function of the parental gene can be maintained (Kaessmann, 2010). More fusion genes with beneficial functions have also been identified in flowering plants and mammals (Brennan et al., 2008; Liu et al., 2009).

Transcription-mediated gene fusions

Another mechanism of gene fusion is through transcription read-through from neighbouring genes and intergenic splicing at RNA level (Figure 1.4b) (Kaessmann, 2010; Latysheva and Babu, 2016). Fusion transcripts can be reverse transcribed and integrated into the host genome (Kaessmann, 2010; Long et al., 2013). The hominoid gene *PIPSL* was found to have arisen through co-transcription and intergenic splicing of two adjacent parental genes *PIP5K1A* and *PSMD4*, followed by a retrotransposition event (Babushok et al., 2007). This new fusion gene *PIPSL* experienced positive selection shortly after emergence, which indicates the beneficial role it encodes (Babushok et al., 2007; Kaessmann, 2010). Using the approach of SSNs, McCartney et al. (2019) identified 45 novel genes from seven animals that were formed through transcription-mediated gene fusions and 64.4% of these fusion genes had annotated transcripts.

1.2.2 Network Thinking in Analysis of Composite Genes

Composite genes could result from juxtaposition of non-homologous genes, which step over the boundary of single gene families and become obstacles of tree-based models (Corel et al., 2016; Haggerty et al., 2014). To investigate this process, the network-based model such as SSN is equipped with several advantages (Figure 1.3). First, it can depict composite and component genes in genomes or communities systemically. Second, the convergent and divergent events can be distinguished by taxonomical distribution comparison. Third, it can be used to evaluate the conservation overlapping degree of composite and component fragments (Corel et al., 2016).

Composite genes are widespread in nearly all huge and dominant communities, especially those giant connected components (GCCs) of SSNs. Haggerty et al. (2014) identified composite genes from 15 eukaryote genomes using a network approach. They found that in the GCC, approximately one-quarter of the sequences belonged to composite genes and one-tenth of the sequences were detected as multi-composite genes. Multi-composite genes are formed from two or more composite genes. Nonetheless, in the rest of the network, only 6% sequences are identified as composite genes (Haggerty et al., 2014). In the same way, Coleman et al. (2015) reconstructed SSNs of using published data (Kim and Yi, 2012) containing 319 actinobacterial genomes and detected 13 composite genes. This network was decomposed into 10 small CCs and one GCC which occupied 82% of the network and covered all 13 composite sequences. Jachiet et al. (2013) mined the 591,439 sequences from three domains of life, viruses and plasmids through constructing SSNs. Gene fusions have been found from both cellular organisms and MGEs, and the frequency of composite genes in eukaryotes was significantly higher than in prokaryotes and MGEs. It has been shown that 53% of triplets were detected from cellular organisms and 42% detected between prokaryotes and MGEs. In particular, genes of mobile elements were involved in a large portion of gene fusion events. However, transferred mobile elements can only maintain and propagate in host lineages with quite a low probability (Graham et al., 2008). The sources contributing to mobile element gene remodelling therefore need to be investigated further. In the viral world, Jachiet et al. (2014) applied the mobilome network into 3,008 complete viral genomes. Probably due to the blurry boundaries between viral gene families, there were 8-15% composite genes observed from the network. Additionally, these mosaic genes distributed extensively in all functional categories in viruses and almost all were crucial functional encoding sequences, for instance the family of DNA polymerase beta/AP endonuclease proteins and multi-domain helicase/methyltransferase proteins. Furthermore, network analysis also promotes our understanding of the origin of composite genes. Eukaryotic-specific multi-domain nitrate reductase EUKNR was found to have arisen from fusion of eukaryotic SUOX, Cyt-b5, and NADH reductases using a SSN-based model (Ocaña-Pallarès et al., 2019). In the origin of Haloarchaea, 320 new composite genes were identified in early history and then subsequently proliferated to descendent groups (Méheust et al., 2018). Among these composite genes, almost 40% have parental genes from bacterial genomes, involved in metabolic pathways and promoting the adaptation to oxygenic and salty niches (Méheust et al., 2018).

1.3 Mechanisms and Evolution of CRISPR-Cas Systems

1.3.1 The History of CRISPR Discovery

Direct repeat structures were first identified by Ishino and colleagues in 1987, when they analysed the *iap* gene in *E. coli*. They noticed highly conserved DNA fragments of 29 nucleotides were spaced by DNA fragments of 32 nucleotides (known as spacers) on the 3' flank of the *iap* gene (Ishino et al., 1987). A similar pattern was also seen in the Gram positive bacterium *Mycobacterium tuberculosis* complex (MTBC) (Van Soolingen et al., 1993). Later, Mojica et al. (1995) discovered interspaced dyad symmetry repeats in the archaeal species *Haloferax mediterranei* and *Haloferax volcanii*. This work was credited by Jansen et al. (2002) and together they referred to the class of repeats as one family is now known as: Clustered Regularly Interspaced Short Palindromic Repeats (the acronym CRISPRs). Jansen et al. (2002) suggested spacers with similar length but unique sequences could potentially be important in prokaryotes and remain to be deciphered.

In addition, Jansen et al. (2002) firstly recognised CRISPR-associated (*cas*) genes. They found four homologous genes (*cas1* to *cas4*) that associated with CRISPR loci, adjacent to the repeat arrays, in a specific order. Functional analysis of Cas3 and Cas4 proteins implied that they might play a role in DNA modifying and cleavage. Additionally, Jansen et al. (2002) noticed different species with dissimilar repeats also possessed different sets of *cas* genes. Furthermore, the analysis of homologies indicated that these Cas genes were functionally associated with CRISPR loci. This work was a breakthrough for the construction of the model of CRISPR-Cas system and indicated that a diverse classification of this system was needed.

In 2005, Mojica et al. (2005) observed the homology of spacers in CRISPR loci with external sequences like bacteriophages and conjugative plasmids, which firstly predicted its biological functions in defending foreign DNA like the eukaryotic RNA interference system. At the same time, Pourcel et al. (2005) and Bolotin et al. (2005) also independently reported that spacers were derived from extrachromosomal elements in *Yersinia* and *Streptococcus*. More Cas genes were then identified (Haft et al., 2005) and the role that the CRISPR-Cas system may play in prokaryotic immunity was becoming clearer (Koonin et al., 2006). In 2007, a laboratory experiment observed novel spacer integration in *Streptococcus thermophiles* after challenge by bacteriophages (Barrangou et al., 2007). They suggested that Cas proteins in CRISPR-Cas system could mediate adaptation of novel as well as direct invader immunization. The same research also reported that bacterial resistance to phages was correlated with similarity between spacers and phage. Since then, research has focussed on the mechanisms by which CRISPR-Cas systems work, and the potential applications of CRISPR-Cas in gene editing (Adli, 2018; Cong et al., 2013; Deveau et al., 2008; Mali et al., 2013a; Sapranauskas et al., 2011). Thrillingly, the 2020 Nobel Prize in Chemistry was rewarded to Emmanuelle Charpentier and Jennifer Doudna for their discovery and application of CRISPR-Cas technology in precise gene editing (Strzyz, 2020). CRISPR-Cas technology was also employed in rapid disease infectious diagnosis (Strich and Chertow, 2018), which has made great contributions to SARS-CoV-2 detection (Broughton et al., 2020).

1.3.2 Classification and Mechanisms

CRISPR-Cas is an adaptive immune system that exists in most Archaea and half of Bacteria (Horvath and Barrangou, 2010). It contains a CRISPR array and a cluster of *cas* genes. A CRISPR array is composed of nearly identical direct repeats and interspaced diverse spacers. Spacers are derived from different foreign MGEs, derived from situations where the host species is under attack (Shmakov et al., 2017b). The process of acquiring new spacer sequences is one of three stages of the CRISPR-Cas mediated defence process, and it is called Adaptation (Figure 1.5). The other two processes are known as Expression and Interference. In the expression stage, an array containing spacers is transcribed into a long precursor CRISPR RNA (pre-crRNA) and this RNA is then processed into smaller mature CRISPR RNAs (crRNAs). In the subsequent interference stage, the invading sequence is combined with complementary matured crRNA and is destroyed by incorporated effector Cas complex (Rath et al., 2015). Upstream of the CRISPR arrays is an AT-rich leader sequence. Though it does not directly participate in elimination of invading genetic elements, it is essential in regulating spacer integration and crRNA biogenesis (Hille et al., 2018; Yosef et al., 2012).

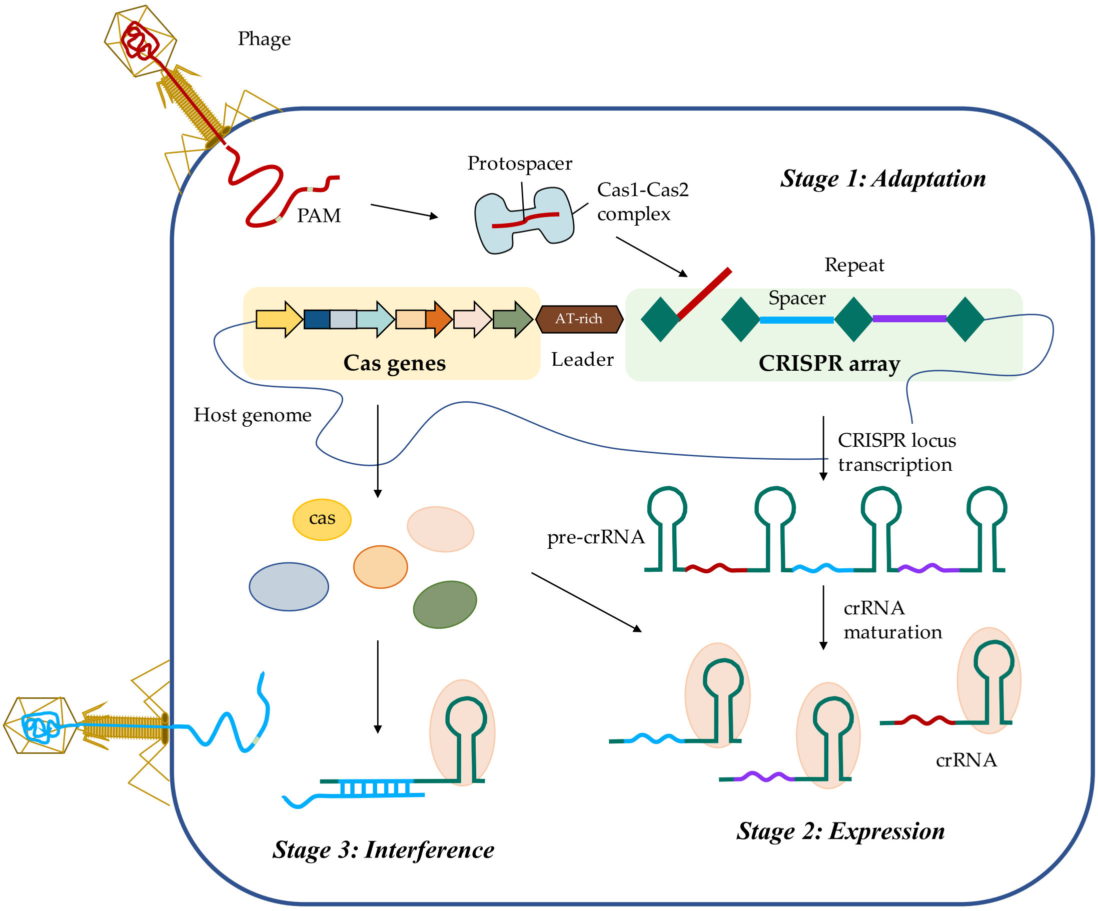


Figure 1.5 Overview of CRISPR-Cas in prokaryotic immunity. The defence process of CRISPR-Cas system can be divided into three stages. Upon integration, part of foreign segment is adapted into CRISPR locus that is located in host genome while under attack from a novel invader. Subsequently, if a known phage is encountered, CRISPR locus will be transcribed into pre-crRNA and processed into matured crRNAs. These crRNAs guide a cluster of Cas proteins and protect the host by cleaving invading sequences.

According to studies that have been published to date, CRISPR-Cas systems are classified into 2 classes, 6 types and 33 subtypes. The classification of class 1 includes type I (subtype A, B, C, D, E, F, G), type III (subtype A, B, C, D, E, F) and type IV (subtype A, B, C) while class 2 includes type II (subtype A, B, C), type V (subtype A, B, C, D, E, F, G, H, I, U) and type VI (subtype A, B, C, D) (Makarova et al., 2020). The main difference between these two classes is the effect modules where the class 1 system employs multiple Cas proteins on crRNA binding and target cleavage whereas the class 2 system uses a single but multi-domain Cas protein on these activities (Figure 1.6) (Makarova et al., 2020). Class 1 systems occupy widely around 90% of all CRISPR-Cas while the remaining 10% are class 2 systems and mostly discovered from Archaea (Makarova et al., 2012, 2020). The famous Cas9 protein that is used in genetic engineering is derived from type II system that belonging to class 2 (Hsu et al., 2014). This subsection will briefly introduce the mechanisms of three stages in two CRISPR-Cas classes.

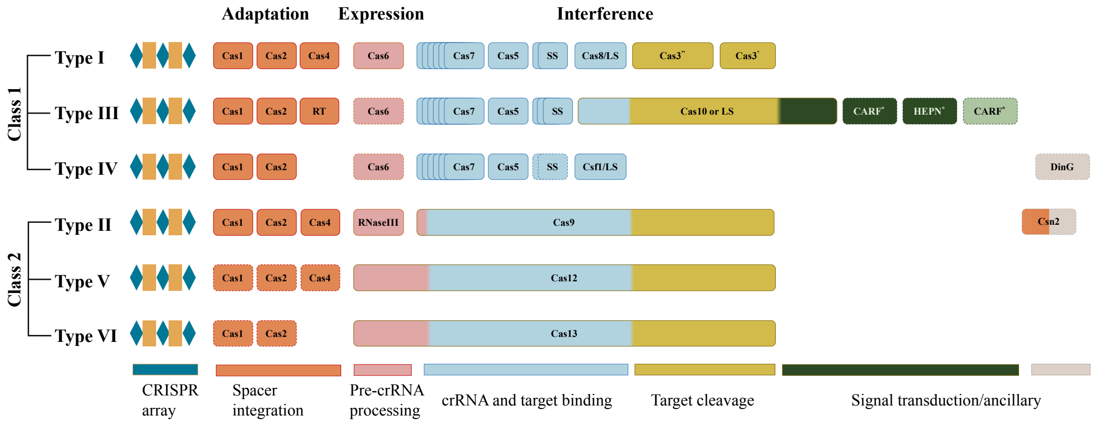


Figure 1.6 The function modules of two classes of CRISPR-Cas systems. Multiple effector modules are involved during expression and interference of class 1 system while single but multi-domain Cas protein is encoded in class2 system. RT, reverse transcriptase; LS: large subunit; SS, small subunit. Adapted from Koonin & Makarova (2019).

1.3.2.1 The Mechanism of Adaptation

As an adaptive immune system, adaptation (also known as integration or acquisition) is the key stage that endows CRISPR-Cas systems with the ability to store memories of invaders, and enables the inheritance of the adaptation by offspring (Figure 1.7). Integration in both classes are mediated by two core proteins: Cas1 and Cas2. These two proteins are highly conserved in almost all subtypes (Koonin et al., 2017). The integration stage can be concisely divided into two steps: 1) identify and extract the foreign segments (called "protospacer") from the exogenous genome; 2) integrate the spacer into the right position of the CRISPR array in the host genome. Cas1 and Cas2 form as a heterohexameric $[(\text{Cas1}_2\text{-Cas2})_2]$ complex (hereafter, Cas1-Cas2) and play important roles through both phases of integration (McGinn and Marraffini, 2018). Here, we introduce the spacer adaptation in two phases based on the current best-studied models: type I-E and type II-A.

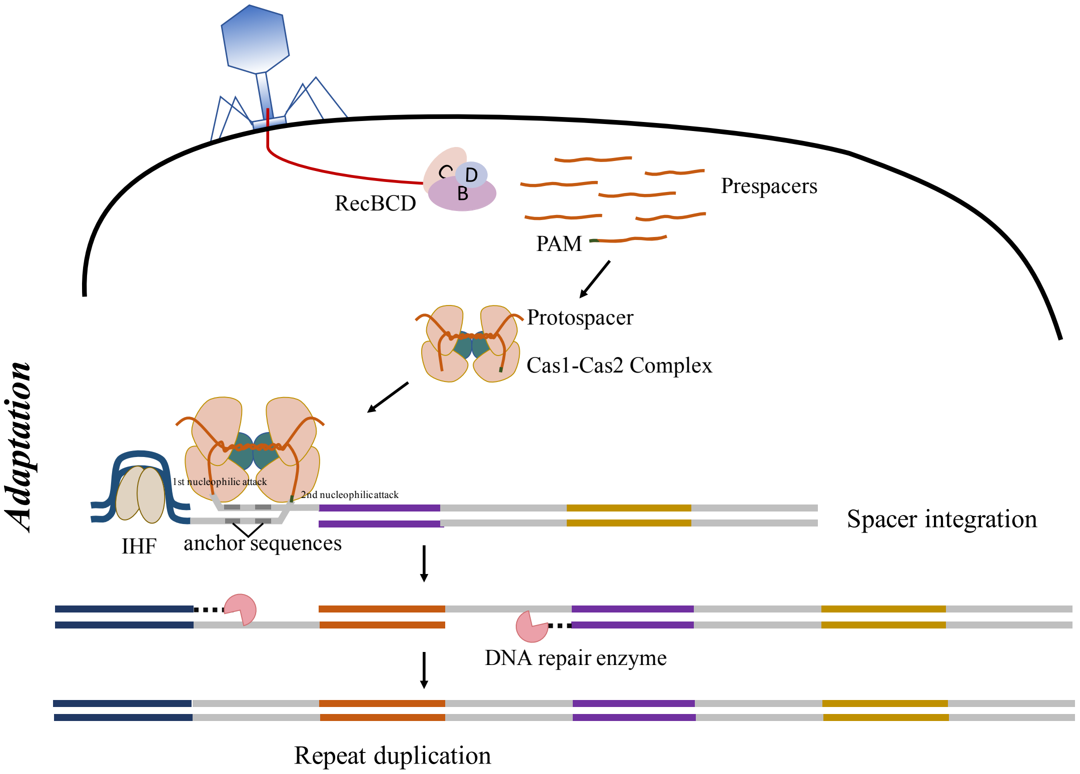


Figure 1.7 The mechanism of spacer integration in CRISPR-Cas system. Protospacer substrates (prespacers) are generated by DNA repair system like RecBCD or AddAB. Cas1-Cas2 then binds and cleavage specific length of protospacer. With the help of integration host factor (IHF) in type I-E or leader anchoring sequence (LAS) in type II-A, Cas1-Cas2 with protospacer complex causes two nucleophilic attacks between leading site and the first repeat, which result in precisely integration and first repeat duplication.

Adaptation - Phase 1

During the first phase of adaptation (Figure 1.7), protospacers from foreign genetic elements are first recognised and processed before being integrated into CRISPR loci. In Gram positive species, substrates that contain protospacers are possibly generated from phage genomes by DNA repair system RecBCD (Bernheim et al., 2019; Levy et al., 2015; Radovčić et al., 2018). RecBCD can unwind and degrade DNA fragments during the repair of double-stranded DNA (dsDNA) breaks (DSBs), and often works on replication forks (Wigley, 2013). Meanwhile, RecBCD activity is controlled by asymmetry crossover hotspot instigator (Chi) sites. A Chi site is an eight-nucleotide length motif and serves as a repressor. In other words, RecBCD degrades linear DNA to smaller segments until it reaches Chi sites (Smith, 2012). Chi works in an asymmetrical manner, which requires RecBCD only to interact from one side, and RecBCD from the opposite direction would be limited by reverse complement of Chi sites (Smith, 2012). The degraded segments can then be recognised by Cas1-Cas2 complex in order to facilitate subsequent integration (Ivančić-Baće et al., 2015). This proposal is supported by identification of protospacers hotspots between stalled folks and Chi sites (Levy et al., 2015). Similar results also were found in Gram positive organisms who use RecBCD’s paralogs AddAB as the DNA repair machinery (Modell et al., 2017). The follow-up study further clarified that the helicase activity of RecBCD is important to spacer adaptation rather than the nuclease activity and inactivation of 5’ single-stranded (ssDNA) exonucleases could invoke significant spacer acquisitions (Radovčić et al., 2018). Nevertheless, spacer adaptation was still identified in strains that lacked RecBCD or AddAB (Levy et al., 2015; Modell et al., 2017). This indicated the possibility of other routes that were also able to provide protospacer substrates, such as restriction-modification systems (Dupuis et al., 2013b; McGinn and Marraffini, 2018).

Chi sites are highly enriched in host genomes compared with phage genomes and MGEs. This enrichment is an important method for differentiating self from non-self genomes. The activity of RecBCD (or AddAB) in prokaryotic organisms is restrained by a high density of Chi sites in order to protect species from integrating self DNA and autoimmunity. By contrast, invading phages or plasmids that are devoid of Chi sites can have their DNA mostly or fully degraded and supply sufficient protospacer substrates to Cas1-Cas2 (Levy et al., 2015).

In addition to Chi sites, protospacer adjacent motifs (PAM) that are located in phage genomes also assist with locating protospacers and avoiding host autoimmunity. This short motif is normally two to six nucleotides long and situated immediately downstream of targets (Mojica et al., 2009; Shah et al., 2013). In type I-E systems, the protospacer substrates (called "prespacers") that degenerated by RecBCD in last step forms dual-forked structure, and then binds to and is stabilized by Cas1-Cas2 complex. PAM in the 3' overhang of the dual-fork DNA is recognised and cleaved by one Cas1. Considering Cas1-Cas2 complex is symmetrical with two Cas1 proteins, after one Cas1 binds with one side, the 3' overhang on the other side is subsequently bound and cleaved by the other corresponding Cas1. This results in two-side 3'OH on each overhang (Li et al., 2015). In this process, the length of captured protospacer is likely to be determined by the Cas1-Cas2 complex which serves as a molecular ruler (Li et al., 2015; Nuñez et al., 2015). A recent finding reported Cas4 might also be functional with PAM in type I CRISPR-Cas systems (Kieper et al., 2018). In type II system, Cas9 is also required to specifically interact with PAM-adjacent protospacers along with the Cas1-Cas2 complex and accessory protein Csn2 (Heler et al., 2015). By contrast, some type III systems have been identified that favoured integrating RNA transcripts as spacers rather than DNA segments. This preference for RNA transcripts is encoded by a fusion protein that is comprised of reverse transcriptase (RT) and Cas1 (Figure 1.6) (Silas et al., 2017a).

Adaptation- Phase 2

In the second phase of spacer adaptation, processed protospacers integrate into the leader end of CRISPR array. This is thought to record past infections in a chronological order (Barrangou et al., 2007). To ensure precise acquisition, an AT-rich leader sequence that lies upstream of CRISPR arrays and plays a vital role. In type I-E systems, to facilitate the recognition of boundary between leader sequence and the first repeat by Cas1-Cas2 complex, an Integration Host Factor (IHF) specifically binds with the leader sequence and forms a U-shape structure (Figure 1.7) (Nuñez et al., 2016). Cas1-Cas2 with a protospacer then docks into the leader-proximal repeat. Next, two nucleophilic attacks happen on borders of the repeat. In the first nucleophilic attack, the 3’OH of the protospacer interacts with the leader-repeat junction and ligates to the repeat. The second nucleophilic attack happens between the first repeat-spacer boundary (Nuñez et al., 2016; Yoganand et al., 2017). Moreover, there are two conserved and palindromic inverted repeat motifs in the CRISPR repeat settle Cas1-Cas2 like an anchor (Figure 1.7). The Cas1-Cas2 complex again serves as a molecular ruler and determines the size of the repeat and the second attack site (Goren et al., 2016). After the spacer integrates in a polarized manner, double strands of the first repeat serve as templates, and gaps are filled by DNA repair enzymes, resulting in the duplication of the first repeat (Jackson et al., 2017).

The IHF is essential for the process described above, therefore species who lack IHF require other mechanisms in order to guide spacer integration. In type II-A, a short motif that directly upstream to the CRISPR array determines the site of new spacers. This conserved site is known as the leader-anchoring site (LAS). Mutations in the LAS would lead to ectopic spacer integration in which spacer inserts in the middle of an array (McGinn and Marraffini, 2016). A similar conserved leader motif has also been reported in type I-D (Kieper et al., 2019), but the detailed molecular mechanism still remains to be discovered.

The two adaptation phases as described above are the canonical mechanisms of CRISPR-Cas when encountered a novel invader, which is termed “naïve spacer adaptation” (Fineran and Charpentier, 2012). However, phages can escape elimination through point mutations in PAM or target sequences (Deveau et al., 2008; Westra et al., 2014). To overcome this possibility, prokaryotes carry out a different integration processes called primed spacer adaptation (also known as priming) (Fineran et al., 2014; Richter et al., 2014). For those invaders who perfectly or partially match pre-existing spacers, primed adaptation triggers rapid and efficient integration of additional spacers. Primed adaptation is closely related to interference, which has been reported in laboratory experiments and bioinformatics studies of type I (Fineran et al., 2014) and type II CRISPR-Cas systems (Nicholson et al., 2019; Nussenzweig et al., 2019). In the type I-E system, in order to perform this interference-driven adaptation, the Cas1-Cas2 complex is required to interact with CRISPR-associated complex for antiviral defence (Cascade) and Cas3. After recognising the presence of PAM in exogenous MGE, the Cas1-Cas2 can bind to the invading segment and Cas3 can annihilate it at the interference stage (Figure 1.9) (Richter et al., 2014; Swarts et al., 2012). Cas3 retains helicase and nuclease activities which can generate protospacer substrates for Cas1-Cas2 acquisition (Künne et al., 2016). Similar machinery has also been reported in type I-F which includes a large Cas1-Cas2-3 complex (including a composite protein fusion of Cas2 and Cas3) operating in both the interference and acquisition stages (Fagerlund et al., 2017; Rollins et al., 2017). However, the presence of priming in other types still remains unclear, and the underlying mechanisms in type II system are also waiting to be addressed (Nussenzweig et al., 2019).

1.3.2.2 The Mechanism of Expression

Sequence-specific target elimination by CRISPR-Cas system relies on spacers that are found in the CRISPR loci and are complementary to the target foreign genetic elements. At the expression phase, the CRISPR locus is normally transcribed from a site in the leader sequence, and produces a long pre-crRNA that contains the complete array transcript. The pre-crRNA is then processed by Cas genes into mature crRNAs that include the spacer and part of the repeat (Haurwitz et al., 2010). The process of transcription is similar across the two classes, whereas maturation is carried out using different endoribonucleases (Figure 1.8) (Richter et al., 2012a).

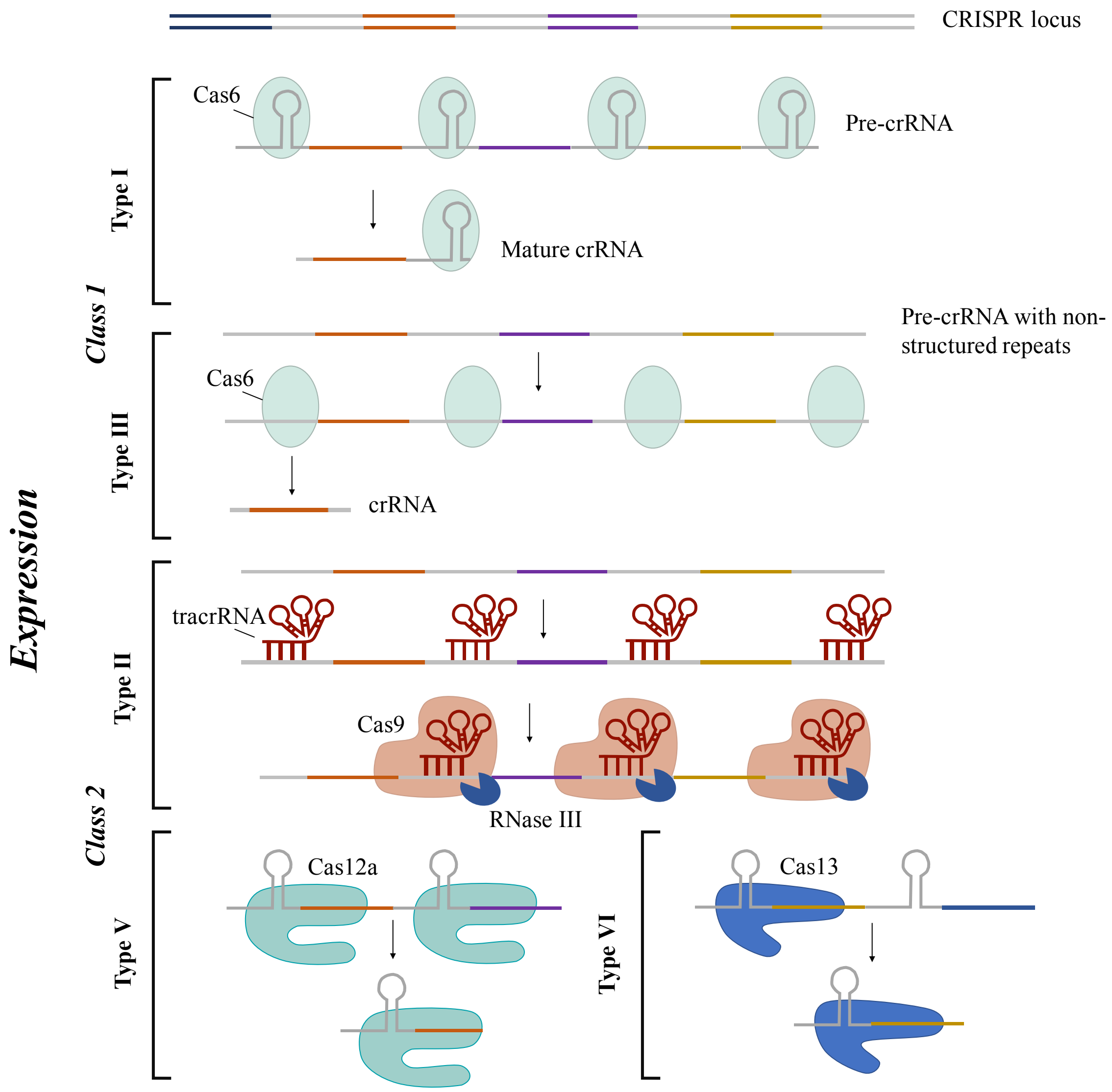


Figure 1.8 Spacer expression in CRISPR-Cas system. Expression varies in two classes. Pre-crRNA in class1 system are mainly processed by Cas6. Repeats with hairpin structures can be directly recognised by Cas6 while the non-structured ones are based on sequence distinguishing. The multi-domain proteins Cas9, Cas12 and Cas13 are employed in crRNA biogenesis. A special trans-activating crRNA (tracrRNAs) is required in type II. It can bind with repeat in crRNA and form a duplex for Cas9 combination. The long pre-crRNA with multiple effector complexes is then cleavage by RNase III and an unknown RNase to short mature crRNAs. By contrast, repeats in type V and VI can form stem-loops and directly be distinguished and cleaved by Cas12 and Cas13, respectively. Adapted from Hille et al. (2018).

In almost all class1 systems, the group of Cas6 and its variants are used during expression. For instance, Cas5d in type I-C (Nam et al., 2012), Cas6f in type I-F (Przybilski et al., 2011) and Csf5 in type IV (Özcan et al., 2019) can all be used. Cas6 can recognise and cleave repeat sequences in pre-crRNA, and generate mature crRNAs. Most CRISPR loci of type I contain palindromic repeats that are able to form stem-loop structures, which provide bonding and cleavage sites for endoribonucleases. After processing, the mature crRNA consists of a full spacer, a 5' repeat handle and a 3' stem-loop that belongs to the next repeat (Figure 1.8) (Haurwitz et al., 2010; Richter et al., 2012a; Sashital et al., 2011). In particular, Cas6 remains associated with crRNAs after cleavage and facilitates Cascade formation. CrRNA serves as scaffold for other Cas proteins combinations in the subsequent interference stage (Nam et al., 2012; Wiedenheft et al., 2011). In contrast, repeats in type I-A and type I-B are non-palindromic (Kunin et al., 2007), and these cannot form stem-loops naturally. Though hairpin structure is very importation for Cas6 cleavage, Cas6 can work independently of repeat sequence recognition (Wang et al., 2016b). However, other studies assumed the interaction was more related with Cas6 remodelling activity and formed a dimerization structure for reposition site recognition (Sefcikova et al., 2017; Shao and Li, 2013; Shao et al., 2016). Similarly, repeats that are found in type III systems are predominantly non-structured or only form week stem-loops (Kunin et al., 2007). In type III-A system, metal ions mediate Cas6 activity for crRNA processing (Hatoum-Aslan et al., 2014; Wei et al., 2019). The further maturation requires a trans-acting non-Cas nuclease protein; however, full elucidation of the underlying mechanism requires more investigation (Hille et al., 2018). In particular, for systems that lack structured repeats, such as I-A, I-B and III-A, Cas6 detaches crRNA after cleavage and do not participate in subsequent Cascade formation (Plagens et al., 2014; Richter et al., 2012b). Meanwhile, for systems such as the III-B variant that lack homologues of Cas6 nucleases, crRNA biogenesis might be carried out by host housekeeping endonuclease RNase E as a replacement (Behler et al., 2018).

CRISPR expression and crRNA maturation in type 2 systems is quite different to class 1. In type II systems, RNAs that are called trans-activating crRNAs (tracrRNAs) mediate maturation and immunity (Deltcheva et al., 2011; Shmakov et al., 2015). These tracrRNAs contain segments that can base-pair with repeats within the pre-crRNA and this compensates for the drawback associated with unstructured repeats. These tracrRNA:crRNA duplexes can be combined and stabilized by the Cas9 effector protein in preparation for interference (Deltcheva et al., 2011). Then, a RNase III is recruited to cleave within the repeat. The tracrRNA:crRNA duplex is subsequently matured through removing 5' repeat-derived tag by an unknown RNase (Hille et al., 2018; Jinek et al., 2012). Transcription in type II-C is slightly different. Promoter elements are located in repeats rather than in the leader sequence, so crRNAs are individually transcribed before binding with tracrRNAs and Cas9 (Zhang et al., 2013). In type V and VI systems, repeats with stem-loops can be recognised by the corresponding effector proteins Cas12 and Cas13 (Figure 1.8). Most systems in type V and VI do not require tracrRNAs for crRNA processing except type V-B (Shmakov et al., 2015). The multi-domain proteins Cas12 and Cas13 are able to cleavage within the repeat and accomplish crRNA biogenesis independently (Hille et al., 2018). It has also been noted that some long pre-crRNAs in type VI-A can directly serve as guides for the interference machinery without maturation (East-Seletsky et al., 2017).

1.3.2.3 The Mechanism of Interference

During the last stage of CRISPR-Cas mediated immunity, mature crRNA guides interference machinery and performs a sequence-specific destruction, which provides robust protection against invaders. Generally, interference can be divided into two steps: 1) target recognition and binding; 2) target cleavage. Both stages of interference involve different Cas proteins, which causes a diverse classification of CRISPR-Cas systems (Makarova et al., 2020).

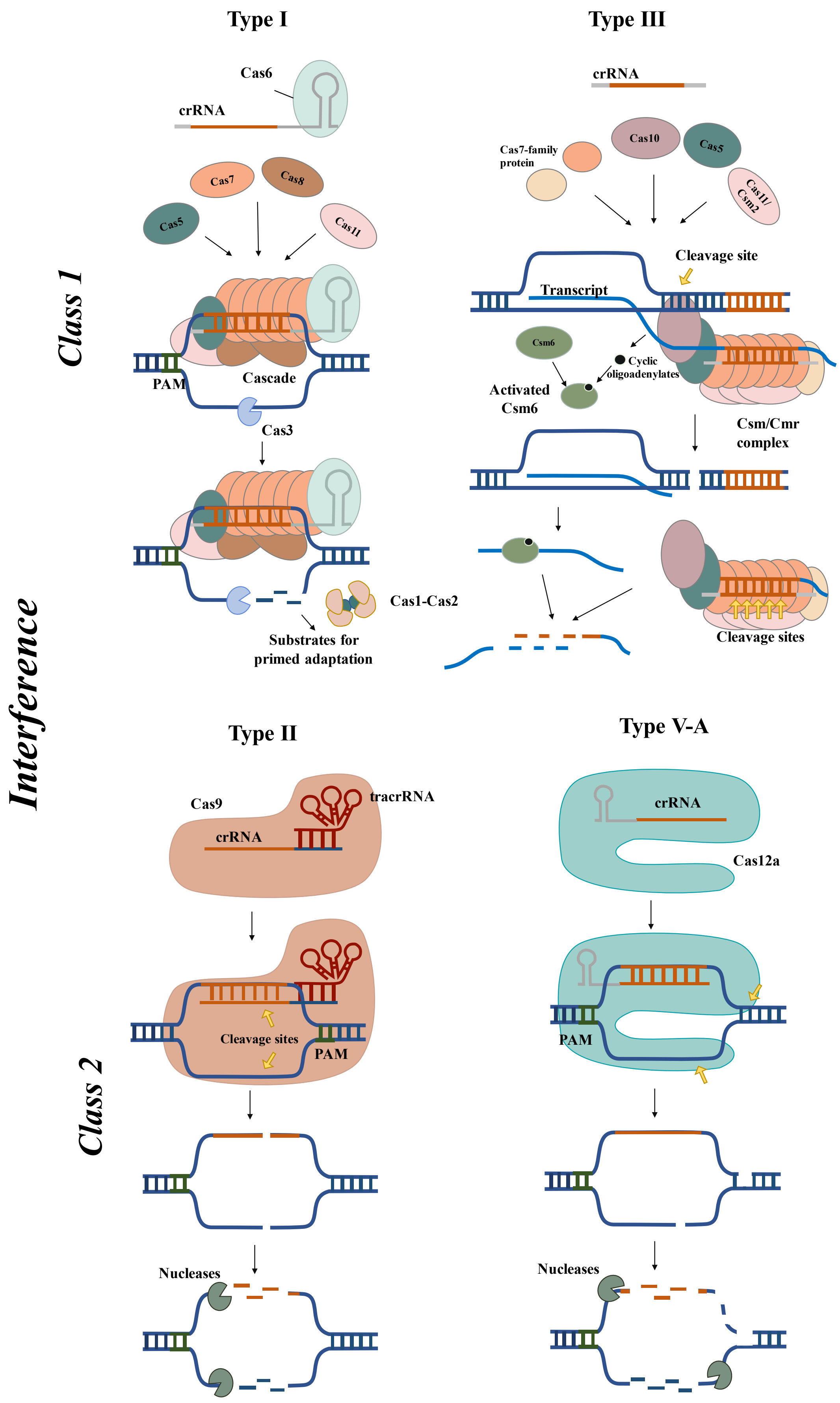

Interference
Class 1
Type I
Cas6
crRNA
Cas7
Cas8
Cas5
Cas11
PAM
Cascade
Cas3
Cas1-Cas2
Substrates for primed adaptation
Type III
crRNA
Cas10
Cas5
Cas7-family protein
Cas11/ Csm2
Cleavage site
Transcript
Csm6
Cyclic oligoadenylates
Activated Csm6
Csm/Cmr complex
Cleavage sites
Class 2
Type II
Cas9
tracrRNA
crRNA
Cleavage sites
PAM
Nucleases
Type V-A
crRNA
Cas12a
PAM
Nucleases

Figure 1.9 Interference in CRISPR-Cas system. Diverse Cas genes are involved during interference stage of two classes. In class1, multiple Cas genes are bonded with crRNA and forms Cascade (or Csm/Cmr) complex. In type I, crRNA with a spacer can quickly combine with target complementary sequences and the dissociative ssDNA cleaved by Cas3. However, type III can target both DNA and RNA sequences. After attaching the target, Cas10 in Csm or Cmr initiate breaks on both stands and activating Csm6 RNA cleavage by mediating cyclic oligoadenylates (cOAs) production. The complementary RNA transcript is abolished by Cas7 in complex. By contrast, interference in class2 is relatively simplified. Individual multi-domain effector remains connected with crRNA (and tracrRNAs in type II) since expression stage. Then, target sites are cleaved by nucleases domains like HNH and RuvC in Cas9 (type II) and RuvC in Cas12a (type V-A). Adapted from Hille et al. (2018).

In class 1 systems, multiple Cas proteins normally would unite together as one large effector complex. The complex works with crRNA and other nuclease in invader binding and annihilation. In type I, the functional machinery is Cascade complex with nuclease Cas3, which was briefly described in expression section 1.3.2.2. Cascades in all subtypes of type I are largely conserved but with minor differences (Makarova et al., 2015). Here, we will mainly introduce Cascade in subtype I-E because it is the most well-studied model (Figure 1.9). Cas6 that works in expression stage remains binding with crRNA, and sequentially serves as a scaffold for assembly of 1 Cas5, 6 Cas7, Cas8 and 2 Cas11, which forms a Cascade in a seahorse-like shape (Van Der Oost et al., 2014; Semenova et al., 2011). Cas5, Cas6 and Cas7 all belong to repeat-associated mysterious proteins (RAMPs) that can bind RNA sequences (Wang and Li, 2012). Multiple Cas7 proteins form the backbone of Cascade with Cas5 at the 5’ end. Meanwhile, Cas6 on the other side is bound with a repeat hairpin (not necessary for non-structured repeat) (Jackson et al., 2014). At the beginning of interference, PAM in the foreign genetic element is recognised by a large subunit (LS) Cas8 and then dsDNA is unwound for crRNA-protospacer binding (Hayes et al., 2016). In this process, a seven-nucleotide region (first eight nucleotides except the sixth, termed the seed sequence) flanking to PAM was found play a key role in protospacer base pairing. Mutations in the seed sequence significantly abolish foreign elements target compared to mutations outside this region, which indicates its function in mediating efficient initial protospacer scanning (Semenova et al., 2011). After target binding, an R-loop including double-stranded DNA-crRNA hybrid and a non-target ssDNA is formed and stabilized by the two Cas11 small subunits (SS) (Mulepati et al., 2014). At the final step, the non-target stand is bulged for single-stranded nuclease Cas3 cleavage (Figure 1.9) (Redding et al., 2015; Xiao et al., 2017).

Interference machinery in type III system can target both DNA sequence and RNA transcript. Type III-A and III-B also employed similar complexes like Cascade for target destruction, which are known as Csm and Cmr, respectively (Figure 1.9) (Hochstrasser et al., 2014; Jackson et al., 2014). In the complex, Cas5 binds the 5' end of mature crRNA and multiple Cas7 family proteins form the backbone. Cas11 and Cas10 are defined as small and large subunits, respectively (Osawa et al., 2015; Taylor et al., 2015). Interference starts with effector complex binding with the nascent transcript that possesses a cognate repeat segment. While the RNA transcript is degraded by Cas7 subunits, Cas10 also performs a DSB of the invader DNA (Figure 1.9) (Osawa et al., 2015; Staals et al., 2014). A recent finding has subsequently revealed a second role played by Cas10 during interference. It can help generate cyclic oligoadenylates (cOAs), which can active RNase (belonging to Csm6) to cleavage RNAs non-specifically (Kazlauskiene et al., 2017). Though PAM is not involved in the mechanism of self versus non-self discrimination in most type III systems, a motif called RNA-PAM that adjacent to target RNA seems crucial for recognition and degradation in type III-B (Elmore et al., 2016; Marraffini and Sontheimer, 2010).

In contrast to the class 1 system which used multi-subunit effector complexes as interference machinery, class 2 prefers single multi-domain proteins that cooperate with crRNA. The use of simple structured proteins by class 2 means it is favoured in the extensive application in the genome-editing field (Mali et al., 2013b). Cas9, Cas12 and Cas13 are the featured effector proteins of type II, type V and type VI in class 2, respectively (Hille et al., 2018). They operate from the expression stage (section 1.3.2.2) and sequentially bind with crRNA until cleaving targets. The most well-characterised interference mechanism of class 2 is in type II system in which the duplex tracrRNAs:crRNA unites with Cas9 and form a ribonucleoprotein complex (Figure 1.9) (Mir et al., 2018). A PAM that is located downstream of the target sequence assists the probe for complementary seed sequences. This results in base-pairing between crRNA and the target strand in an R-loop structure (Jinek et al., 2012; Sternberg et al., 2014). HNH and RuvC domains in Cas9 who have nuclease activities are then triggered to produce blunt breaks in both strands (Jinek et al., 2014; Yamada et al., 2017). Similar multifunction genes were also identified in type V system: Cas12a, Cas12b and Cas12c for type V-A, V-B and V-C, respectively (Shmakov et al., 2015). Unlike Cas9 and Cas12b, the activity of Cas12a does not associate with tracrRNAs (Shmakov et al., 2015). In type V-A, target sequence binds with crRNA from a T-rich PAM recognition and then both strands are degraded by the RuvC domain in Cas12a (Swarts et al., 2017). However, the Ruv-like domain is missing in Cas13 and its activity does not require tracrRNAs either (Abudayyeh et al., 2016; Shmakov et al., 2015). In Cas13, two higher eukaryotes and prokaryotes nucleotide (HEPN)-containing domains act as RNases to destroy the target. After crRNA binding with complementary single-stranded RNA (ssRNA) transcripts, the machinery is activated and a global degradation of all exposed RNA including RNA transcripts from the invader is carried out (Liu et al., 2017).

Intriguingly, the interference module is not restricted to the combination with crRNA from its adjacent CRISPR array. In *Marinomonas mediterranea*, crRNAs from type I-F system can guide the interference machinery from type III-B system to destroy segments from invaders (Silas et al., 2017b). This cooperation provides dual protection for the host. Phages that possess mutant PAM escape defence from type I-F but can still be captured by type III-B since the interference complex of the latter system does not require specific PAM (Marraffini and Sontheimer, 2010) and is more tolerant of protospacer mutations (Staals et al., 2014).

1.3.3 Arms Race between CRISPR-Cas Systems and Phages

Spacers in CRISPR loci provide specific protection to the host in defending a range of MGEs. However, a comparative analysis has reported that the majority of known spacers lack traceable origins, which is known as "dark matter" (Shmakov et al., 2017b). The rest of the spacers that have recognisable protospacers are mostly mapped to phages (80% to 90% based on different subtypes) (Shmakov et al., 2017a, 2017b). This constructs a co-evolution relationship between prokaryote and phages (Westra et al., 2016). The special "cut and paste" mechanism of CRISPR-Cas system determines its high specificity during immunization. Correspondingly, phages also encode several counter-defence strategies against the immunity from CRISPR-Cas. In this section, we are going to briefly introduce several strategies that phages employ during the constant arms race with CRISPR-Cas.

First, phages can circumvent target from CRISPR-Cas through modifying self-genomes such as mutation and deletion. A co-evolution study has revealed that even a single mutation in target sequences (like protospacer) can help avoid cleavage (Heidelberg et al., 2009; Semenova et al., 2011). Additionally, mutations are more likely happen in phage PAM, which avoid annihilation from those CRISPR-Cas subtypes that more tolerant with mutations (Deveau et al., 2008; Pyenson et al., 2017; Sun et al., 2013). As mentioned before (section 1.3.2.1), CRISPR-Cas triggers primed adaptation when cleaving known phages, especially with mutated targeted regions. Under this selective pressure, corresponding segments were observed deleted in phage genome in advantage from escaping capture by crRNA (Watson et al., 2019). Moreover, genetic rearrangement was noticed in phage genomes as well. Multiple phages recombined and formed a mosaic genome, in which the targeted region was exchanged and phages resistance was raised (Paez-Espino et al., 2015; Westra et al., 2016). Under pressure from CRISPR-Cas immunity, point mutations and phages segments rearrangement were also observed could accrue in order to enhance the survival rate (Paez-Espino et al., 2015).

Aside from genome modification, some phages specifically encode a protein called anti-CRISPR (Acr) that interacts directly with interference machinery, including the effector complex or nuclease, and abolishes CRISPR-Cas immunization (Borges et al., 2017; Hampton et al., 2019; Hille et al., 2018). Acr has been identified from both temperate and lytic phages (Borges et al., 2017; Hynes et al., 2017) and target a wide range of CRISPR-Cas subtypes including I, II, III, V and VI in Bacteria (Borges et al., 2017; Hwang and Maxwell, 2019) and type I and III in Archaea (Bhoobalan-Chitty et al., 2019; Peng et al., 2020). Diverse Acr proteins are highly specific to different CRISPR-Cas subtypes. In type I-F, multiple Acr proteins were identified to prevent CRISPR-Cas. AcrF1, AcrF2 and AcrF10 bind subunits of Cascade in order to impede target DNA binding, while AcrF3 directly circumvents cleavage from the Cas3 nuclease (Bondy-Denomy et al., 2015; Guo et al., 2017). Similar inhibition activity was also noticed in type II. AcrIIA4 blocks Cas9 activity in many-fold ways. Firstly, it combines with PAM-interaction domains to abolish DNA binding (Basgall et al., 2018; Dong et al., 2017; Yang and Patel, 2017). Next, it associates with the DNA-melting region to prevent DNA binding and unwinding (Trasanidou et al., 2019). Finally, it interacts with the RuvC domain and the bond that connected RuvC and HNH domain, which repress DNA cleavage sterically (Dong et al., 2017; Shin et al., 2017). The multi-angle regulation by AcrII4 also suggests its potential application in manipulating CRISPR-Cas9 gene editing (Sontheimer and Davidson, 2017). Most discoveries of the Acr functions focussed on interactions with interference machinery and its role in adaptation and expression remains an open question (Hille et al., 2018).

Intriguingly, CRISPR-Cas systems also can be encoded by phage, which serves as an anti-defence counteract against the host genome (Faure et al., 2019; Hampton et al., 2019; Hille et al., 2018). A complete and functional I-F system was identified in *Vibrio cholerae* phage, in which spacers in phage CRISPR loci can target host genomic islands. These islands are normally dominant in prokaryotic phages inhibition (Seed et al., 2013). Moreover, many Cas gene clusters in phages are incomplete and harbour short CRISPR arrays (Faure et al., 2019; Hampton et al., 2019). The range of spacers in phage CRISPR loci covers not only bacterial genomes but also other phages, which hypothetically indicates the role of phages encoded CRISPR-Cas in virus-virus competition (Faure et al., 2019).

1.3.4 Origin and Evolution of CRISPR-Cas Systems

The CRISPR-Cas machinery can be briefly divided into adaptation and effector modules from the perspectives of structure and function (Makarova et al., 2015). As introduced in section 1.3.2, adaptation modules in both classes mainly involve Cas1 and Cas2 proteins whereas effector modules are relatively more diverse. In particular, MGE not only participate in the co-evolution of CRISPR-Cas, but have also made vital contributions to its origin and subsequent diversification multiple times (Koonin and Krupovic, 2015; Mohanraju et al., 2016; Shmakov et al., 2017a).

The Cas1 endonuclease in the adaptation module is thought to be derived from a superfamily of self-synthesizing DNA transposons called casposons (Krupovic et al., 2014). During transposon integration, casposon has been found to employ the transposase activity of Cas1 homologous. This process is very similar to spacer integration and thus casposon is predicted, and experimentally demonstrated, to be relevant to the origin of CRISPR-Cas adaptation module (Krupovic et al., 2014, 2016). Cas2 is thought to have originated from the insertion of a toxin-antitoxin module into an ancestral casposon, which completes the emergence of adaptation module (Mohanraju et al., 2016).

Compared to the adaptation module, the understanding about the origin of effector module is less evident. A recent study revealed the process of effector complex emergency in class 1 using type III system as an ancestral candidate (Koonin and Makarova, 2019). They hypothesized that the ancestor of CRISPR-Cas is a putative signalling system (likely to be an abortive infection module) that was comprised of a Cas10 homologous, CRISPR-associated Rossmann fold (CARF) domain and HEPN domain. This theory is compatible with the previously presented biochemical mechanisms (Kazlauskiene et al., 2017; Niewoehner et al., 2017). cOA molecules that synthesized by effector complex Cas10 can bind Csm6 from CARF domain and activate the Csm6's RNase activity respect to HERP domain. This RNase activity of HERP domain can abolish RNA sequences indiscriminately and potentially regulate dormancy or programmed cell death. Based on the cOA-Cas10 signalling interaction, Koonin & Makarova (2018) proposed that the effector module of CRISPR-Cas may have initialised from a stress-response system that is mediated by cOA.

Class1 CRISPR-Cas system employs multiple Cas proteins in interference machinery and is regarded as the ancestral system considering the widespread across prokaryotes (Koonin and Makarova, 2019). Effector module in class2 requires a single, multiple domain Cas protein. The detailed evolutionary process is still not completely understood, but comparative analysis has revealed that the origin of type II and type V is due to random insertion of TnpB-coding transposons (Shmakov et al., 2017a).

1.3.5 Horizontal Gene Transfer (HGT) and CRISPR-Cas Systems

HGT, including transformation, conjugation and transduction frequently occurs between prokaryotes and has been identified to have affected CRISPR-Cas loci through phylogenetic and comparative analyses (Godde and Bickerton, 2006; Makarova et al., 2013, 2015). Complete CRISPR-Cas systems have also been found in different classes of MGEs, such as plasmids, viruses and transposons (Faure et al., 2019; Newire et al., 2020; Peters et al., 2017; Seed et al., 2013), indicating their mobility between organisms. HGT is thought to contribute to the widespread distribution of CRISPR-Cas systems in Archaea and Bacteria (Makarova et al., 2015).

CRISPR-Cas systems can also affect HGT through several mechanisms. For example, a spacer in the CRISPR loci of *Staphylococcus epidermidis* complements a *nickase* gene that is widespread in staphylococcal conjugative plasmids (Marraffini and Sontheimer, 2008). When a plasmid carrying this gene enters a cell, it is recognised as invading DNA, and its acquisition by the host cell is prevented by the

CRISPR-Cas system. Homologies between spacers and integrative conjugative elements have also been noted in *Pseudomonas aeruginosa* (Wheatley and Maclean, 2020). From a pneumococcal virulence study of *Streptococcus pneumoniae*, natural transformation has also been found to be constrained by CRISPR-Cas, and additionally this study stressed the fact that the environment can drive losses of CRISPR-Cas (Bikard et al., 2012). In addition, a comparative analysis found that species with active CRISPR-Cas enclosed less MGEs and have a narrower spread in environments, which indicated the inhibition force exerted by CRISPR-Cas in order to prevent HGT (Zheng et al., 2020). Consistently, Wheatley & Maclean (2020) proposed a similar hypothesis that species without CRISPR-Cas or those that harboured inactive CRISPR-Cas are better at adapting to new niches through acquiring beneficial genes such as antibiotic resistance genes (ARGs) by HGT. Moreover, another independent study showed the association between *acr* genes and ARGs in *P. aeruginosa* (Shehreen et al., 2019). They hypothesized that *acr* genes in the host that were obtained through HGT might inhibit the activity of CRISPR-Cas, consequently promoting the gains of ARGs. In contrast, HGT by transduction can be enhanced by the presence of short segments in the host genome that match the invasive DNA (Varble et al., 2019; Watson et al., 2018). Interestingly, there does not seem to be a correlation between CRISPR activity (measured by the length of CRISPR arrays) and the extent of HGT on evolutionary timescales (Gophna et al., 2015). From the perspective of long-term evolution, the inhibitory effect of CRISPR-Cas systems on HGT might therefore be balanced by promoting transduction or other mechanisms (Faure et al., 2019).

HGT of CRISPR-Cas is pervasive and the inhibition of CRISPR-Cas to HGT were not significant in long timescale (Bikard et al., 2012; Gophna et al., 2015; Marraffini and Sontheimer, 2008). Therefore, Bernheim & Sorek (2020) proposed a model called the “pan-immune system” that describes how closely related organisms carry different defence mechanisms, giving the bacterial community the ability to resist a large range of invaders. In this model, HGT plays a key role in transferring and maintaining defence systems, when the transfer between species bears low fitness costs. Aided by HGT, defence systems can be regarded as public goods, collectively defending closely related organisms. Correspondingly, phages will mutate constantly to escape bacterial resistance systems (Paez-Espino et al., 2015), which drives rapid and constant co-evolution between phages and prokaryotes (Faure et al., 2019).

1.3.6 Objectives of This Study

Although phylogenetic approaches have been employed as the principal evolutionary tools for more than a century, many complex patterns such as hybridization, HGT, gene fusion cannot be accurately and thoughtfully presented solely by the tree-based model. Gene fusion and HGT have contributed to the origin of many organisms and these processes are still ongoing. To study these patterns, we employed network-based approaches such as SSNs. We hypothesized that networks had great potential in handling large datasets and have applied network thinking in studies about composite genes and CRISPR-Cas systems.

One of the main aims of this study is to assess the power of network approaches. Different large datasets were constructed according to the research objectives. To study composite genes in the three domains of life as well as MGEs, we constructed a dataset containing 182 complete eukaryotic and prokaryotic genomes, 79 viruses and 1614 plasmids. The main objectives and findings for studying composite genes (Chapter 2) are as follows:

(1) I aimed to identify composite genes and parental component genes using SSNs and identified 221,043 (18.57%) through construing networks in program CompositeSearch.

(2) I aimed to investigate the distribution of composite genes in different species.

(3) I aimed to figure out whether composite genes are biased in their distribution, and to explore the underlying reasons through functional annotations. Through Odd Ratios test, composite genes are more likely to derive from eukaryotes rather than prokaryotes in most COG categories.

The other research objects in this thesis relate to CRISPR-Cas systems, which show complex evolutionary patterns within and across prokaryotic organisms. To comprehensively understand the evolution of this system, we collected all available prokaryotic nucleotide genomes from the NCBI RefSeq database and used different approaches, including comparative analysis, phylogenies and networks. The objectives and finding of studying CRISPR-Cas are as follows:

(1) I aimed to compare different CRISPR-Cas identification tools from the perspective of bioinformatics using the same dataset (Chapter 3).

(2) I aimed to investigate the spacer evolution in CRISPR loci and analyse the effects of insertion, deletion, and recombination (Chapter 3). Diverse evolutionary processes were observed in spacers.

(3) I aimed to examine the "pan-immune model" hypothesized by Bernheim and Sorek (2020) and found pervasive repeat sharing and a small number of spacer sharing between species using network and phylogenetic approaches (Chapter 3).

(4) I aimed to analyse the distribution of CRISPR-Cas system and demonstrated related genetic backgrounds through a co-occurrence study (Chapter 4).

(5) I aimed to explore CRISPR-Cas chemical mechanisms through combinations of a network model and a co-occurrence study (Chapter 4).

Chapter 2.

Eukaryote Genes Are More Likely than Prokaryote Genes to be Composites

2.1 Abstract

The formation of new genes by combining parts of existing genes is an important evolutionary process. Remodelled genes, which we call composites, have been investigated in many species, however, their distribution across all of life is still unknown. We set out to examine the extent to which genomes from cells and mobile genetic elements contain composite genes. We identify composite genes as those that show partial homology to at least two unrelated component genes. In order to identify composite and component genes, we constructed sequence similarity networks (SSNs) of more than one million genes from all three domains of life, as well as viruses and plasmids. We identified non-transitive triplets of nodes in this network and explored the homology relationships in these triplets to see if the middle nodes were indeed composite genes. In total, we identified 221,043 (18.57%) composites genes, which were distributed across all genomic and functional categories. In particular, the presence of composite genes is statistically more likely in eukaryotes than prokaryotes.

2.2 Introduction

Reticulation occurs when two or more evolutionary lineages merge, and consequently, reticulation cannot be visualised or analysed using tree-like models of evolution. We see reticulate events occurring during meiotic recombination, horizontal gene transfer (HGT, also known as lateral gene transfer) (Nelson-Sathi et al., 2012), exon shuffling (Oakley, 2017), and hybrid speciation (Linder et al., 2004) for example. Merger events can be seen at multiple levels, such as genes, genomes, operons and gene clusters.

This paper focuses on the combination of genetic fragments from unrelated gene families to produce a single gene. This process of gene fusion occurs when parental (or component) genes merge to form a new gene called a composite (or fused) gene (Corel et al., 2016; Oakley, 2017). Because reticulate evolution cannot be adequately represented using tree-like representations, we constructed sequence similarity networks (SSNs, also known as protein/gene similarity networks) and visualised them using Gephi (Bastian et al., 2009) and Cytoscape (Shannon et al., 2003). In these kinds of networks, gene, genome or species data can be used to detect recombination events. In the SSNs that we have constructed, genes or proteins are represented as nodes while inferences of homology between genes are represented by edges. Within the framework of the SSN, some special relationships, such as non-transitive triplets when two component genes have no overlap, can be identified as motifs in the network. SSNs have been used elsewhere in order to investigate the existence of composite genes (Coleman et al., 2015; Haggerty et al., 2014). In an analysis of 15 eukaryotic genomes, Haggerty et al. (2014) constructed a network that contained a giant connected component (GCC) where one quarter of all sequences were identified as composite genes and approximately 10% of sequences were identified as multi-composite genes (those formed from the union of two or more composite genes). Moreover, Coleman et al. (2015) used SSNs to explore 1642 antibiotic resistance genes derived from more than 100 species. They found 73 fused genes using the FusedTriplets software (Enright et al., 1999; Jachiet et al., 2013), which accounted for 4.43% of the total gene count. In addition, Jachiet et al. (2013) using the MosaicFinder software, found gene fusions in both cellular organisms and mobile genetic elements (MGEs). In another analysis using the same kind of approach, viruses were suggested to consist of only 8–15% of composite genes, with this low number being attributed to the blurry boundaries between viral gene families (Jachiet et al., 2014). In addition, gene fusion has been shown to have played an essential role in the evolution of the cellular life cycle, with composite gene formation seen in genes related to chromatin structure and nucleotide metabolism (Méheust et al., 2016). Also, Ocaña-Pallarès et al. (2019) concluded that there was a significant role for gene fusion in the origin of eukaryotes, as evidenced by SSN built from eukaryotic EUKaryotic restricted Nitrate Reductase (EUKNR) and similar eukaryotic and prokaryotic sequences. The result indicated that EUKNR was formed by a fusion of eukaryotic sulfite oxidases (SUOX, N-terminal) and NADH (C-terminal) reductases. Therefore, while it is clear that gene fusion is a common feature of genes, a comprehensive comparison across a broad range of taxa and molecule types would provide more evidence for its frequency and impact.

In this paper, we describe an approach to identify composite genes using a dataset of 1875 completed genomes, comprising more than one million sequences, from all three domains of life as well as from MGEs. We tested whether the rate of gene remodelling has been uniform across all of life, and all cellular functional categories.

2.3 Materials and Methods

2.3.1 Dataset Construction and BLAST Analysis

A total of 1,190,265 protein sequences were collected from the RefSeq database at the National Centre for Biotechnology Information (Pruitt et al., 2006). We manually selected taxa in order to maximise diversity, while also ensuring computational tractability. The final dataset covered 182 species from the main representative lineages, belonging to 36 eukaryotes (13 phyla, 21 classes), 56 archaea (4 phyla, 9 classes), 90 bacteria (25 phyla, 32 classes), 79 viruses and 1,614 plasmids. Homology between pairs of amino acid sequences was inferred using an all-versus-all protein BLAST (BLASTP version 2.4.0, NCBI, Bethesda, MD, United States), with an E-value cutoff of 1e-5, 5000 max target sequences, and soft masking parameter (the others by default) (Altschul et al., 1997). The dataset species information and download paths are available at https://github.com/JMcInerneyLab/CompositeGenes/blob/master/accession.txt.

2.3.2 Composite Gene Identification

Using the BLAST results as input, we identified composite genes as motifs of triplets in the graph where there was a “non-transitive” relationship between three nodes (Corel et al., 2016). Composite gene detection was carried out by the CompositeSearch program (Pathmanathan et al., 2017) when associated component genes have no overlap theoretically, with default identity cutoff of 30% and 20 amino acid overlaps to limit false negative error. The CompositeSearch output contains information on composite genes, component genes and the families to which they belong. This output was depicted, explored, and manipulated using Gephi (version 0.9.2, The Gephi Consortium, Paris, France) (Bastian et al., 2009).

Because the proportion of composite gene from different domains might be affected by biased sequence database sampling, we randomly sampled 50,000 protein-coding genes from archaea, bacteria, eukaryotes and plasmids respectively. These random samples were taken forward for analysis in the same way as the original data. The major difference between the subsampled datasets and the original data was that in the subsampled datasets, the number of genes from each of the four kinds of dataset was the same. We used CompositeSearch in order to construct an SSN from the BLASTP output of the subsampled datasets containing 200,000 genes. These SSNs were then used in order to identify composite genes. Sampling was repeated 100 times and the results were summarised graphically.

2.3.3 Functional Annotations

We used EggNOG (version 4.5.1, Computational Biology Group–EMBL, Heidelberg, Germany) (Huerta-Cepas et al., 2015) in order to assign gene functional categories. The analysis was carried out through the web interface using the DIAMOND (Buchfink et al., 2015) mapping mode. In the output, genes were assigned to different Orthologous Groups (OGs), and each OG had functional annotations that included Clusters of Orthologous Groups (COGs) functional categories: COG for universal Bacteria, EuKaryotic Orthologous Groups (KOGs) for Eukaryotes and arKOGs for Archaea (Tatusov, R. L., Galperin, M.Y., Natale, D.A., Koonin, 2000); Gene Ontology (GO) terms (Consortium, 2004); Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and SMART/Pfam protein domains. Both composite and non-composite genes were placed into at least one of 23 COG categories and at least one of four GO terms.

2.3.4 Statistical Analysis

In the EggNOG output, each gene has a detailed functional annotation and is associated with at least one general COG category code (A to Z apart from R and X). Because of recombination, the category code for a given gene could be single letter like “A” or multiple letters such as “ABC”. When counting the number of genes that possess a particular function, if a multiple letter category was selected, we counted this gene multiple times. For instance, if the most common COG category for a gene was "ABC", and then this gene was counted three times as A, B and C.

To investigate the distribution of composite genes and non-composite genes among eukaryotes and prokaryotes, an odds ratio (OR) test (Szumilas, 2010) was carried out. OR tests are normally used to test the strength of the association between two events. Here, for each protein function, we used an OR test to test the association between gene origin and the likelihood of fusion (See Equation 1).

$$OR = \frac{a/c}{b/d} = \frac{ad}{bc} \tag{1}$$

where *a* is the number of composite eukaryote genes, *b* is the number of non-composite eukaryote genes, *c* is the number of composite prokaryote genes, *d* is the number of non-composite prokaryote genes. The 95% confidence intervals (CI) were calculated by

$$Upper\ 95\%\ CI = e^\wedge\left[\ln(OR) + 1.96\sqrt{\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right)}\right] \tag{2}$$

$$Lower\ 95\%\ CI = e^\wedge\left[\ln(OR) - 1.96\sqrt{\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right)}\right]$$

Considering all OR tests that were carried out on 24 COG categories, simultaneous tests are likely along with false statistical inference (Chen et al., 2017). Therefore, we used a conservative Bonferroni correction (Sedgwick, 2012) to limit type I error. Bonferroni correction is method that has been widely used for multiple comparisons to adjust p-values through controlling the familywise error rate (Bland and Altman, 1995). Since the size of our tests is moderate, the conservativeness of Bonferroni adjustment should be tolerant. The critical level of significance was initially set as α = 5%, we corrected it as α/2N, N is the number of performed tests, which in our case is 24. The new significance level is 0.1% and corresponding confidence coefficient of 99.9% is 3.09 standard deviations, using the standard normal distribution table. The corrected CI was calculated by

$$Upper\ 95\%\ CI = e^\wedge\left[\ln(OR) + 3.09\sqrt{\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right)}\right]$$

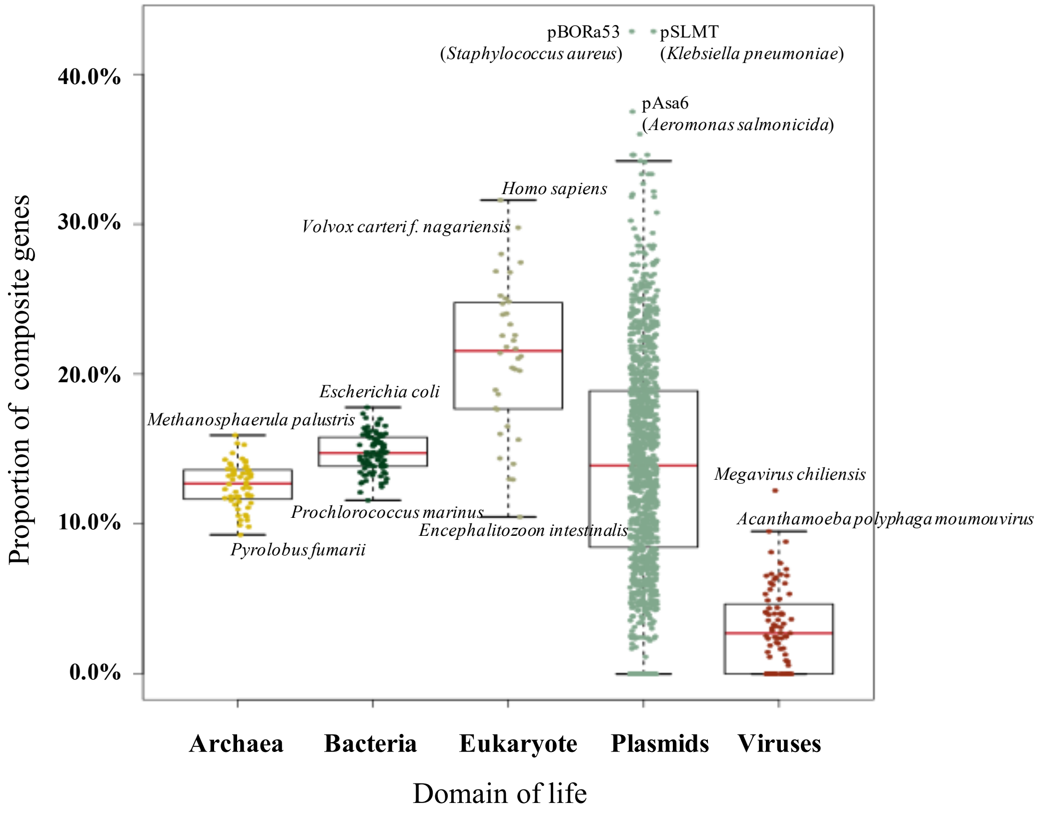
$$Lower\ 95\%\ CI = e^\wedge\left[\ln(OR) - 3.09\sqrt{\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right)}\right]$$

(3)

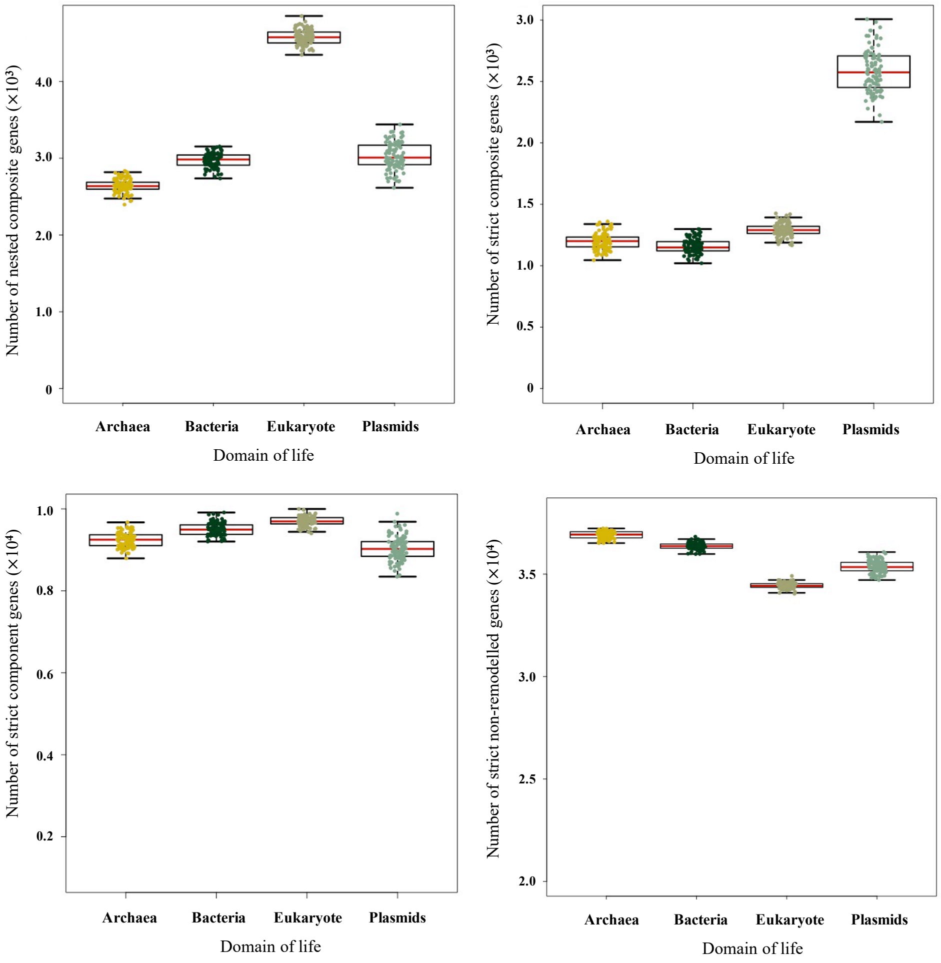
2.4 Results

2.4.1 Pervasive Existence of Composite Genes across All of Life

We assembled a dataset of 1,190,256 genes from 36 eukaryotes, 56 archaea, 90 bacteria, 79 viruses and 1614 plasmids from more than 60 taxonomic classes. Following an all-versus-all BLAST, a total of 540,325,758 significant hits were detected. Using CompositeSearch, an SSN containing 1,025,263 nodes and 109,650,422 edges was constructed. In this network, 221,043 composite genes (18.57% of the gene dataset, Figure 2.1a) were identified, linked to 603,604 component genes. Collectively, these genes were assigned to 360,981 gene families.



(a)



(b)

Figure 2.1 Pictures of composite and component genes among different domains. (a) Proportion of composite genes within different domains and mobile genetic elements. Dots represent individual genomes. **(b)** Numbers of nested composite, strict composite, strict component, and non-remodelled genes within different domains for each of the 100 replicates of equal sampling. All analyses were replicated 100 times and each replicate is represented by a dot.

To gain a better understanding of those genes involved in non-homologous recombination, all genes were categorized into four groups: nested composite genes, strict composite genes, strict component genes, and non-remodelled genes (Figure 2.2). Nested composite genes have been formed by the merging of at least two sequences but are additionally involved in other non-transitive triplets as components; that is to say they themselves are composites but also form other composites. In contrast to nested composites, strict composite genes only act as composite genes in the network, similar to strict component genes. Non-remodelled genes do not show evidence of having participated in any recombination events. In our dataset, 181,157 genes as nested composite genes, 39,886 genes were identified as strict composite genes, 422,447 as strict component genes, and 546,766 as non-remodelled genes (Figure 2.3a).

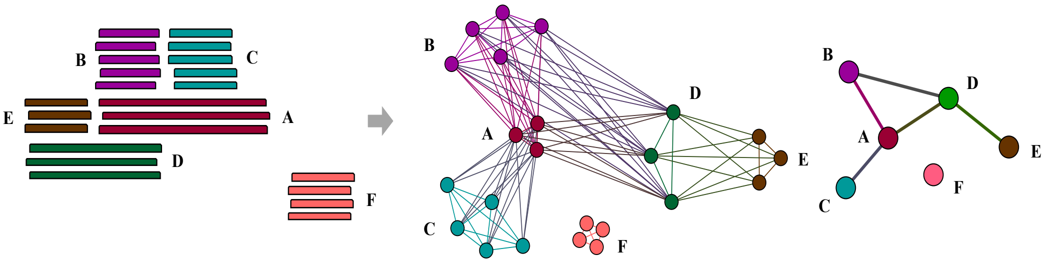


Figure 2.2 Sample patterns of nested composite, strict composite, strict component, and non-remodelled gene families. Genes in family A not only participate in the fusion of genes in family D as component genes but are also formed by genes from family B and C as composite genes; this is regarded as nested composite genes. In contrast, genes in family B, C and E belong to strict component families which only act as component genes in this network. Similarly, genes in family D as members of a strict composite family. In additional, family F is non-remodelled gene. Also, because there is no overlap between gene family A and C, gene family B and E so "A-B-C" and "B-D-E" can be regarded as non-transitive triplets.

Within 182 species across the three domains of life, remodelled composite genes were discovered in all species, indicating that gene fusion is, and has been, widespread across all life on Earth. Overall, 23.66% (205,913 composites identified from 870,120 eukaryotic and prokaryotic genes) of examined genes were identified as composite. However, there was a considerable amount of variation in the proportion of composite genes across species and molecule type. Table 2.1 presents the ten genomes with the highest and lowest rates of composite genes among eukaryotes and prokaryotes. Composite genes account for almost one third of the genomes of *Homo sapiens*, *Volvox carteri* f. *nagariensis* and *Aureococcus anophagefferens*.

Table 2.1 The ten species that contain the highest and lowest proportions of composite genes. The ten species that contain the highest (left) and lowest (right) proportions of composite genes. All species that contains more than 24% composite genes are from eukaryotes, whereas most species that contain less than 11% composite genes are from archaea (Crenarchaeota family, mostly).

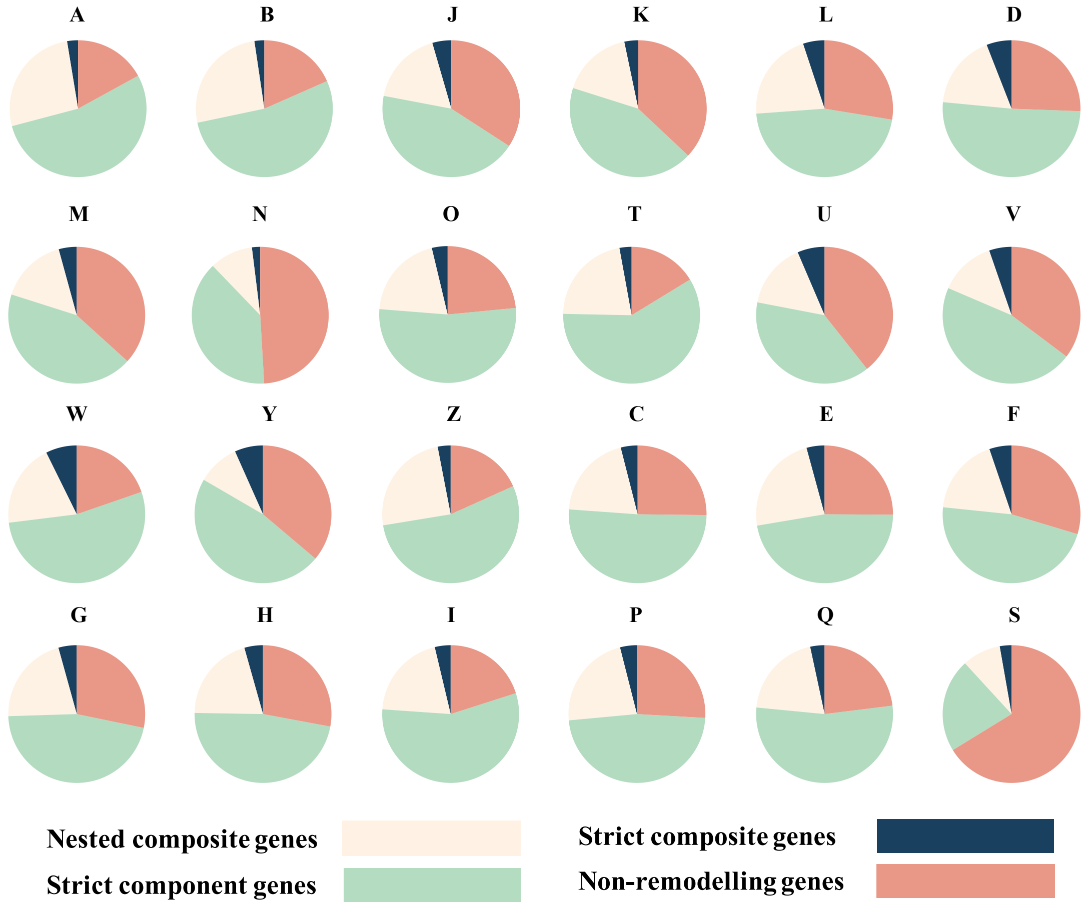
	Species	**Total Number of Genes**	**Number of Composite Genes**	**Proportion**
Ten species with highest proportions of composite genes	*Homo sapiens*	109018	34455	31.60%
	Volvox carteri f. *nagariensis*	14436	4298	29.77%
	Aureococcus anophagefferens	11520	3227	28.01%
	Capsaspora owczarzaki	8792	2413	27.45%
	Chlorella variabilis	9780	2626	26.85%
	Polysphondylium pallidum	12367	3313	26.79%
	Monosiga brevicollis	9203	2322	25.23%
	Salpingoeca rosetta	11731	2939	25.05%
	Allomyces macrogynus	19446	4829	24.83%
	Tetrahymena thermophila	10626	2625	24.70%
Ten species with lowest proportions of composite genes	*Fervidicoccus fontis*	1385	152	10.97%
	Thermoproteus uzoniensis	2112	224	10.61%
	Nanoarchaeum equitans	540	57	10.56%
	Staphylothermus marinus	1598	168	10.51%
	Encephalitozoon intestinalis	1939	203	10.47%
	Ignisphaera aggregans	1930	198	10.26%
	Methanopyrus kandleri	1687	173	10.25%
	Pyrobaculum neutrophilum	1966	195	9.92%
	Hyperthermus butylicus	1681	165	9.82%
	Pyrolobus fumarii	1885	175	9.28%

As shown in Figure 2.1a, the proportion of composite genes often shows a wide distribution, depending on the classification of the genome in which the gene is found. Among cellular lifeforms, eukaryote genomes contain the highest proportion of composite genes on average (22.66%), followed by bacteria (14.76%) and then archaea (12.78%). However, the distributions are quite wide though prokaryote species manifested a narrower distribution of composite frequency when compared with eukaryotes. When considering mobile genetic elements, the average percentage of composite genes in plasmids (14.69%) is almost the same as bacteria but is noticeably higher than the average seen for virus genes (4.82%).

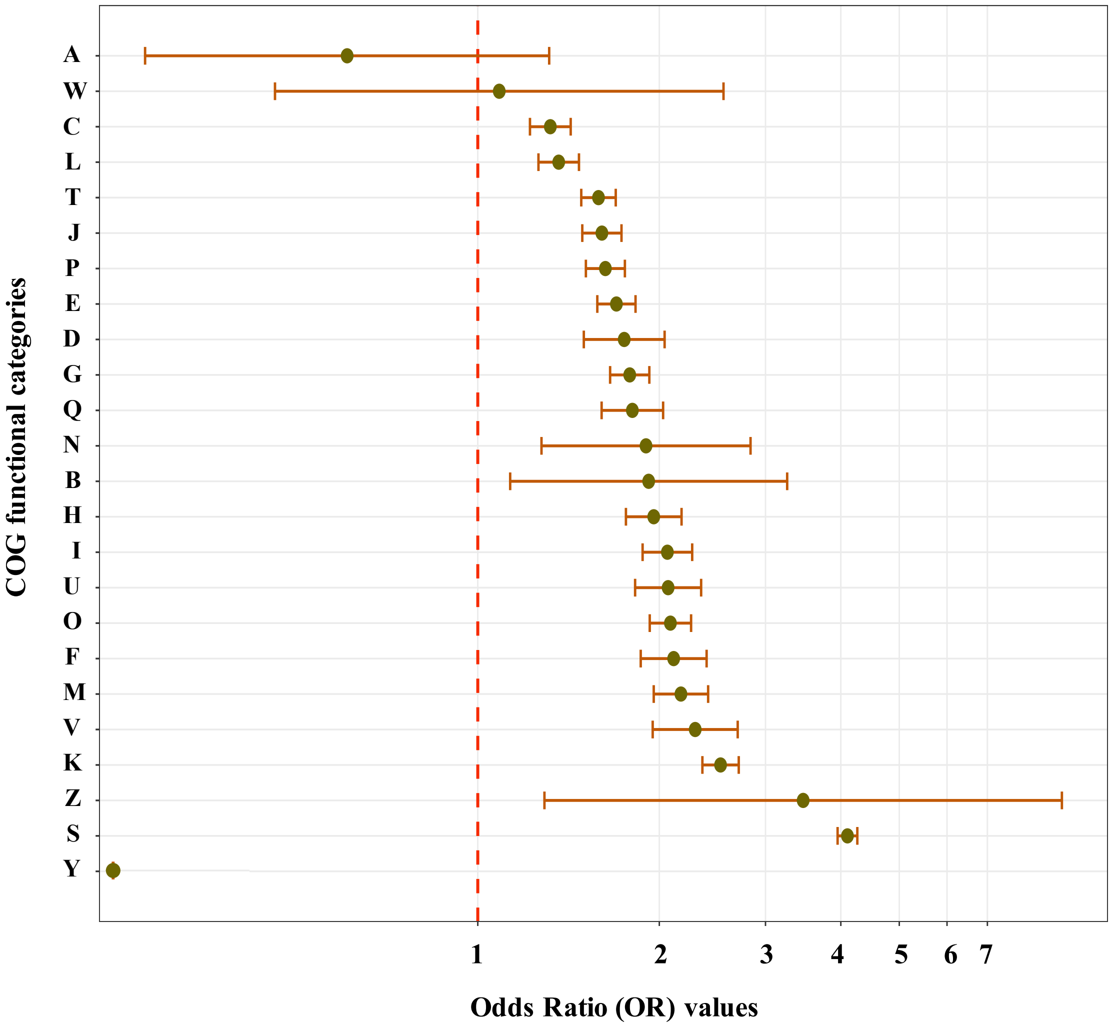
To avoid the effects of unequal sampling in large dataset, we used a jackknife resampling approach in order to generate datasets of 50,000 sequences each from eukaryotes, archaea, bacteria and plasmids. With these uniformly-sized gene sets we used the same analysis methods as for the large dataset: sampling, identifying homologs and constructing SSNs. We then replicated this process 100 times. On average, across all replicates, 19,443 (9.72%) genes were identified as composite genes (Figure 2.1b), which is approximately half the percentage identified from the large dataset (18.57%). The difference indicates that the detection rate of composite genes is related to genomic sequence sampling size and therefore, the reporting of composite genes is always a lower bound for the actual percentage. The resampling procedure was designed to analyse composite gene distribution while attempting to normalise for the difference in data size for each of the four main classifications (eukaryote, bacteria, archaea and plasmids). Plasmids have the highest proportion of strict composite genes while eukaryotes have the largest proportion of nested composite genes (Figure 2.1b). Nonetheless, even though there is no obvious difference between eukaryotes and prokaryotes in terms of the number of nested composite genes, strict composite genes are approximately twice as likely in eukaryotes as in archaea and bacteria. Bacteria and archaea are quite similar, in terms of their proportions, for all four categories of remodelled and non-remodelled genes. Finally, strict component genes do not show much difference across any of our genome types though eukaryotes have the highest number of strict components but the lowest number non-remodelled genes.

2.4.2 Sequence Functional Annotations

The EggNOG mapper program (Huerta-Cepas et al., 2015) was used to assign functions to all sequences. For all results, COG and GO annotations were used to evaluate functional categories. First, composite genes were found to be widespread across all functional categories (Figure 2.3, Appendix A - Figure S2.1). Gene distributions show different patterns across different functions (Figure 2.3a, Appendix A - Table S2.1). Genes with unknown function (category S, 66.23% non-remodelled) are less likely to have been remodelled. The category of genes that have the second-lowest rate of remodelling is cell motility (N, 49.08% non-remodelled). Genes in RNA processing and modification (A, 26.52%) and dynamics (B, 25.96%) had the highest rate of nested composites. Conversely, genes involved in signal transduction (T, 59.05%) tend to have the highest proportion of strict component genes whereas genes involved in extracellular structures (W, 7.3%) are more likely to be strict composite.



(a)



(b)

Figure 2.3 Function analysis of composite and non-composite genes. (a) Numbers of nested composite, strict composite, strict component, and non-remodelled genes across all COG categories. **(b)** Numbers of OR, upper 95% CI and lower 95% CI value (after Bonferroni correction) across all COG functions. The detailed numbers are shown in Appendix A - Table S2.1. There was not composite gene identified from prokaryote in category Y in this dataset, so OR test was not applied. Apart from A and W, which span 1.0, the odds of composite gene presence in all COG categories shows statistically significant tendency in eukaryotes. A: RNA processing and modification; B: chromatin structure and dynamics; C: energy production and conversion; D: cell cycle control and mitosis; E: amino acid metabolism and transport; F: nucleotide metabolism and transport; G: carbohydrate metabolism and transport; H: coenzyme metabolism; I: lipid metabolism; J: translation; K: transcription; L: replication and repair; M: cell wall/membrane/envelope biogenesis; N: Cell motility; O: post-translational modification: protein turnover, chaperone functions; P: Inorganic ion transport and metabolism; Q: secondary metabolites biosynthesis: transport and catabolism; T: signal transduction; U: intracellular trafficking and secretion; V: defence mechanisms; W: extracellular structures; Y: Nuclear structure; Z: cytoskeleton; S: function unknown.

We used an odds ratio (OR) test and Bonferroni correction (see Methods 2.3.4) on composite and non-composite genes from eukaryotes and prokaryotes in different functional categories in order to understand if genes from different classifications were more likely to be remodelled in one or the other. If the OR value and its upper and lower 95% CI value span 1, we take this as evidence that there is no significant difference in composite gene formation between eukaryotes and prokaryotes, and vice versa. If the OR number is greater than 1, this indicates a positive correlation between remodelling and being from a eukaryotic genome, while if the number is less than 1, it indicates an association between remodelling and being a prokaryote. From the results of these analyses, the frequency of composite genes in eukaryotes were found to be statistically higher than from that of prokaryotes for most kinds of gene (Figure 2.3b, Appendix A - Figure S2.1, Appendix A - Table S2.1). Some exceptions were found for genes in extracellular structures (category W) and RNA processing (category A) whose 95% CI was found to span 1 (Figure 2.3b). Therefore, across all the species examined, the odds of a gene being a composite if it is a eukaryote is statistically significant higher than if it is a prokaryote.

2.5 Discussion

Network models such as SSNs have been broadly employed in studies of evolutionary relationships (Alvarez-Ponce et al., 2013) and gene sharing and recombination detection. We carried out a large-scale examination of more than one million genes across 1875 complete proteomes including archaea, bacteria, eukaryote, plasmids and viruses. The results suggest that composite genes exist in all organisms and across all kinds of genes.

Eukaryotes, are known to have originated from the merger of an archaeon and a bacterium (McInerney et al., 2014). On average, more than one fifth of eukaryote genes show evidence of remodelling by gene fusion and the probability of a gene in our dataset being composite if it is derived from a eukaryote genome are significantly higher than the probability if the genes comes from a prokaryote genome. What is not known at this stage is the process that has led to the change in frequency of gene remodelling.

Candidates for the process include the combination of homologous recombination during meiosis, combined with the relatively lower level of horizontal gene transfer (HGT) in eukaryotes compared with prokaryotes. The lower level of HGT means that evolutionary innovation via HGT is more restricted in eukaryotes and this restriction, combined with the opportunities for illegitimate crossover events during meiosis could account for the elevated levels of remodelling. In other words, restricting HGT sets up a situation where composite gene formation is one of the main routes to evolutionary innovation. These findings are consistent with Jachiet et al. 2013) who found that eukaryote sequence evolution was highly influenced by gene fusion.

Although evidence of remodelling is quite high in eukaryote genes, plasmid genes also show evidence for a large number of gene fusion events. The average percentage composite genes found in plasmid genomes in our dataset is 14.69%, which is almost as high as the percentage recorded for bacteria. In 2013, Jachiet et al. (2013) mined a data set from three domains of life and MGEs, discovered 42% of composite genes were included at least one MGE gene as a component. Likewise, Halary et al. (2013) found that the plasmids in Borrelia genes behaved like "private genetic goods" (McInerney et al., 2011) and were much less likely to be involved in gene remodelling or sharing with other taxa. It has been suggested that this restriction in gene sharing contributed to the survival of Borrelia against the host immune environment (Barbour et al., 2006; Chaconas and Kobryn, 2010). The high level of remodelling seen in plasmid genes would suggest that MGEs act as a source for remodelling. Corel et al. (2018) also found that gene externalization (gene fusion between cellular organism and MGE) played an important role in microbial evolution (Sibbald et al., 2019).

In our dataset, compared to non-composite genes, fusion genes are more likely to be involved in chromatin structure and dynamics, extracellular RNA processing and modification, as well as cytoskeleton. It has already been shown for eukaryotes that composite genes have been foundational (Méheust et al., 2016), particularly in photosynthetic lineages (such as ubiquitin-nickel superoxide dismutase fusion protein in algae) (Sibbald et al., 2019). Further, a recent published work by McCartney et al. (2019) suggested novel functional protein coding genes in human could emerge through transcription-derived gene fusion. Novel composite genes also have been reported in the origin of haloarchaeal lineages contributed by bacteria, which is named as chimeric (ChiC) genes (Méheust et al., 2018). ChiC genes are more likely to be involved transport and metabolism whereas other composite genes more likely to be involved in replication, recombination and repair, both functions have high composite gene portion in my dataset. In additional, the research from Corel et al. (2018) also suggested that recent externalized genes in abundant in replication, recombination, and repair but hard to accumulate, which could be the result of transposon. Moreover, composite genes in viruses tend to be found in nucleotide metabolism and transport, replication and repair, cell wall, membrane and envelope biogenesis as well as post-translational modification. This finding is consistent with Jachiet et al. (2014).

In conclusion, we applied a network approach in order to investigate composite gene in species across all of life, although the results of this study really only provide a lower-bounds estimate of the extent of gene remodelling, we have been able to show that it is a pervasive and important element of evolutionary history.

Chapter 3.

Dynamics of Spacer Evolution in CRISPR-Cas Systems

3.1 Abstract

CRISPR-Cas is a common immune system that exists in prokaryotes. Many studies have focused on its application to gene editing, but the process through which CRISPR systems arise and are maintained is not completely clear and needs to be studied. In this study, we collected 12,184 prokaryotic species and identified CRISPR arrays using four different computer programs. Using a series of conservative filters, CRISPR arrays were identified in 82.7% of Archaea and 40.6% of Bacteria. Using a combination of mutation tracking in CRISPR repeats, and a comparative genomic analysis using arrays who share multiple similar spacers, different evolutionary processes including polarised integration, middle (or ectopic) spacer integration, deletion, recombination and horizontal gene transfer (HGT) were detected that jointly produce a very complex pattern of CRISPR arrays. Our results suggest that a large proportion of spacers appear in CRISPR arrays in non-chronological order, which probably results from recombination or insertion in the middle of an array. Also, although spacer insertions and deletions occur continuously over time, the last spacer on the end of an array is likely to be more strongly conserved than any other spacer. From network and phylogenetic approaches, we found spacers are rarely shared between distant species compared to universally similar repeats, which potentially indicates little effect resulting from HGT in CRISPR-Cas systems.

3.2 Introduction

In nature, approximately 50% of Bacteria and 90% of Archaea use Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-CRISPR associated protein (Cas) as a guard to defend against invading prophage or other mobile genetic elements (Horvath and Barrangou, 2010). As a defence system, CRISPR-Cas can integrate modified genetic segments from viruses (called protospacers) into CRISPR loci located in the host genome. The integrated sequence is called a spacer (Barrangou et al., 2007) and this process is called adaptation (also known as integration or acquisition). Therefore, CRISPR-Cas can be regarded as a system for generating a library that stores short fragments of foreign DNA from invaders, that are separated by highly conserved direct repeats. Upon subsequent infection, a spacer and part of its associated repeat can be transcribed and matured as CRISPR RNA (crRNA), which integrates and guides adjacent assembled Cas nucleases to destroy complementary invading DNAs (Brouns et al., 2008).

However, protospacers are not randomly selected during adaptation. In type I, II and V CRISPR-Cas systems, the protospacer is only sampled near protospacer adjacent motifs (PAM) in the viral genome (Gleditzsch et al., 2019). PAMs can be thought of as a marker in the invader that the Cas protein recognised during adaptation. PAMS also help Cas proteins distinguish between self and invaders and avoid autoimmunity during interference (Horvath and Barrangou, 2010; Mojica et al., 2009).

Through comparative analysis and phylogenetic comparison, horizontal gene transfer (HGT) has been seen to affect the CRISPR-Cas systems of many species (Godde and Bickerton, 2006; Makarova et al., 2013, 2015). Also, the discovery of complete CRISPR-Cas systems in mobile genetic elements such as plasmids or viruses suggests that they can be seen as “public goods” to some extent (Koonin et al., 2017; McInerney et al., 2011). However, as a defence system, the effects of CRISPR-Cas on transferred genes are still debated. As we know, there are three main mechanisms of HGT: transformation, conjugation and transduction, which show different interactions with CRISPR-Cas. On one hand, it has been reported that conjugation and transformation can be inhibited by CRISPR-Cas systems. Marraffini and Sontheimer (2008) discovered a homologous region between a CRISPR spacer and a widespread plasmid in *Staphylococcus epidermidis*. They also showed that an active CRISPR-Cas system in the host inhibited potential conjugation of plasmids that have spacer-homologous segments. Secondly, natural transformation has been reported to be prevented by CRISPR-Cas in *Streptococcus pneumoniae* (Bikard et al., 2012). They have also shown that species that possess a CRISPR-Cas system would face different selective pressure; which means species who lack CRISR-Cas systems may possess fitness advantages. These results are consistent with the research of Zheng et al. (2020). Their experiments in *Bacillus cereus* suggest that strains that have inactive CRISPR-Cas systems, or that lacked them completely, could better adapt into hosts or environments by promoting HGT of beneficial genes. However, on the other hand, a contradictory result with respect to HGT was observed in the case of transduction. For instance, CRISPR-Cas in *Pectobacterium atrosepticum* facilitates plasmid and chromosomal gene transfer through generalized transduction (Watson et al., 2018). A similar result was shown in *Staphylococci*, where matches between viral genomes and spacers could enhance specialized transduction of CRISPR-adjacent genes (Varble et al., 2019).

As described above, the effects of CRISPR-Cas systems on HGT can be different, depending on the different mechanisms of HGT. Intriguingly, an early study reported that significant HGT cannot be observed in long timescales (Gophna et al., 2015), potentially owing to the rebalance between promotion from CRISPR-mediated transduction and inhibition from conjugation and transformation (Faure et al., 2019; Watson et al., 2018).

In natural environments, prokaryotes and bacteriophages synchronously evolve in an arms race with one another (Hampton et al., 2019). Hence, a comprehensive understanding of the evolution of CRISPR-Cas systems would be beneficial both from the perspective of ecological and medical applications. Diverse processes including insertion, deletion and recombination have been reported in CRISPR loci (Lopez-Sanchez et al., 2012). The most common insertion is naïve adaptation which happens when a cell encounters a new invader (McGinn and Marraffini, 2018). One of the best-known adaptation mechanisms is seen in CRISPR subtype I-E (Figure 3.1a). During spacer adaptation, the Cas1-Cas2 complex needs to recognise the boundary between the leading sequence and the first repeat. This activity is dependent on an integration host factor (IHF) that combines with the leading AT-rich segments and forms a U-shape structure for recognition (Nuñez et al., 2016; Wright and Doudna, 2016). The other example is seen in subtype II-A, where there is a highly conserved leading-anchor sequence (LAS) located upstream of the first repeat, which makes sure all newly integrated spacers are inserted at the leading side (McGinn and Marraffini, 2016). In addition, the unique integration of foreign DNA, apart from naïve acquisition, can be triggered by an existing spacer in a CRISPR locus, and this process is called primed spacer acquisition (Amitai and Sorek, 2016).

When a prokaryote encounters a phage that contains a perfect or partial match to a spacer region, one or more additional protospacers can be acquired through primed acquisition upon interference (Hille et al., 2018). This process has been discovered in both type I and type II CRISPR-Cas systems *in vitro* and *in silico* (Fineran et al., 2014; Nicholson et al., 2019; Nussenzweig et al., 2019). However, both known integration paths are thought to add CRISPR loci to the leading side of the array, that is to say, the order of spacers in CRISPR arrays is thought to reflect the chronological order in which integration has occurred (Yoganand et al., 2017). McGinn & Marraffini (2016) also regarded CRISPR locus as "a molecular fossil record of infections"; the leading spacer records the most recent attack while downstream spacers match more ancient threats (McGinn and Marraffini, 2016). Their research focussed on the mechanism and physiological significance of type II-A polarized integration. They showed that mutations in the LAS would lead to ectopic spacer integration. In this process, a sequence in the middle was recognised as the anchor and consequently the spacer was inserted on the downstream rather than on the leading side. Additionally, they found, under the same phage attack, a corresponding spacer located on the leading side (mimicking polarised spacer integration) confers selective advantages to the host compared to it being located in the middle (mimicking ectopic spacer integration). In particular, it has been noticed that an ectopic integrated spacer can become the first spacer by deleting all segments between the mutated LAS and the new anchor (McGinn and Marraffini, 2016).

Considering spacers that located in leading side encoded more robust immunization against invaders, we speculated within a community, some prokaryotes carried the canonical CRISPR locus while some carried the locus with reordered spacers by recombination and deletion to boost overall immunity against phage. In this scenario, the order of spacers in an array may not be chronological and unable to reflect the timeline of invaders but instead be more dynamic and adaptive as a community. In addition, when we regard a strain or population as a whole, recombination between loci (Kupczok et al., 2015) and HGT across species (Bernheim and Sorek, 2020; Van Houte et al., 2016) can both increase spacer diversity and expand immunity range. Although research from Kupczok et al. (2015), in an analysis of two Gammaproteobacteria (type I) and two Streptococcus species (type II), suggested that the evolution of spacer arrays in CRISPR loci is shaped more by deletion and acquisition than by recombination, we hypothesized that the role of recombination should still be studied.

To date, multiple processes - such as addition, deletion, recombination and HGT - have been reported in the history of the CRISPR-Cas system, and these have the potential to result in an extraordinarily complex evolutionary pattern. In this chapter, in order to understand those evolutionary processes in a more comprehensive way, we have introduced different approaches to categorise and visualise spacer insertion, deletion, recombination and transformation within and between CRISPR loci. We have compiled a dataset containing 12,184 complete Archaeal and Bacterial genomes and used four well-known CRISPR-Cas identification programs to identify CRISPR-Cas loci. In total, we used three different approaches to investigate spacer evolution. First, we traced spacer insertion and deletion in one CRISPR locus through tracking the adjacent duplicated repeats. However, this can only reflect changes within a single locus. In order to investigate the possibility of recombination and HGT between species, we then clustered similar CRISPR arrays and constructed species phylogenies using core genes as taxonomic markers. Core genes represent genes that are highly conserved in all members of the strain (Charlebois and Doolittle, 2004). These genes normally correspond to the most essential housekeeping genes such as those responsible for translation and transcription (Koonin and Makarova, 2013) and are strongly conserved (Segata and Huttenhower, 2011), making them ideal for constructing organismal phylogenetic trees (Ciccarelli et al., 2006; Creevey et al., 2011; Szollosi et al., 2012). Thus, we built species phylogenies by identifying and aligning core genes in each group of array clusters. Combining phylogenetic trees with manually aligned CRISPR loci clusters, we observed spacer gains, losses, and duplications in the CRISPR loci of the various array families. In particular, we found spacers that are located on the end of an array, far from the leading end, show high sequence conservation. Finally, to investigate CRISPR array sharing between species, we used a network approach to study repeats and spacers simultaneously. Significantly similar repeats and spacers were clustered and were used to construct weighted sequence similarity networks. Here, we investigated three questions: 1) Do CRISPR spacers in an array always appear in the chronological order in which they were inserted? If not, why not? 2) How do spacers evolve in a CRISPR array? What processes shape it more: insertion, deletion or recombination? 3) Do closely related species in a community share their CRISPR-Cas systems with one another in order to help obtain immunization prior to attack? We hypothesized that CRISPR-Cas evolution is affected by complex, diverse processes (such as through insertion, deletion and recombination), that help determine the immunity of an individual strain, and immunity of the community (through HGT).

3.3 Method and Materials

3.3.1 Dataset Construction

To construct our dataset, all available Archaeal and Bacterial complete nucleotide genomes were collected from the National Center for Biotechnology Information(NCBI) RefSeq database (Pruitt et al., 2006) in January 2019. In total, there were 277 genomes belonging to Archaea and 11,907 genomes belonging to Bacteria. All data were stored on the University of Manchester high performance computing cluster for further analysis. All species information and download path are available from Appendix B - Table S3.1 and https://github.com/JMcInerneyLab/CRISPRsharing/tree/master/dataset.

3.3.2 CRISPR Array Mining

Four published programs have been widely used to identify CRISPR arrays in prokaryotic genomes: MinCED (Bland et al., 2007), PILER-CR (Edgar, 2007), CRISPRDetect (Biswas et al., 2016) and CRISPRCasfinder (Couvin et al., 2018). I will briefly describe them here, as well as the ways in which they were used in order to identify CRISPR arrays.

3.3.2.1 MinCED

MinCED (version 0.2.1) is short for Mining CRISPRs in Environmental Datasets, a Java-based program that derives from program CRT (Bland et al., 2007). It is designed for detecting CRISPR repeats and spacers from complete genomes or environmental databases. Repeats are nearly identical in CRISPR loci. The sizes of repeats are normally 23 to 47 bp (Horvath and Barrangou, 2010) but extra-large repeats of 50 bp also have been identified (Biswas et al., 2016). Therefore, in order to find all possible CRISPR arrays, the lengths of repeats were set to between 23 to 55 bp during execution of all programs.

3.3.2.2 PILER-CR

PILER-CR (version 1.06) is a C++ based program that can be used to identify CRISPR arrays rapidly and accurately (Edgar, 2007). On top of detecting CRISPR arrays, PILER-CR can also cluster similar direct repeats into groups with the help of software MUSCLE (Edgar, 2004), in which the minimum identity is set as 75%. The cut-off for repeat length was set to the same values as was used in MinCED.

3.3.2.3 CRISPRDetect

CRISPRDetect (version 2.4) (Biswas et al., 2016) is a part of the CRISPRSuite, a CRISPR detection, viewing, analysis, and comparison package. It supports both web and command platforms, and input can be fasta, gff or gbk files. CRISPRDetect uses the CRISPRDirection algorithm (also in the package CRISPRSuite) to identify CRISPR array direction (Biswas et al., 2014). In particular, CRISPRDetect has a scoring system that evaluates arrays based on nine known biological properties such as repeat length, similarity to referenced sequences, *cas* genes and so forth. The detailed calculation can be found in Appendix B - Table S3.2. The developers have suggested that an array with a score above 4.0 can be graded as good. However, CRISPRDetect can only predict *cas* genes from gbk formatted input rather than the fasta format that was used in this study. Therefore, 3.0 was set as a conservative quality score cut-off during the execution of the program. In contrast to other programs, the threshold of spacer lengths is not fixed during execution, cutoffs change with repeat length of the array.

3.3.2.4 CRISPRCasFinder

CRISPRCasFinder (version 4.2.17), an update of the web program CRISPRFinder, is one of the newest programs designed to predict CRISPR arrays (Couvin et al., 2018). It is similar to CRISPRDetect, CRISPRCasFinder seeks to verify the direction of the

CRISPR array (using CRISPRDirection). CRISPRCasFinder also has a scoring system that ranks four evidence levels based on the number and conservative level of repeats and spacers. Detailed level rules are listed in Appendix B - Table S3.2. Any potential CRISPR array that satisfies the two criteria that the evidence level is higher than level 2 and also it has a known direct repeat sequence, is retained. In particular, CRISPRCasFinder can call the dependency program Macromolecular System Finder (MacSyFinder) during execution to predict Cas proteins. Then the identified cluster Cas proteins are used to classify system subtypes (Abby et al., 2014).

3.3.3 Resample Dataset Construction

Different programs show very different results for CRISPR array identification. In order to minimise the possibility that we are analysing false positives and to maximise the likelihood that we are analysing “true” CRISPR arrays, all arrays that were identified by at least three programs were collated into a “resample dataset”.

A strain could contain multiple CRISPR loci and a locus is composed of several to hundreds of repeats and spacers (Couvin et al., 2018). Considering the huge size of these results, this study applied a unique indexing system on spacers and repeats to quickly track and compare different loci. Each spacer was named based on its taxonomic ID, species name, RefSeq accession number, array number, starting position and identified program. An example of the naming system looks like this: “79885_Bacillus_pseudofirmus_NC_013791_3_SP3792003_CAS”, which is the name given to a spacer that was identified using CRISPRCasFinder (CAS), and is located at position 3792003 bp of CRISPR array 3 on NC_013791 chromosome (or plasmid) of species *Bacillus pseudofirmus* (taxid number is 79885). As for the corresponding repeat, since only conservative repeats were collected, the index is similar but lacks a start position. This naming system helped us to easily track back the origin and a location of sequence in the raw data.

3.3.4 Spacer Acquisition

Typically, as described above, during the adaptation stage of CRISPR array formation, new spacers insert between the leader sequence and repeat, and the first repeat is duplicated (Wang et al., 2016a). It is expected, therefore, that spacers and repeats should be ordered chronologically according to infection time. That is to say, if a mutation in the first repeat happens before a new integration, the mutation will be duplicated and will present itself in the subsequent repeats (Figure 3.1a). Hence, mutations within an array could reflect spacer uptake and removal processes to some extent.

In this study, to track spacer evolution in a detailed way, we presented an approach of depicting mutations in repeats. Here, we have chosen to draw the repeats in an alignment, with each repeat being found in the alignment in the order in which they are observed in the array. In other words, the first sequence in the alignment is the first repeat in the array, the second row of the alignment is the sequence of the second repeat in the CRISPR array, etc. All repeats are ordered from 5' to 3', but many CRISPR loci are in the reverse orientation. Therefore, we noted orientation (predicted by program CRISPRDirection that is embedded in CRISPRCasFinder and CRISPRDetect) and adjacent Cas gene cluster position next to the y-axis. In Figure 3.1b, any nucleotide that is identical to the array's conservative repeat is simply coloured grey. If a mutation has arisen, we have coloured the nucleotide according to the colours in the figure legend. In this way, new mutations can be easily observed. By representing the arrays in this way, we can characterise the patterns of mutation. Considering the conserved repeat length in a CRISPR loci, the patterns are very similar to heatmaps.

To filter meaningless mutations, we only collect loci that contain identical mutations that happened at least three times in the same position. In total, 685 CRISPR loci were identified that have trackable mutated repeats and 685 repeat heatmaps were drawn using R's ComplexHeatmap package (Gu et al., 2016).

Initially, we only expected one kind of pattern of repeats in which a mutation arises and persists from an internal position within the array, until the beginning of the array. However, we observed a more diverse and complex picture after classification. Accordingly, we categorised all 685 repeat heatmaps into the following five patterns (Figure 3.1b):

Pattern 1: A mutation arises in the repeat and this mutation is subsequently conserved through at least two duplications of the repeat.

Pattern 2: A mutation arises, it persists through at least two duplication events and then there is a reversal back to the original nucleotide character state.

Pattern 3: A mutation arises; a reversal occurs and then the mutation arises once again.

Pattern 4: The same mutation appears to arise and then reverts to the original character state several times.

Edge: Mutations that only occur on the edge of the array.

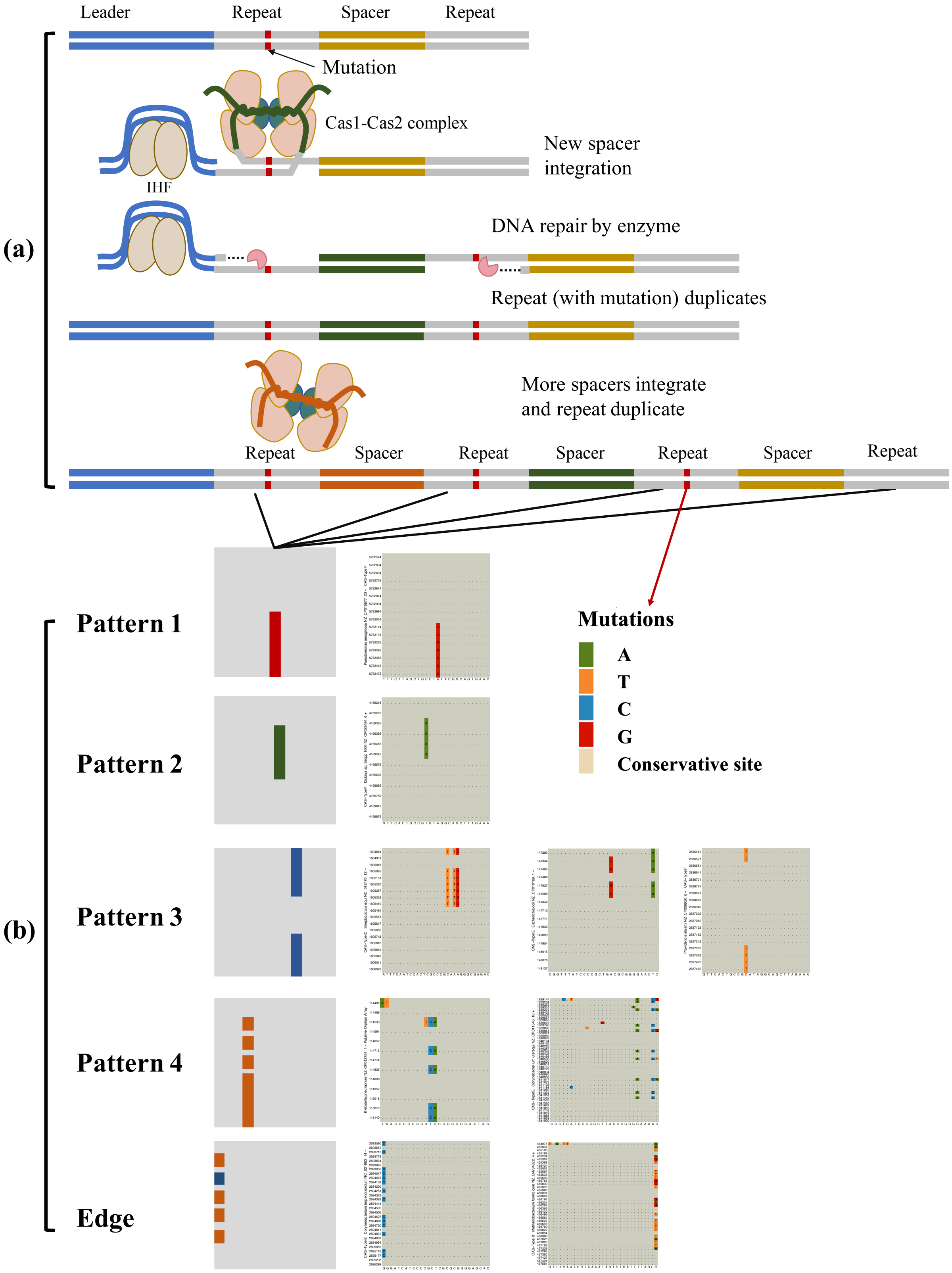

(a)
Leader
Repeat
Spacer
Repeat
Mutation
Cas1-Cas2 complex
IHF
New spacer integration
DNA repair by enzyme
Repeat (with mutation) duplicates
More spacers integrate and repeat duplicate
Repeat
Spacer
Repeat
Spacer
Repeat
Spacer
Repeat
(b)
Pattern 1
Pattern 2
Pattern 3
Pattern 4
Edge
Mutations
A
T
C
G
Conservative site

Figure 3.1 The mechanism of integration and clues of constructing a model of repeat-mutation patterns. (a) General schematic of spacer integration. With the help of integration host factor (IHF, in eggshell colour), Cas1-Cas2 complex that binds to the protospacer proceeds two nucleophilic attacks between the leading sequence and the first repeat. A new spacer was inserted between the gap and two strands of repeat serve as templates for duplication. **(b)** In one array, repeats that have trackable mutation are ordered based on their position and coloured in a heatmap format. Grey nucleotides are regarded as conserved sites while other coloured nucleotides are regarded as different mutations. The X-axis represents conserved repeat sequences. The Y-axis represents positions of repeats that rank from 5' to 3'. Other information including species, locus, array number, array orientation and associated Cas gene are also labelled next to y-axis. The order of Cas label indicates its location near the CRISPR locus: upside means upstream (near 5') of CRISPR locus and downside means downstream (near 3') of CRISPR locus. Full size pictures of example repeat patterns are shown in Appendix B - Figure S3.3.

3.3.5 Spacers Dynamics

There are a limited number of CRISPR loci that possess recognisable mutations, and this cannot reflect the full extent of the dynamics between species, particularly multi-spacer transformation through HGT. Therefore, to investigate how spacers evolve between loci within chromosome or between cells, we categorised similar arrays and placed them into families based on their spacer similarity. Along with phylogenetic analysis, we can trace the route of spacer gains and losses along the evolutionary history of each CRISPR locus.

3.3.5.1 Similar Spacer/Repeat Clustering

In order to group similar CRISPR arrays, we first searched and grouped homologous spacers and repeats. Even though repeats in CRISPR arrays are highly conserved, from last section we can see that mutations still have been observed. Here, we used a global alignment approach, taking consideration of the short size of spacers (mostly 27 to 45 bp) and repeats (mostly 23 to 47 bp). A UCLUST (version 2.1) search was performed to identify spacers with significant similarity to one another and group them into clusters. The identify cut off was set as 90% for spacer comparisons and 92% for repeat comparisons.

3.3.5.2 CRISPR Array Family Allocation

To cluster CRISPR loci based on similarity, we used the Markov Cluster (MCL) algorithm which evaluates and generates clusters based on stochastic flows in graphs (Van Dongen, 2000). In our case, the strength of the links between nodes is based on the similarity between two CRISPR arrays. Every shared spacer between two loci increases the strength of association between two nodes. For example, if two arrays have 20 significantly similar spacers, it would be identified as very similar and the e-value would be set to 1e-20. Similarly, if two arrays only share 1 spacer, the link between them would be very weak and the e-value would be set to 0.1. Only pairs with more than two similar spacers were selected and CRISPR array families were clustered using the MCL algorithm (inflation value: 2.0).

3.3.5.3 CRISPR Array Visualisation and Alignment

To visualise each group of clustered CRISPR arrays, we used CRISPRStudio (Dion et al., 2018) which is a Python-based program that can colour clustered homologous CRISPR spacers in a duo colour square-diamond mode. However, the CRISPRStudio software is limited to the use of complete .gff results that are produced as output from the CRISPRDetect program. Therefore, all other CRISPR identification results were formatted into GFF3 format to fit CRISPRStudio using bespoke scripts. The scripts to perform these conversions were written in Python (Version3.5.3) and are available at https://github.com/JMcInerneyLab/CRISPRsharing/blob/master/crispr_gff.py.

In order to reconstruct the evolution of individual CRISPR loci, gene families were manually aligned to track spacer changes using the vector graphic tool Inkscape (Bah, 2007).

3.3.5.4 Phylogenetic Tree Construction

CRISPR arrays that clustered into a single family were regarded as having been derived from the same origin. To track the process of spacer dynamics in an array family, an independent species tree is required to show an evolutionary timeline. Here, we chose the core genes (those genes found in every member of the species) as phylogenetic markers to construct phylogenies for each family.

Following the MCL analysis of the array similarity graph, we only selected array families that containing at least 10 members. Of note is the fact that we observed that the species in each array group all belonged to the same genera, thus the outgroup species for each family was selected from different genera, but the same taxonomic family. Additionally, all genomes were annotated, using Prokka (version 1.13.7) (Seemann, 2014), and these annotated genomes were used as input files for Roary (version 3.13.0) (Page et al., 2015) in order to search for core genes across each array family with an identity threshold of 80%. After this, core genes were aligned using MAFFT (version 7.453) (Katoh et al., 2005) and a gene tree was inferred using IQ-TREE (version 1.6.1) (Nguyen et al., 2015) employing the Maximum Likelihood GTR+I+G model. Confidence in phylogenetic hypotheses was evaluated using bootstrap resampling (1,000 replicates).

3.3.5.5 Gain and Loss Analysis

To analyse the gain and loss of spacers within each gene family, it is necessary to reconstruct the ancestral character states at each internal node in a phylogenetic tree (Creevey et al., 2011; Segata and Huttenhower, 2011). In this way, if we know the ancestral state, we can track whether there were subsequent gains or losses of spacers. We used the program Tree Analysis Using New Technology (TNT) (version 1.5) (Giribet, 2005) in conjunction with a spacer presence-absence (0/1) matrix and the phylogenetic tree that was constructed tree as outlined in section 2.5.4. The dynamics of spacer gains and losses were plotted onto the corresponding tree nodes and presented using the iTOL online tool (Letunic and Bork, 2007).

3.3.6 CRISPR Loci Networks

To investigate the effects of recombination and HGT on CRISPR-Cas families, we applied a network approach to out CRISPR loci, in addition to phylogenetic analysis. We constructed a weighted spacer sharing network where nodes represent CRISPR loci, and edges represent the property that those loci share similar spacers. To underline the effects of HGT and limit the number of spacers deriving from the same foreign genetic elements, only loci that were found in different family or genera but share similar spacers were included. The weight of an edge is represented by the number of spacers that are shared between two CRISPR loci. The Python package ETE3 in combination with the NCBI Taxonomy (Huerta-Cepas et al., 2016) were used to quickly retrieve the full taxonomic description of any given species. Nodes are coloured according to the taxonomic order from which the species originates. The Gephi program (Bastian et al., 2009) was used to render and display networks.

The exact repeat sharing pattern between species is likely different to the spacer sharing pattern. To investigate this possible difference, a similarity-based tree-network pattern for shared spacers and repeats was constructed. Although a spacer sharing network has been built, the evolutionary relationships between species could not be fully presented just through colouring nodes by the taxonomic origins. Here, we used the public All-Species Living Tree Project (LTP) 16S ribosomal RNA (rRNA) prokaryotic tree as a resource (Yarza and Munoz, 2014). However, this also limits the capacity of samples because there are only 843 species in this project that possess identified CRISPR arrays. Therefore, connections were drawn between these 843 species if at least one spacer or repeat was detected in both species. Trees were displayed through the iTOL webtool (Letunic and Bork, 2007).

3.4 Results

3.4.1 Resample dataset

In total, four programs (MinCED, PILER-CR, CRISPRDetect, CRISPRCasFinder) were used to identify CRISPR-Cas systems present in 277 Archaeal and 12,184 Bacterial complete nucleotide genomes from NCBI RefSeq database. The results obtained from all four programs indicate a similar proportion of putative CRISPR arrays across all Archaea. However, the four programs produced quite different results when used to analyse bacterial genomes (Table 3.1). PILER-CR and CRISPRCasFinder identified more CRISPR loci in Bacteria compared with CRISPRDetect and MinCED. To conservatively compare and analyse CRISPR loci across species, a resampled dataset was constructed which only contained CRISPR loci that were identified by at least 3 programs. As shown in Table 3.2, 889 CRISPR loci were identified from 229 archaeal strains (82.7%) and 10,703 CRISPR loci were identified from 4,947 bacterial strains (40.6%) in the resample dataset. The most common length of a spacer is 32 bp, followed by 36 bp, whereas the most common length of a direct repeat is 29 bp, followed by 36 bp (Appendix B - Figure S3.1).

Considering that *cas* genes are only predicted by the CRISPRCasFinder software, the *cas* information in the resample dataset is only retrieved using this software program. According to the output of CRSPRCasFinder, across the 11,592 putative CRISPR arrays, 1,426 arrays (12.3%) were found without adjacent *cas* genes (these are known as “Orphan CRISPR” arrays (Makarova et al., 2015)).

In addition, when analysing the position of *cas* genes in all 4,283 CRISPR arrays that have recognisable *cas* genes in the neighbourhood (<= 10,000 bp window), 3,012 CRISPR loci were located downstream of *cas* genes, 1,064 upstream, and interestingly, 207 loci were found to be situated between protein coding *cas* genes (Appendix B - Figure S3.2) with *cas* gene orientation changed.

Table 3.1 Numbers of species who have identified CRISPR arrays in Archaea and Bacteria by four programs.

	Archaea	Bacteria
MinCED	239 (86.3%)	6,079 (49.9%)
PILER-CR	259 (93.5%)	9,804 (80.5%)
CRISPRDetect	237 (85.6%)	5,399 (44.3%)
CRISPRCasFinder	252 (91.0%)	9,718 (79.8%)

Table 3.2 Lists of different CRISPR related results identified in the resample dataset.

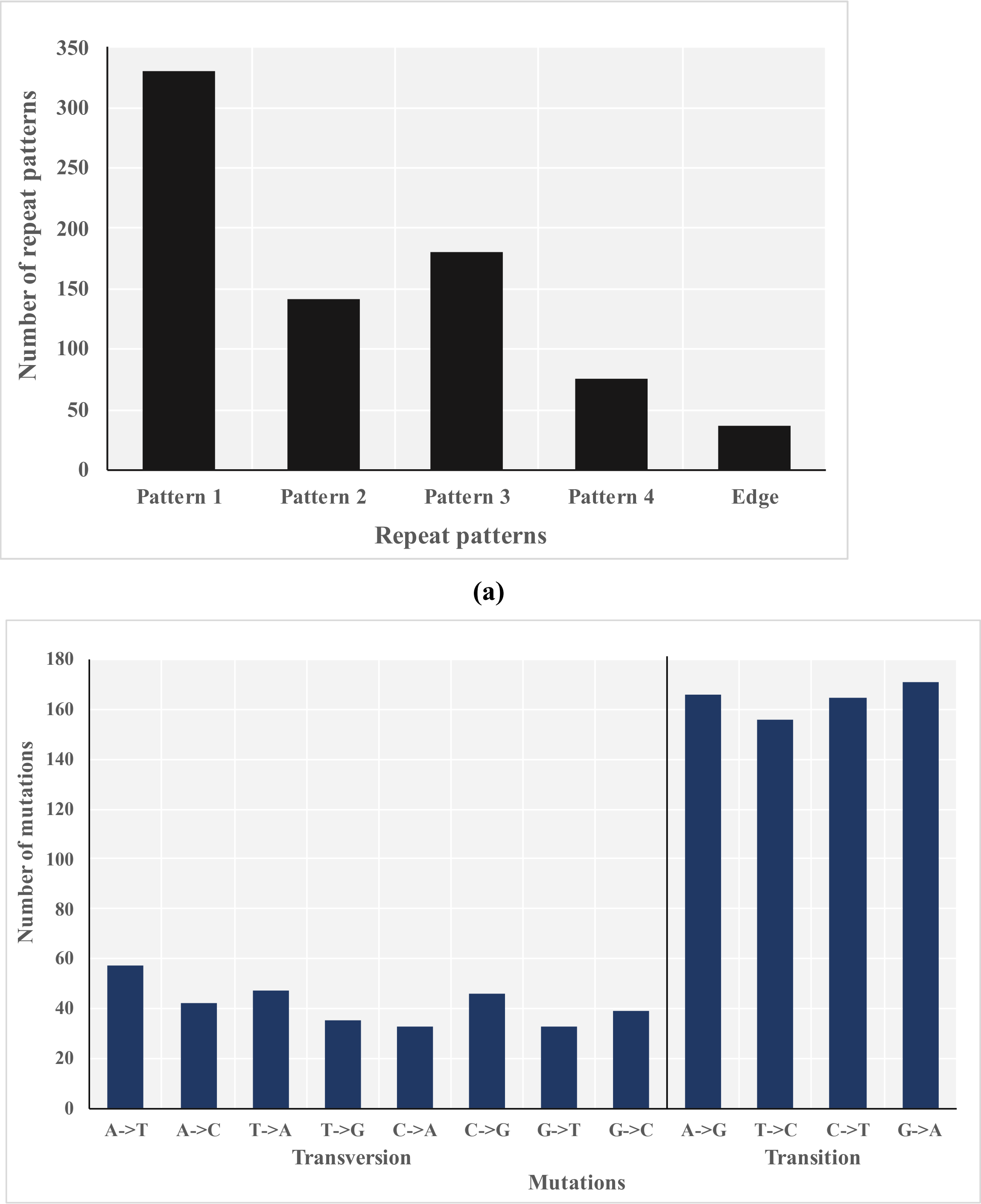
	Archaea	Bacteria
Species that have CRISPR arrays	229 (**82.7%)**	4,947 (**40.6%)**
Species that have identified *cas* genes	193 (69.7%)	4,545 (37.3%)
CRISPR loci in plasmids	13 (9.6%)	160 (1.5%)
Cas genes in plasmids	8 (5.9%)	114 (1.1%)
Total number of CRISPR arrays	889	10,703
Range of spacer number in a locus	2 to 245	2 to 587

3.4.2 Direct repeat tracking

In addition to constructing a conservative CRISPR-Cas dataset from the complete set of available genomes, we also wanted to investigate the process of spacer evolution within and between CRISPR loci. First, we depicted the evolution of spacers by tracing changes in the proximal repeat. As noted previously (see Introduction 1.3.2.1, Figure 3.1a), when a spacer adapts into a host genome, both the innate and primed procedures trigger the duplication of the first direct repeat (McGinn and Marraffini, 2018). That is to say, if mutations occur before duplication, this mutation would be reproduced in the following repeats unless a reverse mutation occurred. From all putative CRISPR arrays in the resample dataset, we filtered 685 arrays that have trackable mutations in their repeat region and plotted these mutations as a form of heatmap image, owing to the character of the most highly-conserved length of repeats. We used the heatmap to show the locations of mutations in all repeats of each CRISPR array (Figure 3.1b). The X-axis represents the consensus repeat sequence, whereas the y-axis represents the position of each repeat (ordered 5’ - 3’). One array could contain multiple patterns and all arrays were categorized into different patterns based on their characters (see Methods 3.3.4).

Pattern 1 conforms to the current understanding of the spacer adaptation process. This pattern indicates that a mutation arose once as the CRISPR array was growing and this mutation was subsequently retained throughout further duplications. This pattern is observed in 331 (43.55%) CRISPR loci, which is also the most common among all the patterns (Figure 3.2a). Interestingly, however, we also observed many instances of patterns 2 (141 loci, 18.55%) and 3 (179 loci, 23.55%), and together these amount to approximately half as many as the number of observations of pattern 1. Patterns 2 and 3 could either arise as a consequence of nucleotide mutations and the reversal of the mutations or alternatively, they could arise as a consequence of insertions of spacers into the middle of the array. Pattern 4 is observed the least often (74 loci, 9.74%) among four normal patterns, and this pattern probably results from nucleotide flip flop in the direct repeats or gene recombination within CRISPR arrays. All these unusual patterns (2,3,4) suggest the effects of more complex processes like recombination, middle spacer integration besides polarized insertion and deletion. Finally, the Edge pattern, which is present in the smallest number (35 loci, 4.61%), could arise by either of two mechanisms. One, previously reported in subtype I-E system (Swarts et al., 2012), suggests that PAM can donate its last nucleotide during spacer integration. To test this hypothesis, we identified the subtypes of all 685 samples (Table 3.3) based on results from CRISPRCasFinder Cas cluster. We found that type I-E was present in 48.6% of all edge samples. The remaining samples which show edge pattern may indicate the role that PAM played in this subtype. The other reason is possibly due to software programs incorrectly identifying break points between spacers and repeats. During identification, the first (or last) nucleotide of spacer might be accidently included in the repeat sequence, and reflected flaws of CRISPR loci identification by bioinformatic tools.

In addition to allocating all mutations into different patterns, we also analysed the patterns of nucleotide substitution (Figure 3.2b). In total, 990 nucleotide mutations were observed containing 658 transitions and 332 transversions. The ratio of transitions to transversions is 1.98.



(a)

(b)

Figure 3.2 Patterns of repeat mutation in CRISPR arrays. (a) Distribution of different patterns across all 685 CRISPR repeat samples. **(b)** Number of different mutation types.

Table 3.3 Numbers of CRISPR-Cas subtypes in 685 repeat mutation examples (One array may include multiple patterns).

CRISPR-Cas subtypes	Repeat pattern					Total
	1	2	3	4	Edge	
Type I-A	10	3	6	9	0	28
Type I-B	52	15	17	9	5	98
Type I-C	35	9	6	1	4	55
Type I-D	6	2	2	2	0	12
Type I-E	99	34	47	18	18	216
Type I-F	37	34	39	7	2	119
Type I-U	3	1	1	3	0	8
Type II-A	6	3	3	1	2	15
Type II-U	14	3	8	2	0	27
Type III-A	10	7	9	3	1	30
Type III-B	9	7	11	3	2	32
Type U	2	0	3	0	0	5
Orphan CRISPR	48	23	28	18	3	120
Total	331	141	180	76	37	765

3.4.3 Spacer dynamics

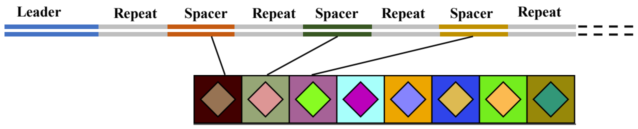
From the previous repeat pattern analysis, we observed that spacer integration was not always in a polarized manner. New spacer integration does not seem to proceed from edge side, which could result from recombination within or between similar loci, HGT between species or undiscovered mechanism of ectopic spacer integration. However, this repeat-mutation trace analysis can only reflect dynamics within one array and the number of arrays who have recognisable mutations are limited. Therefore, to investigate further the evolutionary history of spacers in arrays, we clustered homologous CRISPR arrays based on the similarity of spacers (see Methods 2.5.2) and analysed spacer dynamics along the phylogenetic tree. A total of 87 array families were found to contain 10 or more array members. The program CRISPRStudio was used to visualise the dynamics of spacer changes between closely related loci across genomes. Similar spacers in each array family are coloured by the same colour (two-colour square-diamond pattern in Figure 3.3a) and all loci were manually aligned (Figure 3.3b, 3.3c).

In our hands, almost all array families could be aligned well, except for two families that have extra-long arrays and aligning these long arrays exceeded our manual processing ability. That is to say, in one family, spacers shared between loci were basically in the same ordered. Since we ignored all spacer order information while clustering, recombination within CRISPR arrays makes the process of alignment impossible. Thus, the fact that we could align CRISPR spacers in 85 families suggested that rate of recombination within a locus during spacer evolution is rare.

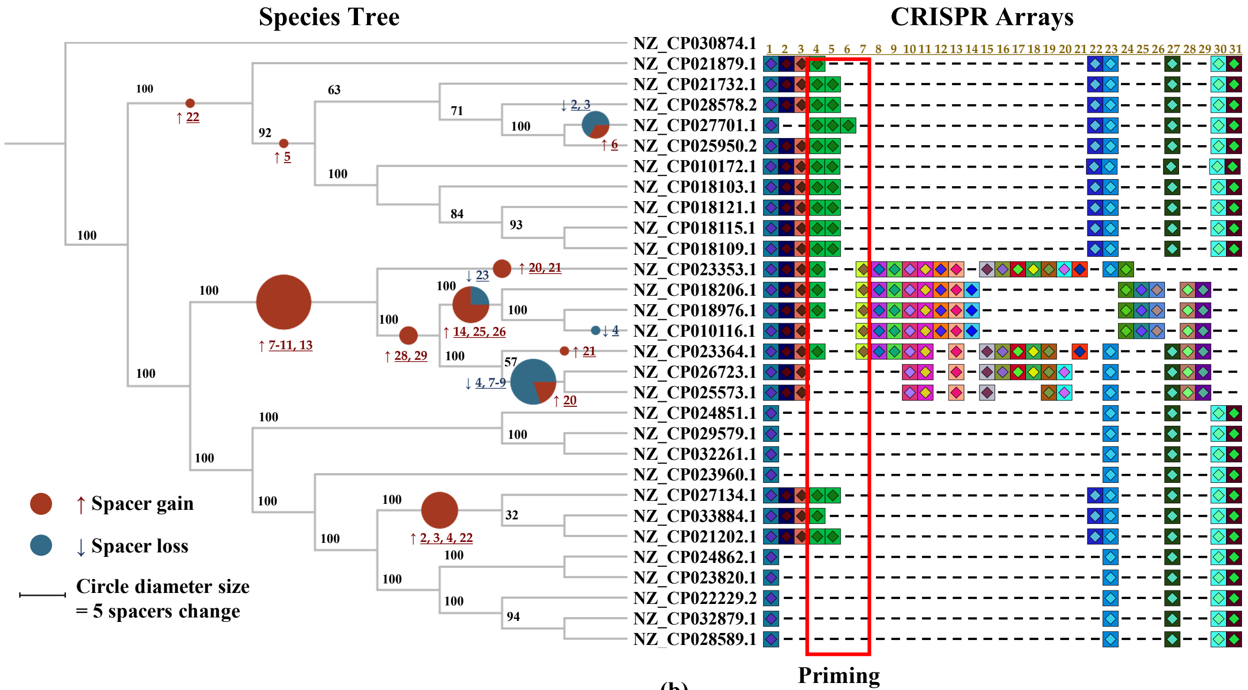
In order to test the effects of ectopic spacer integration and HGT in CRISPR arrays, we constructed a core gene phylogenetic tree for each family. Along with aligned spacers, the evolutionary history can be tracked over time. Core genes were identified using the program Roary (Page et al., 2015) and aligned using MAFFT (Katoh et al., 2005). Using core genes alignment as input, IQ-TREE (Nguyen et al., 2015) was used to infer the species maximum likelihood tree. Together with the spacer presence and absence matrix, the gain and loss of spacers were calculated using what program TNT (Giribet, 2005) (see Methods 3.3.5).

Here, we presented two examples (Figure 3.3b, 3.3c) from all 87 sample families. We selected these two examples because the array lengths are appropriate to show in one figure and also because they exemplify particular evolutionary processes, such as priming and duplication. Spacers in the manually aligned CRISPR loci families were marked with integers. Together with the phylogenetic tree, first, Family 23 (Figure 3.3b) is presented that contains 29 strains of *Escherichia coli* with *Raoultella sp. X13* used as an outgroup. Gains and losses of spacers occurred independently multiple times but most in recent history. Evidence for the gain of new spacers support the possibility that there are other methods of integration apart from polarised integration. For example, spacers 14, 25 and 26 in strain NZ_CP018206.1 (and NZ_CP018976.1, NZ_CP010116.1) are located in the middle but their restricted phylogenetic distribution suggests that they were inserted on a later branch than spacers 28, 29 on the leading side. This is the most parsimonious explanation for the observed pattern, though, of course, it does not rule out more complex, unobserved events of gain and loss. Also, a series of priming event (spacer 4-6 in the red square of Figure 3.3b) can be tracked though CRISPR alignment. Similarly, evidence of priming was detected in 33 out of 87 samples. Another example is Family 36 (Figure 3.3c) that includes 18 stains of *Salmonella enterica,* with *Lelliottia amnigena* being used as an outgroup. There are multiple gains along the history of these strains and minor losses only in the recent history of the strains. Similarly, in Family 36, spacers 4, 11 in NZ_CP022034.1 were integrated in recent history but located downstream of spacers 22, 23 that were integrated earlier in the history of the genomes, providing additional evidence for non-chronological spacer integration. In particular, spacers 7, 8, 10 are shared between two distant genomes NZ_CP022019.1 and NZ_CP024165.1 (or NZ_CP030288.1) with independent inheritance, which indicate the effects of HGT in spacer evolution. Moreover, as shown in the bottom of Figure 3.3c, a five-spacer gain pattern (spacers 46-50) detected in NZ_CP032816.1 indicates an independent array duplication. Duplications were detected in 23 out of 87 samples and may result from recombination (Kupczok et al., 2015). Overall, different evolutionary processes including polarized integration (including priming), ectopic integration, duplication, loss, and HGT are reported, which cooperatively lead to an intricate spacer history cooperatively.

To visualise the spacer dynamics thoroughly, we analysed the evolutionary history by combining repeat patterns with aligned CRISPR loci. In our 87 examples, only Family 59 (Figure 3.4a) has recognisable mutations in direct repeats. Among loci in Family 59, repeats adjacent to spacers 2-6 can be traced with repeat-mutation heatmaps. All loci in this family are in the reverse orientation, which means spacer 1 which is located at the 5'-end is the earliest spacer. Also, since CRISPR array identification follows a "repeat-spacer-repeat" pattern (Figure 3.4f), a spacer is located between two repeats. That is to say, for a spacer in an array, two adjacent repeats can be traced in the corresponding repeat heatmap that are vertically ordered. The upper repeat is the older repeat and serves as the template for the subjacent repeat duplication. In Family 59, the earliest branching of this CRISPR locus is found in NZ_CP027440.1 (Figure 3.4b), where we see spacers 5 and 6 integrated chronologically as evidenced through the fact that we can trace a single nucleotide mutation (from T to A) that is found in all four repeats that flank the spacers. As we move along the phylogenetic tree, we see the insertion of spacer 2 on the branch subtending Node 1, the node that subtends all groups on the tree except for NZ_CP027440.1 and the outgroup. As shown Figure 3.4c, in NZ_CP025859.1, site A mutated to G and conserved to the following integration of spacer 5. However, abnormal repeat duplications (pattern 3) were identified in Node 2 while spacers 3 and 4 integrated (as shown in NZ_CP010169.1, Figure 3.4d). The repeat that is next to spacer 4 still followed previous mutations while spacer 3, located upstream next to the locus, conserved the repeat. Considering that the evolutionary history of spacer 5 was missed, we proposed three possible conjectures that may explain this pattern. First, spacer 3 initially integrated downstream of spacer 5 but was shuffled to the middle through recombination. Second, a replacing recombination may happen between loci who share the same repeat; this could result in deletion of the repeat with mutation and insertion of the original conserved repeat. The third is probably due to ectopic spacer integration with an unknown repeat duplication mechanism. In the most recent strain NZ_CP020058.1, there were 5 spacer losses (spacers 4 to 8), which only left repeats with two spacers in Figure 3.4e. Together with the previous two examples (Family 23 and 36), we found that spacer deletions only occur in the middle of an array. Intriguingly, we found spacers 1, 5, 6 are shared across all loci in Family 59 in the same order but integrated with different repeats, which may result from HGT or indicate a robust fitness benefit of these three spacers. Together, these results indicate that CRISPR array evolution is much more complex than previous appreciated.


Leader
Repeat
Spacer
Repeat
Spacer
Repeat
Spacer
Repeat

(a)


Species Tree
CRISPR Arrays
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
NZ_CP030874.1
NZ_CP021879.1
NZ_CP021732.1
NZ_CP028578.2
NZ_CP027701.1
NZ_CP025950.2
NZ_CP010172.1
NZ_CP018103.1
NZ_CP018121.1
NZ_CP018115.1
NZ_CP018109.1
NZ_CP023353.1
NZ_CP018206.1
NZ_CP018976.1
NZ_CP010116.1
NZ_CP023364.1
NZ_CP026723.1
NZ_CP025573.1
NZ_CP024851.1
NZ_CP029579.1
NZ_CP032261.1
NZ_CP023960.1
NZ_CP027134.1
NZ_CP033884.1
NZ_CP021202.1
NZ_CP024862.1
NZ_CP023820.1
NZ_CP022229.2
NZ_CP032879.1
NZ_CP028589.1
100
100
↑ 22
63
92
↑ 5
71
100
↓ 2, 3
↑ 6
100
84
93
100
100
↑ 7-11, 13
100
↓ 23
↑ 20, 21
100
↑ 14, 25, 26
100
↓ 4
↑ 28, 29
100
57
↑ 21
↓ 4, 7-9
↑ 20
100
100
100
100
100
↑ 2, 3, 4, 22
32
100
100
100
94
↑ Spacer gain
↓ Spacer loss
Circle diameter size
= 5 spacers change
Priming

(b)

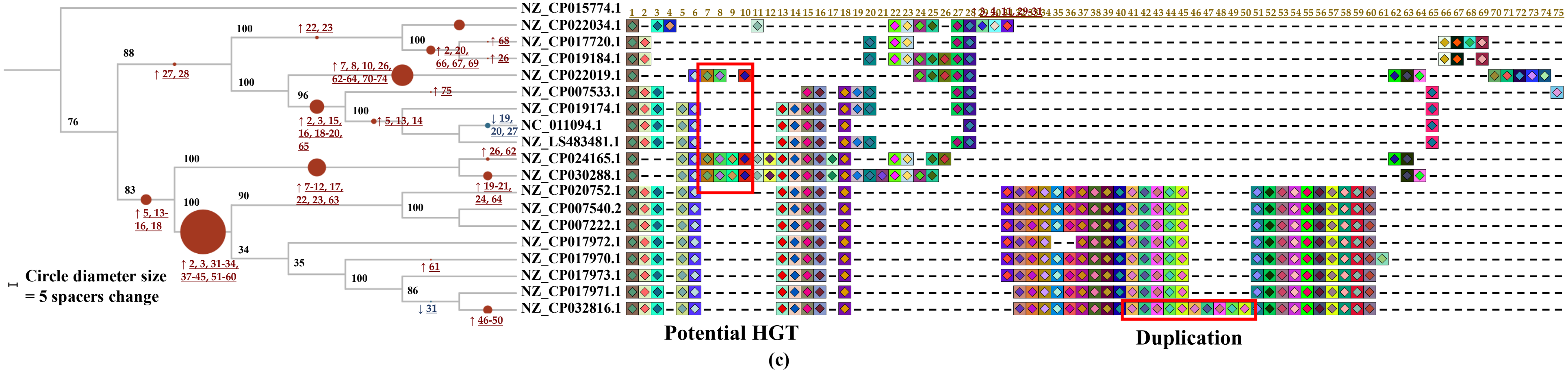


Figure 3.3 Aligned CRISPR arrays with the associated phylogeny for Family 23 (b) and 36 (c). (a) shows the idea of how spacers in one array was drawn for each Family. Phylogeny is constructed from core genes in each family. Gains and losses of spacers in strains and nodes are represented by the pie chart. Red represents gain and blue represents loss. The size of pie chart is related with the total number of changes. Integers next to the pie charts represent changing spacers and the evolution process be tracked in the alignment. Phylogenetic trees are depicted by iTOL and CRISPR loci are manually aligned through software Inkscape.

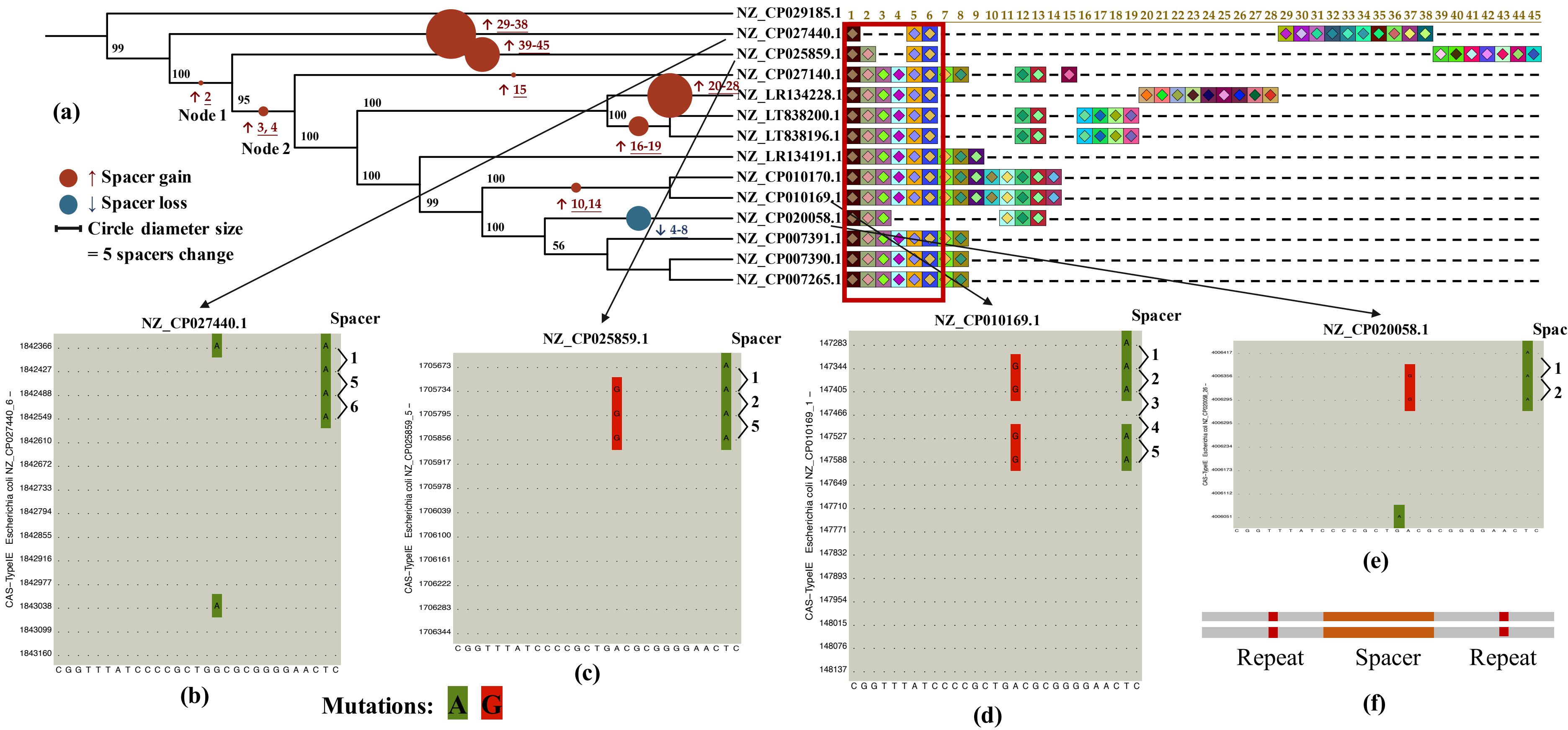
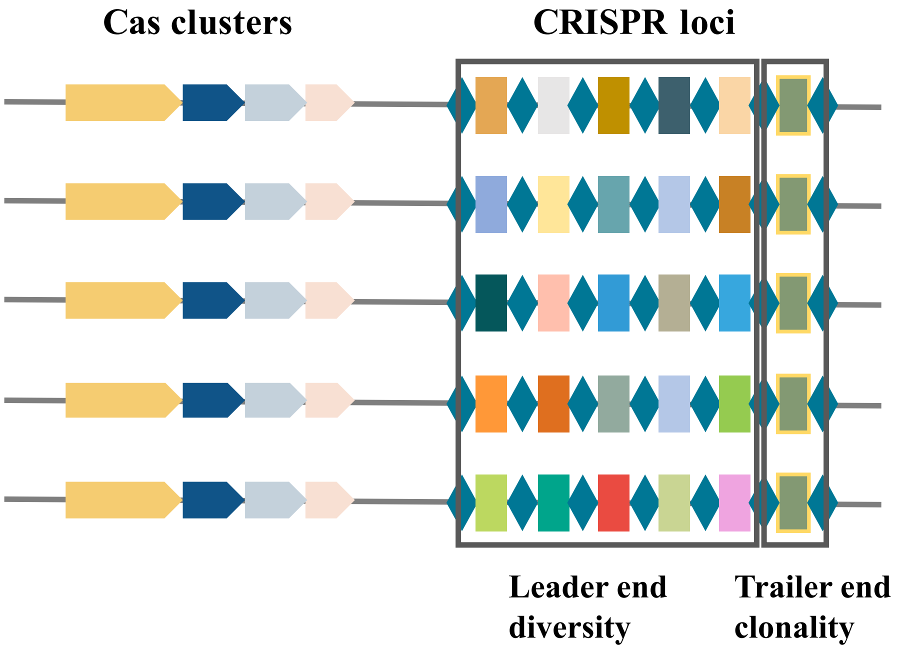


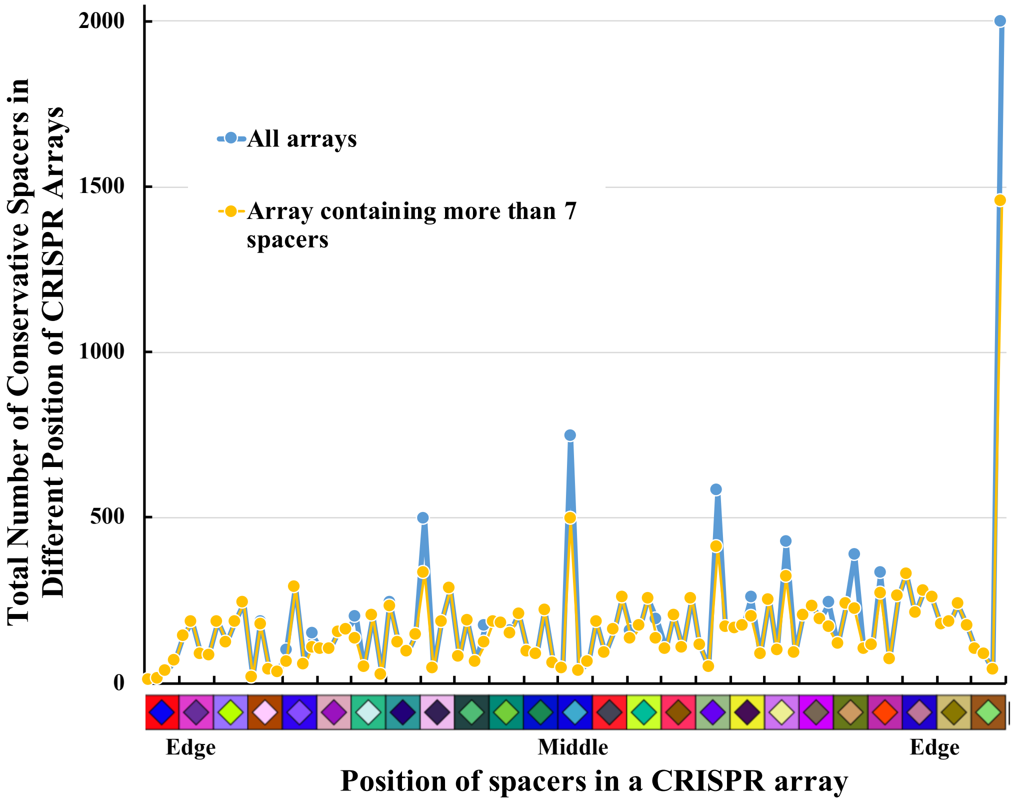
Figure 3.4 CRISPR array alignment of Family 59 (a) with recognisable mutation patterns (b-e). One array can be visualised through two different ways: repeat heatmap **(b-e)** and spacer alignment **(a)**. Considering that CRISPR locus identification requires repeats on both sides, the number of repeats is greater than the number of spacers by 1. The two repeats that adjacent by spacer order vertically in repeat heatmap and the bottom one is the newly duplicated repeat.

3.4.4 Anchor spacer

The evolution of spacers within CRISPR loci is diverse in order to defend against challenges from the environment or host. However, as seen with conservative spacers 1, 5 and 6 in Family 59 (Figure 3.4), spacers seem to face different fitness costs during losses. After observation of all 87 CRISPR spacer alignment samples, we noticed a pattern that for loci in one family, along with rapid spacer turnover, it is also common to have a shared spacer situated near the edge, like an anchor (Figure 3.5a). To test this hypothesis, we calculated the number of conservative spacers (defined as spacers that presenting in more than 85% of strains) and its position in an array. During calculation, we used percentage instead of index numbers to mark the spacer's position, which may increase false positive when evaluating the relationship between position and chance of conservative site in short arrays. Therefore, we calculated arrays from the full dataset and subset the results of arrays that containing more than 7 spacers. Surprisingly, the conservative spacer located at an edge is dramatically higher than in other positions (Figure 3.5b), followed by the site that located in the middle. Also, the conservative edge is normally an ending edge, which is very likely to be the oldest spacer. This also indicates the uneven selective pressure of spacer loss based on position.



(a)



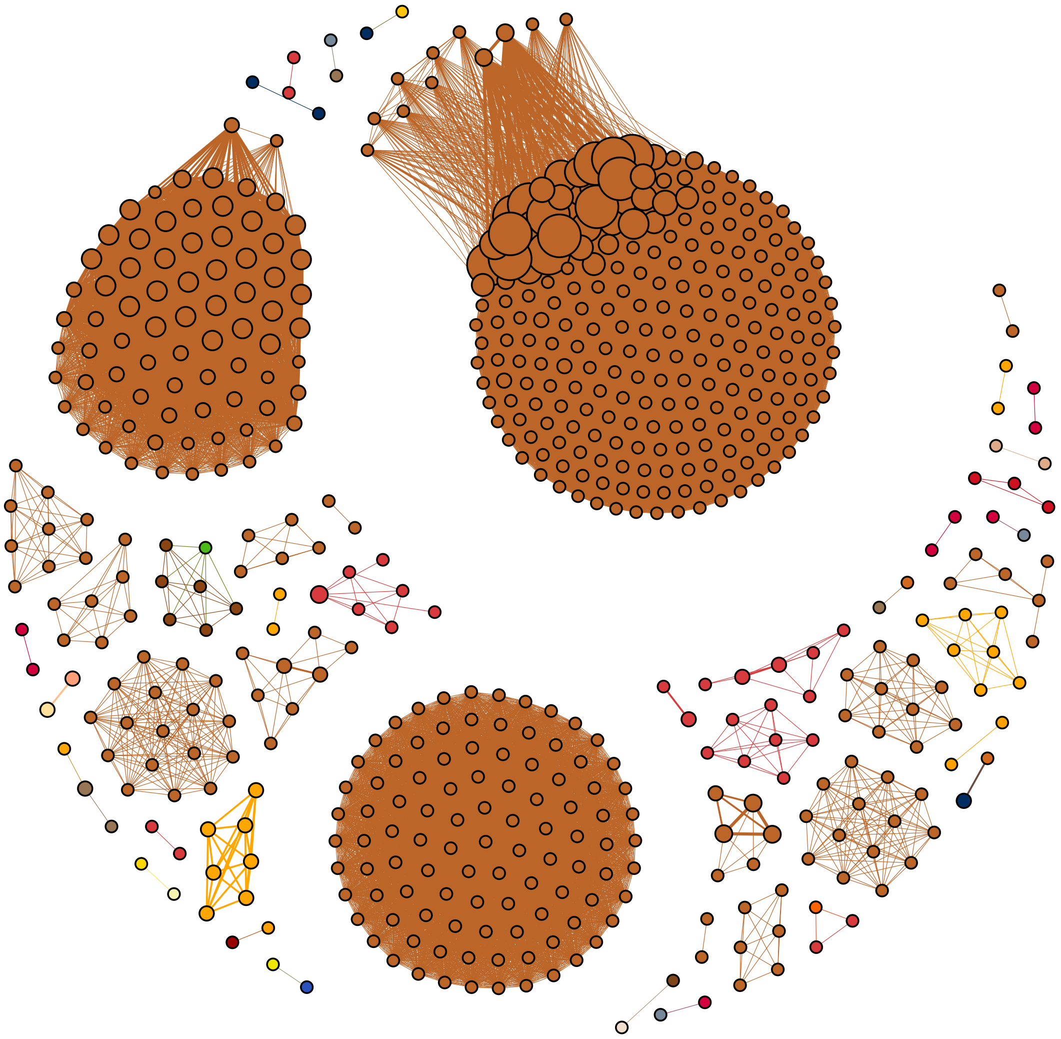
(b)

Figure 3.5 The scenario and number of conservative (anchor) spacers in CRISPR arrays. (a) Spacers located on one edge (normally the earliest integrated spacer) is highly conserved across loci compared to spacers in other positions. **(b)** The number of conservative spacers in different positions of CRISPR loci.

3.4.5 Shared spacers and direct repeats

In addition to integration, deletion, and recombination in CRISPR loci, HGT may potentially play an important role in CRISPR-Cas evolution. Evidence of HGT of CRISPR loci between strains has been presented in array Family 36 (Figure 3.3c) and complete CRISPR-Cas systems in plasmids (Table 3.2) also indicates CRISPR-Cas loci are shared through conjugation. Although HGT in CRISPR-Cas has also been reported by other studies (Godde and Bickerton, 2006; Makarova et al., 2013, 2015), the timescale of HGT and how distantly CRISPR-Cas loci can be shared across species is still controversial issue. Here, we have displayed the sharing patterns of repeats and spacers between CRISPR loci across all prokaryotes using both network and treelike approaches.

Firstly, we used the results from the clustered spacer families (see Methods 2.5.1) identified in the resample dataset, in which a global alignment and clustering was performed by UCLUST (identity cut off 90%). In total, 169,329 spacer families were identified, and 24,236 families were composed of more than 1 member. Using the clustered families as input, we depicted a weighted spacer similarity network that visualised similar spacers. Edges in this network can exist for three possible reasons. First, considering successful phage infection is specific particular receptors on the surfaces of particular bacteria (Seed, 2015), same spacers could derive from the same phage attacks. Second, spacers with extremely high fitness benefit in ancient CRISPR loci would be inherited by their offspring. Last but not least, spacers or even complete CRISPR loci might spread via HGT. The last relationship is what we wished to investigate in the spacer network. Thus, to reduce edges caused by the first two possibilities, we only select edges shared by spacers from different taxonomic genera and families. In this network, nodes represent CRISPR loci, and are coloured based on taxonomic order. The size of a node represents the size (number of spacers) of the locus while the weight of an edge represents the number of shared spacers between two loci. There are 36,416 edges in the genera spacer network (Figure 3.6a) and the three largest connected components are all among loci in Enterobacterales. By contrast, the network of spacers shared between loci from different taxonomic families is much smaller, containing only 50 edges (Figure 3.6b). The dramatic reduction in connections between loci suggests infrequent HGT of CRISPR-Cas between remote species. Also, we found most loci that are found to have similarity across different families only share a single spacer but there are 4 pairs of loci that have multiple spacers in common, such as the khaki coloured edge (Figure 3.6b) between *Rhodocyclaceae bacterium* and *Zoogloeaceae bacterium* Par-f-2.



(a)

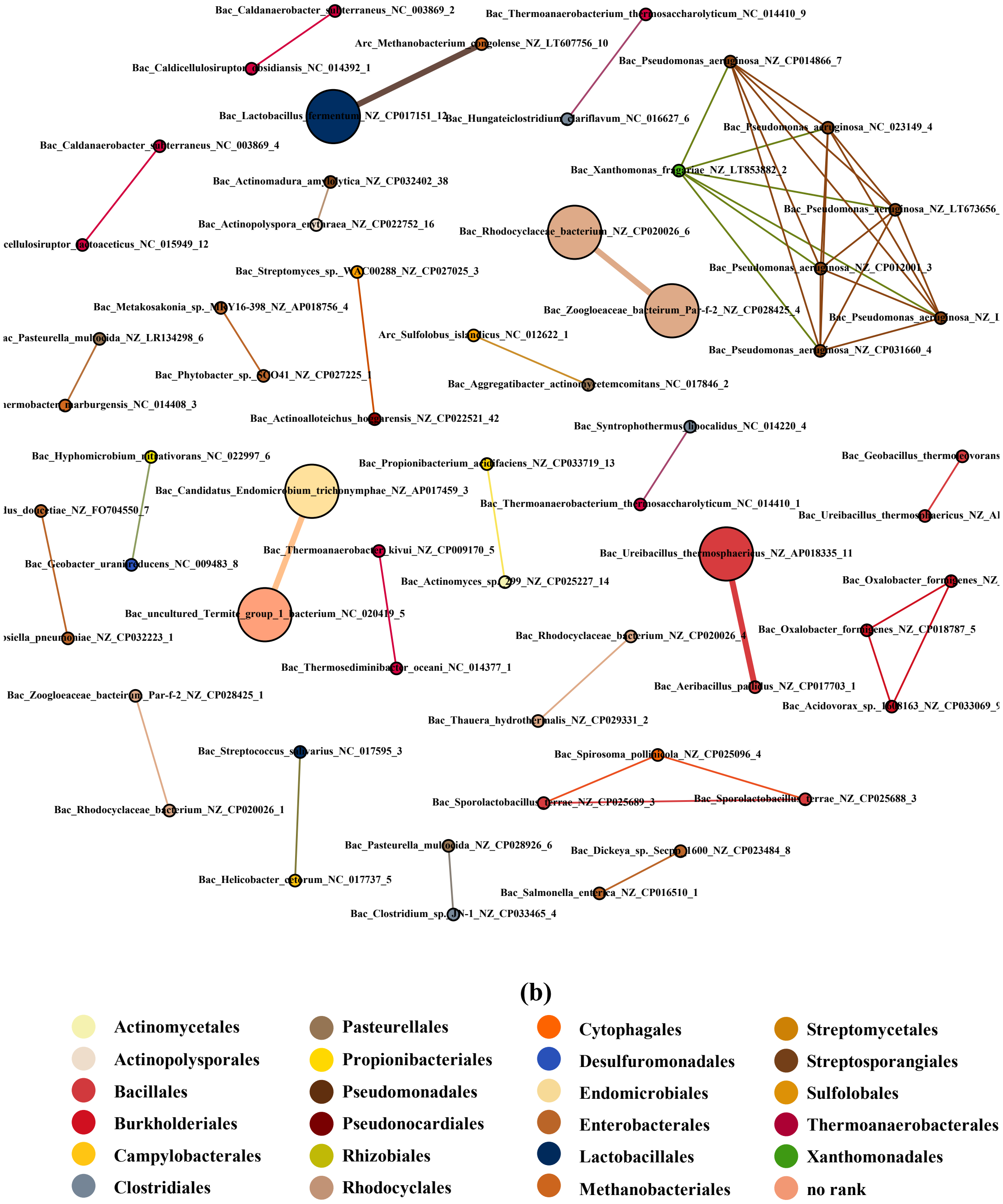
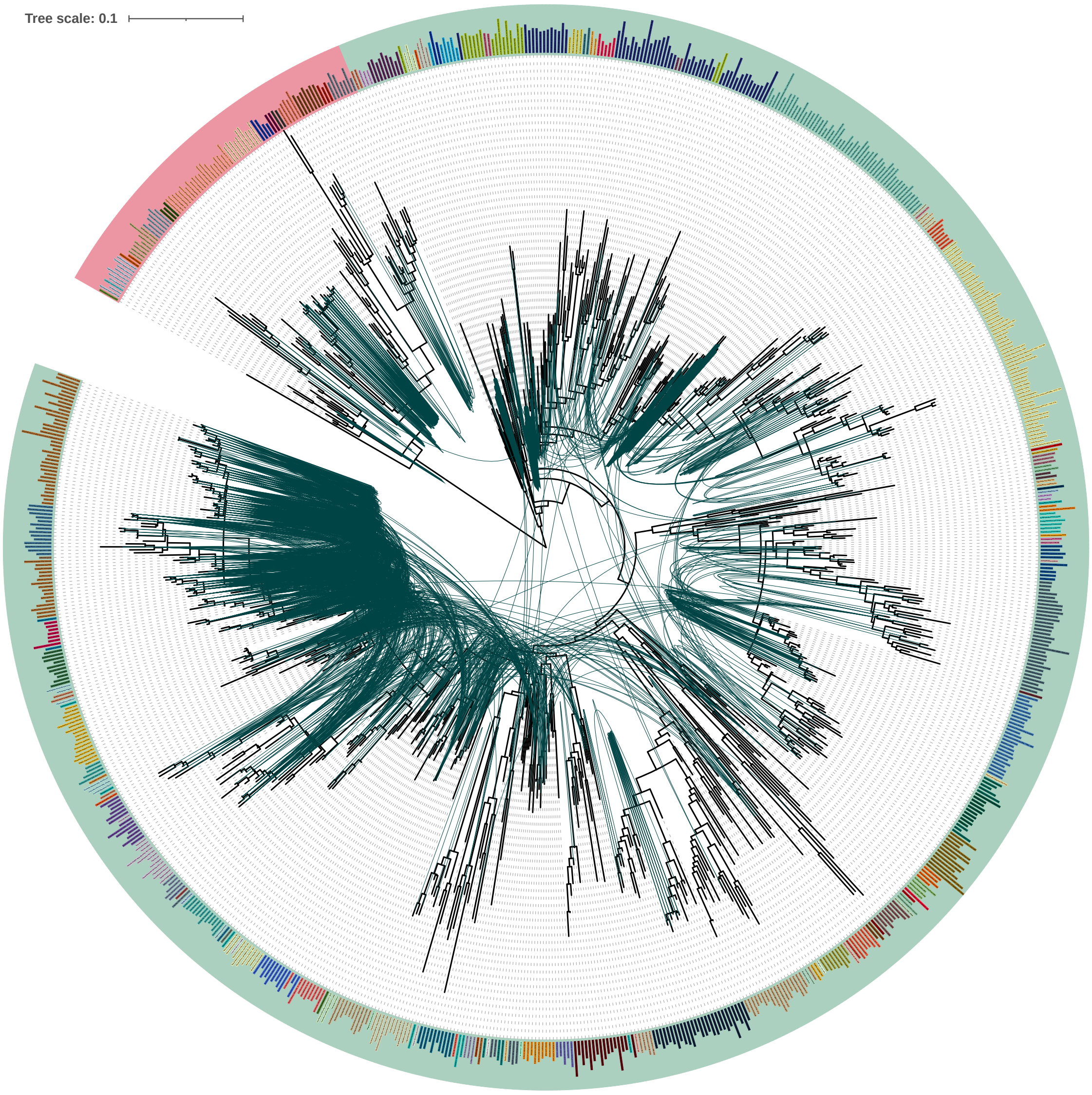
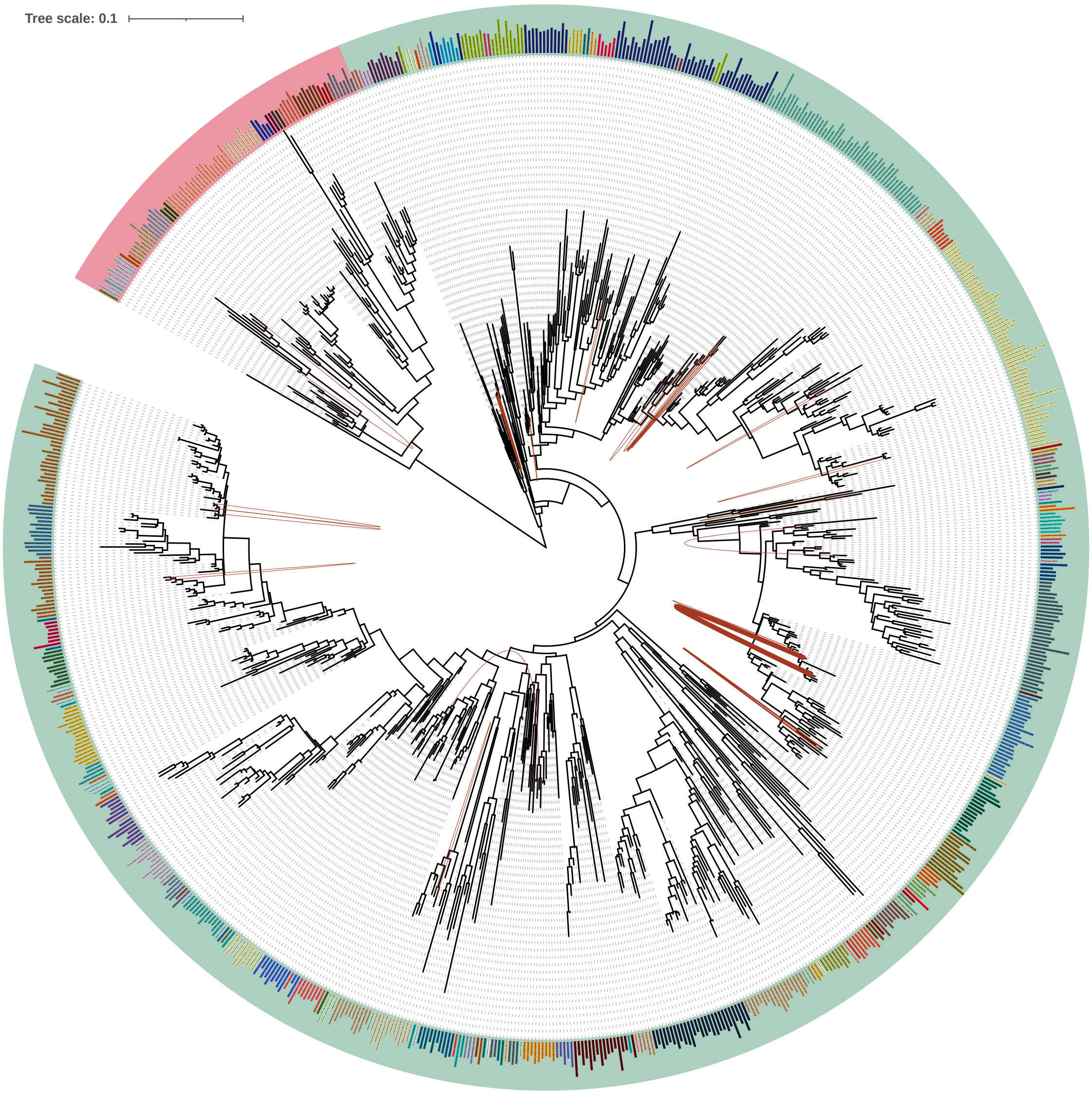

Bac_Caldanaerobacter_subterraneus_NC_003869_2
Bac_Thermoanaerobacterium_thermosaccharolyticum_NC_014410_9
Arc_Methanobacterium_congolense_NZ_LT607756_10
Bac_Caldicellulosiruptor_obsidiansis_NC_014392_1
Bac_Pseudomonas_aeruginosa_NZ_CP014866_7
Bac_Lactobacillus_fermentum_NZ_CP017151_12
Bac_Hungateiclostridium_clariflavum_NC_016627_6
Bac_Pseudomonas_aeruginosa_NC_023149_4
Bac_Caldanaerobacter_subterraneus_NC_003869_4
Bac_Xanthomonas_fragariae_NZ_LT853882_2
Bac_Actinomadura_amylolytica_NZ_CP032402_38
Bac_Pseudomonas_aeruginosa_NZ_LT673656_
Bac_Actinopolyspora_erythraea_NZ_CP022752_16
Bac_Rhodocyclaceae_bacterium_NZ_CP020026_6
cellulosiruptor_lactoaceticus_NC_015949_12
Bac_Pseudomonas_aeruginosa_NZ_CP012001_3
Bac_Streptomyces_sp._WAC00288_NZ_CP027025_3
Bac_Zoogloeaceae_bacteirum_Par-f-2_NZ_CP028425_4
Bac_Metakosakonia_sp._MRY16-398_NZ_AP018756_4
Bac_Pseudomonas_aeruginosa_NZ_I
ac_Pasteurella_multocida_NZ_LR134298_6
Arc_Sulfolobus_islandicus_NC_012622_1
Bac_Pseudomonas_aeruginosa_NZ_CP031660_4
Bac_Phytobacter_sp._SCO41_NZ_CP027225_1
Bac_Aggregatibacter_actinomycetemcomitans_NC_017846_2
iermobacter_marburgensis_NC_014408_3
Bac_Actinoalloteichus_hoggarensis_NZ_CP022521_42
Bac_Syntrophothermus_lipocalidus_NC_014220_4
Bac_Hyphomicrobium_nitrativorans_NC_022997_6
Bac_Propionibacterium_acidifaciens_NZ_CP033719_13
Bac_Geobacillus_thermoleovorans
Bac_Candidatus_Endomicrobium_trichonymphae_NZ_AP017459_3
llus_doucetiae_NZ_FO704550_7
Bac_Thermoanaerobacterium_thermosaccharolyticum_NC_014410_1
Bac_Ureibacillus_thermosphaericus_NZ_Al
Bac_Thermoanaerobacter_kivui_NZ_CP009170_5
Bac_Ureibacillus_thermosphaericus_NZ_AP018335_11
Bac_Geobacter_uraniireducens_NC_009483_8
Bac_Actinomyces_sp._299_NZ_CP025227_14
Bac_Oxalobacter_formigenes_NZ_
Bac_uncultured_Termite_group_1_bacterium_NC_020419_5
Bac_Oxalobacter_formigenes_NZ_CP018787_5
siella_pneumoniae_NZ_CP032223_1
Bac_Rhodocyclaceae_bacterium_NZ_CP020026_4
Bac_Thermosediminibacter_oceani_NC_014377_1
Bac_Aeribacillus_pallidus_NZ_CP017703_1
Bac_Zoogloeaceae_bacteirum_Par-f-2_NZ_CP028425_1
Bac_Acidovorax_sp._1608163_NZ_CP033069_9
Bac_Thauera_hydrothermalis_NZ_CP029331_2
Bac_Streptococcus_salivarius_NC_017595_3
Bac_Spirosoma_pollinicola_NZ_CP025096_4
Bac_Rhodocyclaceae_bacterium_NZ_CP020026_1
Bac_Sporolactobacillus_terrae_NZ_CP025689_3
Bac_Sporolactobacillus_terrae_NZ_CP025688_3
Bac_Pasteurella_multocida_NZ_CP028926_6
Bac_Dickeya_sp._Secpp_1600_NZ_CP023484_8
Bac_Helicobacter_cetorum_NC_017737_5
Bac_Salmonella_enterica_NZ_CP016510_1
Bac_Clostridium_sp._JN-1_NZ_CP033465_4
(b)
Actinomycetales
Actinopolysporales
Bacillales
Burkholderiales
Campylobacterales
Clostridiales
Pasteurellales
Propionibacteriales
Pseudomonadales
Pseudonocardiales
Rhizobiales
Rhodocyclales
Cytophagales
Desulfuromonadales
Endomicrobiales
Enterobacterales
Lactobacillales
Methanobacteriales
Streptomycetales
Streptosporangiales
Sulfolobales
Thermoanaerobacterales
Xanthomonadales
no rank

Figure 3.6 Networks of CRISPR loci across different genera (a) and families (b) that contain similar spacers. Nodes are regarded as CRISPR loci and coloured based on taxonomic orders. Nodes are connected by edges if a similar spacer is identified between two loci. The size of a node is correlated with spacer numbers within a locus and the weight of an edge is correlated with the number of shared spacers. Both networks are depicted by software Gephi (Bastian et al., 2009) with Fruchterman-Reingold layout (Fruchterman and Reingold, 1991).

Although we did not find many relationships between distant taxa groups in our networks, we wanted to investigate further by putting the network results in the context of the phylogenetic tree of life. Our full dataset contains thousands of strains, which can be easily handled using a network approach, but exceeds calculation capacity of the phylogenetic approach. Therefore, we used the published 16S rRNA tree from the LTP project and mapped to our resample dataset. To maintain species diversity, we selected 843 species (91 archaea and 752 bacteria) from 125 taxonomic orders (labelled in different colours, Figure 3.7) from the LTP dataset that contain identified CRISPR-Cas systems and pruned the phylogeny to create a subtree. The previous clustered outputs of spacer and repeat families from UCLUST (see Methods 2.5.1) were employed here. In both phylogenies, edges were drawn between species if they have commonly shared spacers or repeats, respectively. The weight of an edge represents the number of shared elements. The tree-networks revealed universal shared repeats between species (1,965 connections, Figure 3.7a), even including 1 remote hit between archaea *Ammonifex degensii* KC4 and bacteria *Sulfodiicoccus acidiphilus*. It has been reported that even short sequence homology between phages and plasmids (like newly required spacer) can promote transduction rates of plasmids dramatically (Deichelbohrer et al., 1985; Maniv et al., 2016; Varble et al., 2019). Also, complete CRISPR-Cas systems have been reported from bacteriophages (Naser et al., 2017; Seed et al., 2013). Therefore, we conjecture that the extensive shared direct repeats might provide the opportunity for transduction between CRISPR loci who have similar repeats. However, spacer sharing between species reduced dramatically (28 connections, Figure 3.7b) compared to repeat sharing. This result is consistent with previous network results but conflicts with our shared repeats results, which may indicate that other elements are at play that inhibit HGT of CRISPR-Cas systems.



(a)



(b)

Figure 3.7 Patterns of repeats and spacers sharing across prokaryotic species. The Phylogenetic tree is pruned from 16S rRNA tree from the LTP project that contains 843 species who possess identified CRISPR-Cas systems. The red shade represents species from Archaea while green shade represents species from Bacteria. Species names are coloured in the outer ring based on its taxonomic order. In the network, edges connect species who share similar repeat **(a)** or spacers **(b)**.

3.5 Discussion

In this study, we collected 12,184 fully sequenced prokaryotic genomes from the entire NCBI RefSeq dataset and performed CRISPR-Cas system identification using four published CRISPR identification tools (PILER-CR, MinCED, CRISPRCasFinder and CRISPRDetect). We found the results from the four programs quite different. Therefore, to reduce the bias of false positive results from different programs, we constructed a resample dataset by only selecting CRISPR-Cas systems that were identified using at least three software programs. The resample dataset then is regarded as a conservative CRISPR-Cas dataset from the perspective of bioinformatics predictions. However, among these four programs, only CRISPRCasFinder is able to identify *cas* genes and infer their subtypes. Thus, subtype information was derived entirely from the output of CRISPRCasFinder. In total, CRISPR loci were identified from 82.7% of species across Archaea and 40.6% of species across Bacteria whereas *cas* genes were identified from 69.7% and 37.3% of Archaea and Bacteria, respectively. This result is consistent with Makarova et al.'s detection of 13,116 complete genomes (Makarova et al., 2019) but the proportion in Archaea slightly lower than previously observed results from CRISPRFinder (84%) (Grissa et al., 2007). Meanwhile, the numbers of *cas* genes are similar to the recent update of CRISPRCasdb in which across species from Archaea and Bacteria, respectively 75.3% and 36% have identified *cas* gene cluster (Pourcel et al., 2020). Intriguingly, while identifying the position of CRISPR arrays and *cas* gene clusters, we found 227 CRISPR arrays located between protein-coding *cas* genes. This pattern was also identified as a subtype variant of subtype II-C in recent CRISPR-Cas classification (Makarova et al., 2019). However, the pattern in our results is not only limited to subtype II-C but widespread across 11 subtypes. Also, many of them occur with *cas* genes in the reverse orientation.

To explore the evolutionary processes at play in CRISPR loci, we took three different approaches. Processes like insertion, deletion, recombination and HGT were depicted through repeat-mutation patterns, aligned spacer dynamics with phylogenetic trees, and similarity networks. From repeat-mutation patterns, polarized integration was visualised by conserved mutation in repeats until the end of the array or the occurrence of mutations at the same site. However, we found that liner mutation was frequently interrupted by other activities, which resulted in the same mutation at the same site but scattered along the whole array. We propose that activities such as recombination or novel integration mechanisms can both cause this pattern. It has been reported that mutation in conserved LAS can also result in middle spacer acquisition in type II-A CRISPR-Cas systems, which is termed as ‘ectopic spacer integration”. For strains who lack LAS or with a mutant-LAS can potentially use other sequence within CRISPR loci as an anchor during integration to guarantee a precise position (McGinn and Marraffini, 2016). Although through ectopic spacer integration, mutations in repeats should still be reserved to duplicates, it reveals the possibility of other, undiscovered mechanisms in which spacers might integrate along with repeats duplication in other position rather the adjacent one. Another possible explanation for the patterns we observed is recombination. It could be led from within array rearrangement after all mutations have occurred. However, aligned groups of array families suggest the rigorous order of CRISPR loci were rarely affected by recombination within a locus. Nevertheless, recombination can also occur between loci on one host genome or across different organisms. During recombination, when new spacer acquired from other loci, the original spacer with repeat will be deleted (Deveau et al., 2008; Kupczok et al., 2015). Also, the recombination rate is associated with sequence similarity between the donor and recipient (Majewski and Cohan, 1999). We found that repeats were universally shared across prokaryotes using a phylogenetic approach, which suggests that recombination between different loci is possible. In addition, duplication of spacers, as a result of recombination (Kupczok et al., 2015; Lopez-Sanchez et al., 2012; Nickel et al., 2013), was identified in 26.4% array families. By contrast, the study from Kupczok et al. (2015) suggested that CRISPR evolution was mainly shaped by acquisition and pervasive deletion instead of recombination.

Ubiquitous new spacer gains and losses along history have also been reported here using aligned array families with phylogenetic trees. Similar loci were clustered into families based on spacer similarity. In addition, to trace the timeline of spacer dynamics, core genes from species in each family were detected and used to construct a representative species phylogeny. Gains and losses of spacers were reconstructed at each node. We found acquisition and deletion occurred across the phylogeny. In the examples we presented, spacers that were acquired in recent history are located downstream of old spacers, supporting the previous conclusion that spacers are not always ordered chronologically. Also, we found deletion normally occurred in the middle of an array, which was consistent with previous research (Lam and Ye, 2019). However, although spacers turnover diversely in

CRISPR loci, a special retention pattern was observed. Spacers located on the edge of an array were highly conserved across different array families, like an anchor. A similar model was identified in another study, which called “trailer end spacer” (Lam and Ye, 2019). However, study from McGinn2016 showed that the immunity ability of spacer in CRISPR locus is related with their position. The newly integrated spacer has the highest robustness when being attacked. Also, spacers located at the very beginning had more transcribed crRNA (Deltcheva et al., 2011; Elmore et al., 2013; Nickel et al., 2013; Richter et al., 2012a) and lead to more robust immunity, simultaneously. For example, crRNA from Position 1 is twofold higher than from position 5 (McGinn and Marraffini, 2016). Overall, the location of spacers is very important to a CRISPR locus. The selective rates and fitness effects behind spacer in different positions remain to be studied.

In brief, we applied multiple approaches to investigate and visualise the evolutionary processes like insertion (including polarized integration, ectopic integration, duplication, and priming), deletion, recombination, retention, and HGT of the CRISPR-Cas system. Jointly these results suggest a complex pattern of CRISPR-Cas evolution but the mechanisms behind ectopic spacer integration and potential recombination between different loci are worth further investigation.

Chapter 4.

A Study of Genes that Associated with CRISPR-Cas Systems

4.1 Abstract

CRISPR-Cas system, as an adaptive defence system, has pervasively spread in most Archaea and half of Bacteria. The distribution of CRISPR-Cas systems is non-uniform with respect to phylogenetic relationship. However, the elements and mechanisms that affect this distribution remain to be clarified. CRISPR-Cas has been reported to have high mobility and thus the fixation or abolition is likely to be determined by host and environment (Makarova et al., 2020). In the current study, 12,184 complete nucleotide genomes from Archaea and Bacteria were collected and the CRISPR-Cas identification and classification results were obtained from Chapter 3. To elucidate the effect of host genetic background on CRISPR-Cas system, we then subsampled a smaller dataset of 1,824 species. After annotation, all protein-coding genes were clustered into families, and we assessed the association between CRISPR-Cas and each gene family. Proteins that showed significant association were then functionally classified and grouped based on KEGG BRITE. We have found that the number of genes that co-occur with type II and type III are significantly higher than the number of genes that disassociate these CRISPR-Cas types, and they mostly function in metabolic pathways. Also, genes that are associated with CRISPR-Cas subtypes were likely to be compatible to the phylogeny, which indicates the consequences of shaping CRISPR-Cas distribution by host co-occurring genes.

4.2 Introduction

CRISPR-Cas systems became one of the popular themes in recent years due to their ease of use as a genome editing tool (Adli, 2018). As an adaptive immune system that commonly exists in most Archaea and approximately half of Bacteria, CRISPR-Cas system is composed of CRISPR locus and *cas* gene cluster. CRISPR loci contain diverse short sequences of previous invaders (known as “spacers”) that are separated by identical or near-identical short palindromic repeats (Karginov and Hannon, 2010). CRISPR arrays can be regarded as a memory bank of previous invaders in which spacers are cognate with different mobile genetic elements (MGEs). The defence ability of CRISPR-Cas is encoded by Cas proteins and involves three stages: adaptation, expression and interference (Marraffini, 2015). Adaptation takes place at the beginning of a defence procedure in which part of the invading DNA (or RNA) sequence is recognised and integrated into the leading side of the CRISPR array. In subsequent invasions, the CRISPR array is processed and matured into CRISPR RNAs (crRNAs) containing the spacer sequence. Subsequently, the crRNA can bind a new invading segment under complementary base pairing rule, and guides combined with the Cas protein for target destruction (Hille et al., 2018; Makarova et al., 2011). In prokaryotes, there are several mechanisms that help avoid cell death from autoimmunity. One of the most important ones is through the protospacer adjacent motif (PAM). PAM in phage can be recognised by Cas before cleavage and promotes distinction between self, versus non-self, sequences.

The latest study by Makarova et al. (2020) has revealed 2 classes, 6 types and 33 subtypes of CRISPR-Cas systems, hierarchically defined according to a series of features including signature *cas* gene identification, phylogenetic analysis, and even considering experimental data. They covered 13,116 complete bacterial and archaeal genomes and significantly expanded their finding from Makarova et al. (2015). In their research, CRISPR-Cas systems were identified in 85.2% of Archaea and 42.3% of Bacteria Makarova et al. (2011). The reasons behind this uneven and sparse distribution can be manifold. First, the distribution could be related to the genetic background. For instance, type II CRISPR-Cas systems are mostly identified in bacterial species. This distribution is possibly due to the associated RNase III. This enzyme plays a key role in pre-crRNA expression stage and is bacterial-specific (Garrett et al., 2011). Second, horizontally transferred CRISPR-Cas systems have been found in many species (Godde and Bickerton, 2006; Makarova et al., 2015; Marraffini and Sontheimer, 2008), which promotes the extensive spreading. However, fitness costs in different recipients could determine the maintenance and cause the scattered distribution. Additionally, a study (Bernheim et al., 2019) has proposed that loss or retention of successfully acquired CRISPR-Cas systems might be associated with host double-strand break (DSB) repair system. Another assumption is that CRISPR-Cas limited the horizontal gene transfer (HGT) of beneficial genes like antimicrobial resistance (AMR) genes (Zheng et al., 2020). Thus, the strong selection pressure of AMR genes may drive losses of CRISPR-Cas systems (Shehreen et al., 2019). Also, the discovery of orphan CRISPR loci (no adjacent *cas* genes) might also be result from independent loss in multiple lineages (Faure et al., 2019).

Therefore, we investigated the effects of the presence or absence of genetic elements in organisms on the distribution of different CRISPR-Cas systems. To understand the connection, we applied a co-occurrence study between gene-gene or gene-CRISPR-Cas systems. Genes that associate with, or disassociate one another more often than expected by chance may indicate functional relationships between the genes, or fitness contributions or costs to hosts. So far, gene co-occurrence thinking has been widely used in many studies (Al-Aamri et al., 2019; Kim and Price, 2011; Shapiro et al., 2017; Whelan et al., 2020), especially related to CRISPR-Cas system (Bernheim et al., 2019; Makarova et al., 2020; Shmakov et al., 2018a). Shmakov et al. (2018) predicted 79 genes that occurred with CRISPR-Cas systems. Also, functional analysis revealed the association with membrane proteins and signal transduction, especially in type III systems. Interference machinery of type III CRISPR-Cas systems can cleave double-stranded DNA (dsDNA), single-stranded DNA (ssDNA) and single-stranded RNA (ssRNA) (Peng et al., 2015; Zhang et al., 2016). An important step of RNA clearance is regulated by cyclic oligoadenylate (cOA), generated by the Cas10 subunit from ATP. Then, cOA binds and activates CRISPR-Cas Associated Rossmann Fold (CARF) domain in type III ancillary genes, like Csm6, and mediates an RNA universal degradation (Figure 4.4a) (Makarova et al., 2020; McMahon et al., 2020). This explains previous detected strong association between type III and signalling pathways. Consistently, other studies also discovered the intense link between CARF domains and type III systems (Makarova et al., 2014, 2020; Shah et al., 2019). However, around half of CRISPR-linked genes that identified by Shah et al. (2019) were found to lack of CARF domains and this needs further investigation.

CRISPR-Cas systems not only associate with ancillary genes, but also cooperate with other subtypes or immune systems (Dupuis et al., 2013a; Oliveira et al., 2014). This phenomenon may be attributed to the arms race between prokaryotes and phages (Hampton et al., 2019). It has been found that species that encode type II

CRISPR-Cas and type II restriction-modification (RM) systems simultaneously are more likely to have enhanced defence against infection in host cells (Dupuis et al., 2013a). Significant co-occurrence relationships were also detected between CRISPR-Cas and RM as well as CRISPR-Cas and Argonaute (ARGO) (Oliveira et al., 2014). Considering that foreign MGE-encoded CRISPR-Cas inhibitors (called "anti-CRISPR") are system-specific (Pawluk et al., 2018), a cell with multi-systems is able to target more invaders and provide two-tier protection. Correspondingly, 9% of bacterial genomes were identified with multiple clusters of Cas proteins (Bernheim and Sorek, 2020). Type I system was found to co-occur with type III (Staals and Brouns, 2013). Consistently, type I-F and type III-B have been found to work synergistically in *Marinamonas mediterranea* (Silas et al., 2017b). Phage with PAM mutations could evade clearance from type I-F but will be captured by the interference machinery from type III-B with the same crRNA.

In this study, to understand the association between genome protein-coding genes and CRISPR loci, we constructed a dataset that containing 1,824 diverse prokaryotic species. After identifying all protein-coding genes and CRISPR-Cas subtypes from nucleotide genomes, we performed a gene-locus coincident search. Contingency tables were built for each pair of gene-locus presence-absence numbers. Genes that significantly associated or disassociated with CRISPR-Cas were selected and categorised. In order to explore mechanisms behind those associations, we functionally annotated all co-occurring proteins using EggNOG. In addition, we constructed a heatmap of the distribution of CRISPR-Cas associated genes across different species. Though gene co-occurrence studies have been reported before, questions still remain unanswered and were investigated in this study: 1) Are there genes that are likely to function in CRISPR-Cas systems, but have not found yet? 2) What genes in host genomes are likely to influence the fitness cost of transferred CRISPR-Cas system? 3) Would genes that negatively assocaited with certain CRISPR-Cas genes have potential for being a new defence system?

4.3 Methods and Materials

4.3.1 Complete and Subset Dataset Construction

To investigate the distribution of CRISPR-Cas system across different taxonomic ranks, we collected 277 archaeal and 11,907 bacterial complete nucleotide genomes from The National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq) database (see Methods in Chapter 3) (O'Leary et al., 2016) in January 2019 as the complete dataset and fetched their lineage information from Python ETE3 packages.

4.3.2 CRISPR-Cas Distribution across Different Taxonomic Family

In order to determine how CRISPR-Cas systems are distributed within Archaea and Bacteria, we used the result of CRISPR-Cas identification and classification from Chapter 2 and calculated the proportion of species that contain CRISPR-Cas systems within each family in taxonomy. The family phylogeny was extracted and pruned from 16S Ribosomal RNA (rRNA) tree from The All-Species Living Tree Project (LTP) (Yarza and Munoz, 2014) and depicted using the iTOL webtool (Letunic and Bork, 2007). Also, to explore different subtype distribution, we coloured the CRISPR proportion bar of each taxonomic family based on the number of CRISPR-Cas subtypes it belongs to. As mentioned before, species may encode multiple Cas clusters belonged to more than one subtypes. For these species, we randomly selected one subtype as the representative. Also, the proportion of species who have had two subtypes and more than two subtypes were calculated respectively and marked this information on the periphery.

4.3.3 Proteins Prediction

To comprehensively analyse the association between CRISPR loci and other genes, we firstly need to clarify all protein-coding genes in prokaryotes. To analyse CRISPR-Cas distribution, we collected all available prokaryotic genomes, but this complete dataset is beyond our computer capacity to annotate all genes. Therefore, we retrieved a subset that contains 1,824 respective organisms that spread across 125 taxonomic orders from the complete dataset. The subset dataset comprises 843 species that possess identified CRISPR arrays and 981 species that do not. Next, we identified all potential protein-coding genes using Prodigal (version 2.6.3) (Hyatt et al., 2010) to predict and translate all proteins in the 1,824 species.

4.3.4 Gene Family Clustering

Similar genes were clustered into families for gene coincident search. Considering the size of our subset database, we performed a fast DIAMOND (version 0.9.30.131) all-versus-all search, using an identity cut-off of 30%, and an e-value cut-off 1e-5. Significant similar hits were identified and served as input for clustering into families using the MCL algorithm with the inflation value set to 2.0 (Van Dongen, 2000).

4.3.5 Detection of Genes that Associated and Disassociated with CRISPR

To find gene families that significantly co-occur with, and those that disassociated with CRISPR-Cas systems, we applied an association study that including three parts. First, investigation of gene families that co-occurred with CRISPR-Cas regardless of CRISPR-Cas classification. Second, comparison of gene families that co-occurred with different CRISPR-Cas types. Third, identification of gene families that co-occurred with different CRISPR-Cas subtypes. The test methods were similar, but we sampled different subset based on the questions about CRISPR-Cas system, types or subtypes. For example, to find out whether *geneA* is associated with CRISPR loci, we built a 2Ｘ2 contingency table including counts of species with present and absent of CRISPR loci and *geneA* like Table 4.1. We assumed if the presence of one gene family is significantly associated with the presence of specific CRISPR-Cas type, this gene family is potentially linked with this type from the perspective of function or possible contributions to fitness. On the other hand, if one gene family is more associated with the absence of CRISPR-Cas system, this may be because of negative interactions with CRISPR, or it could be because it is part of different resistance system and there is conflict between the two.

Table 4.1 Example of the contingency table of testing the co-occurrence between *geneA* and the chance of have identified CRISPR loci. A, B, C, D presents the number of species with presence and absence of CRISPR-Cas types and gene family.

		GeneA		Total
		Present	Absent	
CRISPR loci	Present	A	B	A+B
	Absent	C	D	C+D
Total		A+C	B+D	A+B+C+D

*OR = (A/C)/(B/D) = AD/BC

Here, a Pearson's chi-squared test (Python SciPy 1.3.1) is applied to determine the statistically significant association between CRISPR-Cas and gene family present in more than 10 species. P-value was set as 0.01 initially and then decreased to 9.39e-09 after Bonferroni correction. Subsequently, to further investigate how gene background effected CRISPR-Cas system, we classified all significantly related families into positive association, and negative association groups based on Odds ratio (OR) values (Szumilas, 2010). OR was calculated (Table 4.1*) for families who showed significant results. If OR is higher than 1, CRISPR-Cas and this family is positively related, and *vice versa*.

4.3.6 Functional Annotation

Genes in all associated gene families were annotated though the web-based software EggNOG mapper2 (Huerta-Cepas et al., 2015). To compare the function between positively and negatively CRISPR-linked genes, KEGG Orthology (KO) numbers of proteins were retrieved and categorised into different hierarchies (Kanehisa et al., 2019).

4.3.7 Associated Gene Heatmaps

We hypothesised that the distribution of different CRISPR-Cas systems is influenced by the presence or absence of other genes. To test this hypothesis, we created a heatmap of species with genes that co-occurred with different CRISPR-Cas subtypes. However, the entire heatmap is too large to show on paper and more suitable to present in digital form. Therefore, to visualise the trend of CRISPR-Cas subtypes linked gene, we selected the most significant 160 genes for each subtype. We choose 160 as the cut-off because this number is big enough to show the distribution but also within the visualise capacity. All associated genes were included for subtypes that co-occurred with fewer than 160 genes. In total, a smaller heatmap of species with total 1,640 CRISPR-linked genes was depicted by iTOL. Genes were coloured based on different linked subtypes while species were coloured according to taxonomic order.

4.3.8 Gene Association Network

Gene co-occurrence studies can also promote our understanding of mechanisms of CRISPR-Cas system. Species of Mycobacteriaceae have been widely used in studies of CRISPR type III (Grüschow et al., 2019; Wei et al., 2019). Here, to test the potential application of gene association studies for understanding CRISPR mechanisms, we applied a network approach on genes in Mycobacteriaceae that positively associated with subtype III-A as an example. In this example, we summarised all genes that significantly co-occurred with subtype III-A and used their Cluster of Orthologous Group (COG) membership as input for protein association network. I then analysed all potential connections such as gene neighbourhood, gene fusion, or co-expression, between proteins using STRING web-based tool (Szklarczyk et al., 2015).

4.4 Results

4.4.1 CRISPR-Cas Systems in Different Taxonomic Family

In the analysis described in Chapter 3, I analysed the genomes of 12,461 fully sequenced prokaryotic species (12,184 bacteria and 277 archaea) to identify their CRISPR-Cas loci. To minimise biases in evaluating genome content and CRISPR-Cas distributions, we analysed all complete Archaeal and Bacterial genomes available from NCBI RefSeq. This analysis showed that 82.67% of Archaea and 40.60% of Bacteria possess CRISPR systems. However, in this Chapter, closer analysis of this dataset revealed that the prevalence of CRISPR systems and their subtypes vary remarkably across taxonomic units (Figure 4.1). For example, CRISPR-Cas systems are present all species in Sulfolobaceae or Thermotogaceae while absent in all species in Nitrosopumilaceae or Chlamydiaceae. In addition, the most widespread subtype in Archaea is type I-B whereas in Bacteria is type I-E (followed by type I-C). Also, CRISPR-Cas subtypes were not restricted to specific clades (Figure 4.1) and different types of CRISPR-Cas systems often existed simultaneously even within a species. There are 3,792 species where I identified a single type of CRISPR-Cas system, while 666 species encode multiple CRISPR-Cas gene clusters, in which 568 species possess two subtypes and 98 species possess more than two subtypes (Figure 4.1). Overall, CRISPR-Cas systems spread sparsely among prokaryotes and none of these types is found solely in a single clade.

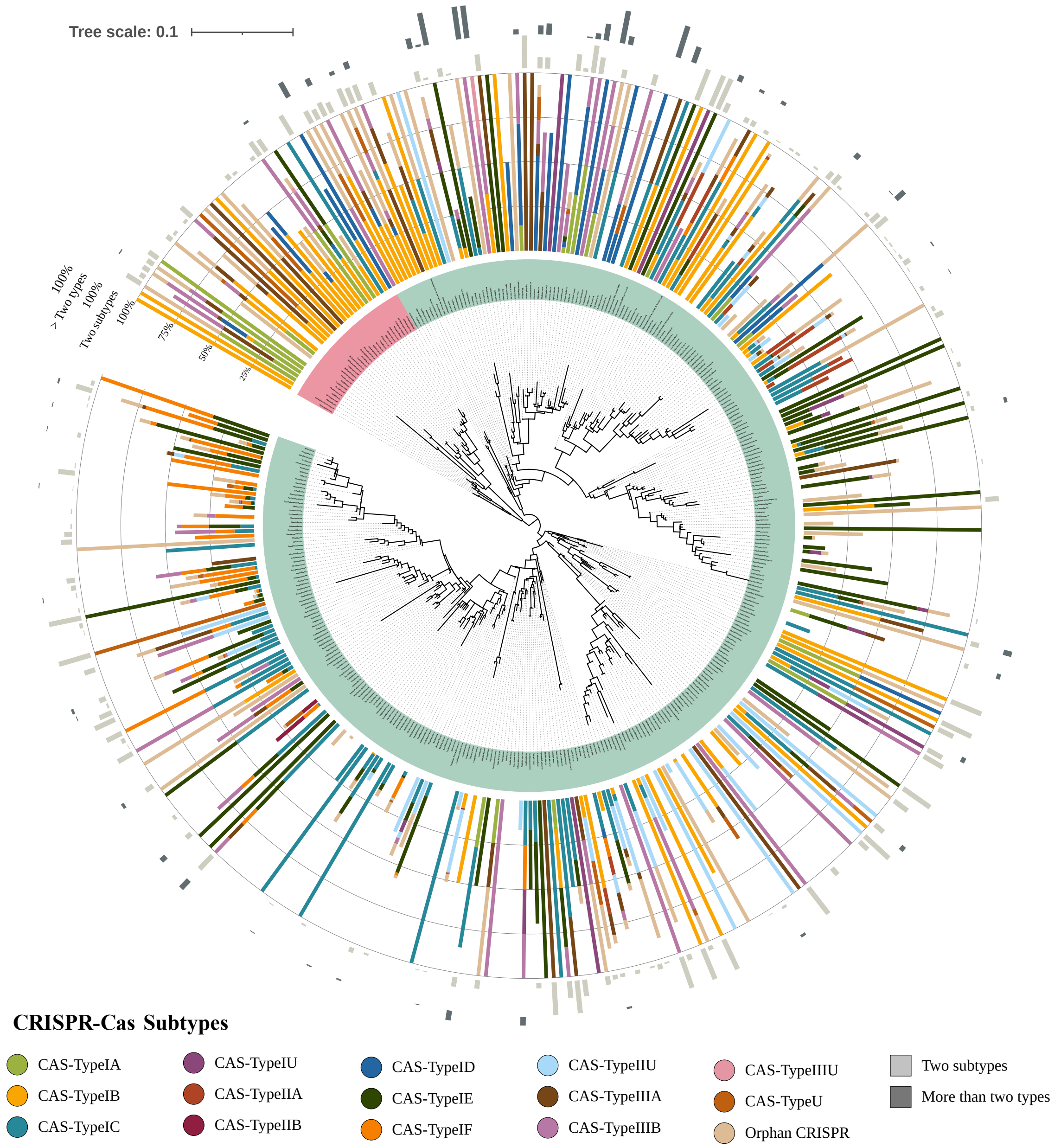


Figure 4.1 Distribution CRISPR-Cas systems across different taxonomic families. Each bar represents the percentages of species that have CRISPR-Cas systems in a taxonomic family and the bar was divided based on subtype proportions. The two bars on outer rings represent proportions of species who encode multiple CRISPR-Cas subtypes in one family. The family phylogenetic tree is pruned and collapsed from published 16s rRNA species tree of LTP project (Yarza and Munoz, 2014). Families with red background are from Archaea whereas in green background are from Bacteria.

4.4.2 Associated Genes with CRISPR-Cas

In order to comprehensively understand how genetic background affects the distribution and evolution of CRISPR-Cas system, we carried out gene-CRISPR co-occurrence studies for a dataset including 1,824 diverse species. CRISPR loci and Cas were identified and classified into subtypes in Chapter 3. To identify the different genes that co-occur with, and/or disassociated with CRISPR-Cas, all protein-coding genes in our dataset were annotated and clustered into gene families. In total, there are 1,064,556 gene families classified. For all gene families present in more than 10 species (59,092 families in total), a Pearson's chi-square test was carried out to detect the association between the gene family and the chance of possessing CRISPR-Cas system regardless of types. Among 59,092 tested gene families, 574 were predicted to co-occur with CRISPR-Cas while 984 were found to disassociate with CRISPR-Cas more often than is expected by random chance.

To investigate the mechanism behind association between CRISPR-Cas and genes, we then functionally annotated all associated genes using EggNOG (Huerta-Cepas et al., 2015). From the list of associated genes, KO numbers of each related gene families were retrieved and categorised according to KEGG BRITE hierarchies for comparison (Figure 4.2a). Since type classification was ignored in this test, we used the term "CRISPR loci" as the label. The number of genes that co-occur and disassociate with CRISPR loci is significantly different (t-test, p-value = 0.0058). Genes that disassociate with CRISPR loci are widely represented in most KEGG functions except terpenoids and polyketides metabolism, as well as translation. Meanwhile, co-occurring genes are mostly involved in metabolism and genetic information processing (Figure 4.2a). In functions such as carbohydrate metabolism, lipid metabolism as well as glycan biosynthesis and metabolism, genes that disassociate with CRISPR loci highly outnumber co-occurred genes.

Although different types of CRISPR-Cas systems all act as adaptive immune system in prokaryotes, they utilized different genes during adaptation, expression and interference. There are no genes that are found commonly encoded by all CRISPR-Cas systems (Makarova et al., 2020). For instance, in the interference stage, class I normally employ multiple genes as effector modules, while class II systems use single a gene with multiple domains (Makarova et al., 2015). Therefore, to understand gene-CRISPR associations for each type, we carried out chi-square tests to find out genes that significantly associate or disassociate with the three most frequent types of CRISPR-Cas systems (type I, II, III). These types are found in 81.14% of species in our dataset. In total, there were 353 families (131 positively, 222 negatively) linked with type I, 1,350 families (1,340 positively, 10 negatively) linked with type II and 299 families (290 positively, 9 negatively) linked with type III. Compared to CRISPR loci and type I, the number of genes that disassociate with type II and type III systems are very low.

Following a similar approach to the one used for the complete set of CRISPR-Cas systems, gene families that co-occurred or disassociated with each CRISPR-Cas type were functionally annotated using EggNOG and classified based on belonged KO numbers. Consequently, we observed a discrepancy on genes that related to different types. For all functions, gene families that co-occur with type I CRISPR-Cas loci did not show significant difference compared to families of genes that tend to disassociate with type I CRISPR-Cas (paired two-tailed t-test, p = 0.0186). By contrast, for the other two types, the difference between numbers of positively and negatively associated gene families are statistically significant (type II p-value = 0.0054, type III p-value = 0.0038). In addition, genes that co-occurred with type II CRISPR-Cas systems play important roles in metabolism including cofactor, amino acid and nucleotide metabolism, as well as signal transduction. A similar tendency also was observed in families that were positively associated with the presence of type III systems, except with more families involved in xenobiotics biodegradation and fewer in Carbohydrate metabolism.

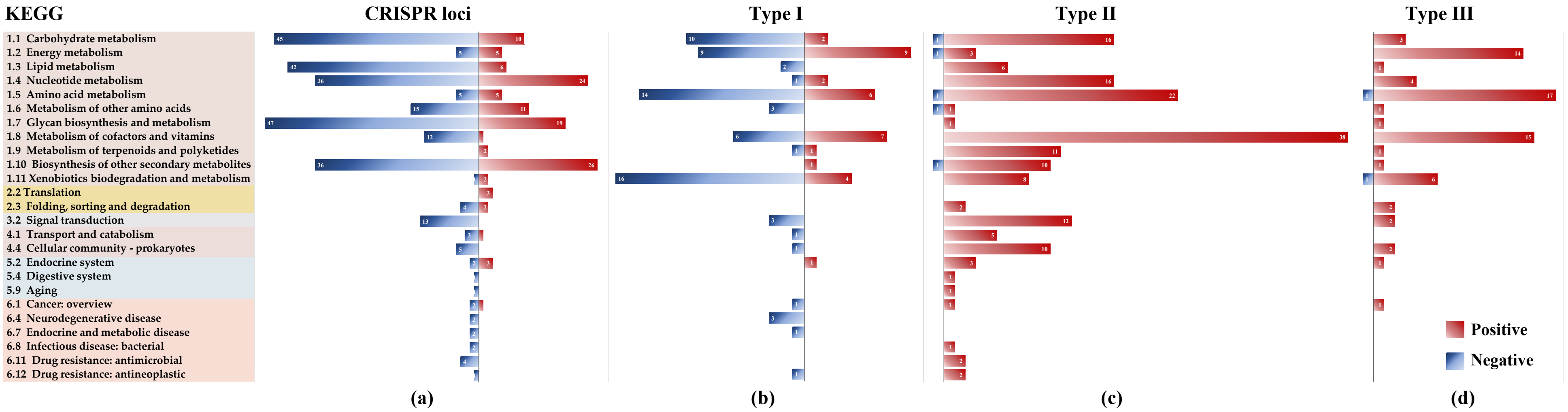


Figure 4.2 Comparison of KO numbers between genes that positively and negatively associate with CRISPR-Cas systems. Genes that had a statistically significant association with **(a)** CRISPR loci (regardless of types), **(b)** type I, **(c)** type II and **(d)** type III systems were functionally annotated and quantified with respect to their positive or negative association. Gene's KO numbers from functional annotations were retrieved and grouped based on KEGG BRITE hierarchies. The six main hierarchies are 1. Metabolism; 2. Genetic Information Processing; 3. Environmental Information Processing; 4. Cellular Processes; 5. Organismal Systems; 6. Human Diseases.

4.4.3 Genes Associated with Subtypes

Genes related to different types were observed have different functional preference. However, it is still unclear how these genes affect the distribution of CRISPR-Cas systems. Therefore, we analysed associations in more detail, by identifying genes that co-occur or disassociate with different CRISPR-Cas subtypes. The procedures were similar to those for detecting type-linked genes: gene families that showed significantly associations with each CRISPR-Cas subtype as assessed using a chi-square test, were selected and functionally annotated through EggNOG.

In total, we have identified 8,144 gene families that significantly associated with various CRISPR-Cas subtypes and 399 gene families that significantly disassociated with being with them in the same genome (detailed numbers of different subtypes are shown in Appendix C - Table S4.1). Genes that significantly associated with different subtypes were visualised via iTOL webtool and coloured based on associated subtypes. The complete heatmap is too large to display on paper and is available at https://itol.embl.de/tree/78145146138451531594157733. Thus, we subsampled the most significant 1,640 gene families that associated with CRISPR-Cas subtypes (Figure 4.3). As shown in the figure, for many subtypes, associated genes are mostly compatible with phylogeny, like I-A, I-F and II-A. Intriguingly, for subtypes like I-E dozens of genes with very significant associations are widely distributed across the tree, whereas other genes were restricted to Actinobacteria. By contract, genes that co-occur with subtypes like I-B are widespread across the tree of prokaryotes, except certain phyla like Actinobacteria and Proteobacteria.

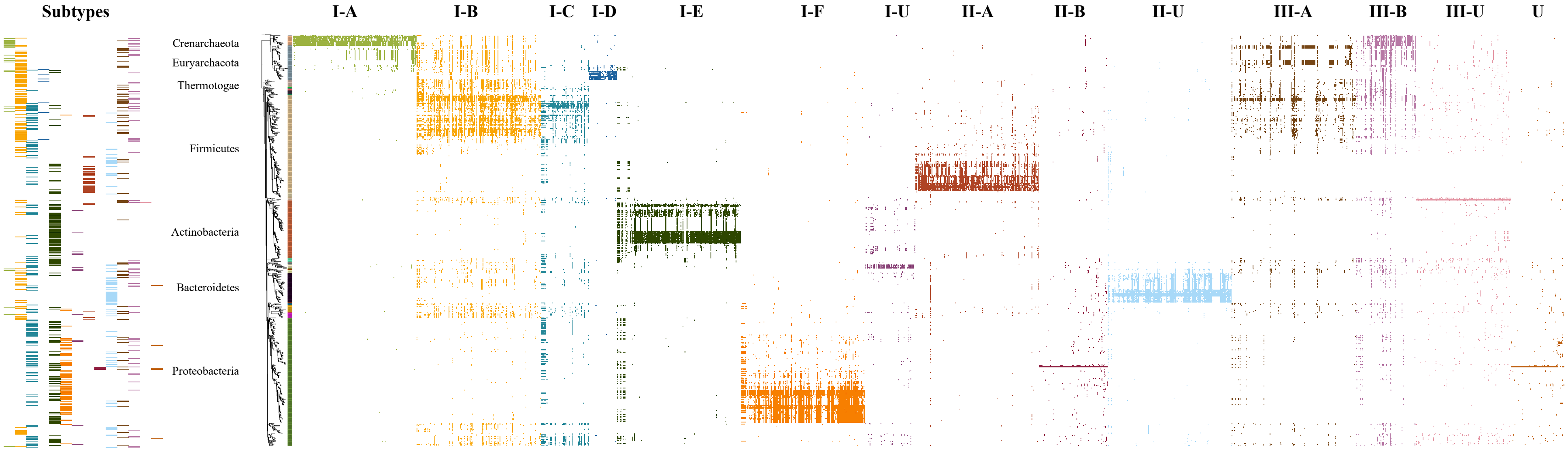


Figure 4.3 Heatmap of subset of genes that significantly associated with different CRISPR-Cas subtypes. It includes the most significant 1,640 genes that associated with different CRISPR-Cas subtypes. For each subtype, genes were ordered based on increasing p-value. That is the most significant co-occurring gene is situated on the very left side. The encoded CRISPR-Cas subtypes in each species are shown on the left of phylogeny, which are in the same colour with their corresponding associated genes. The species' phylogeny was pruned from LTP 16s rRNA phylogenetic tree, including 843 species that have identified CRISPR-Cas systems. Species in phylogeny were coloured based on their taxonomic phylum.

4.4.4 Subtype III-A in Mycobacteriaceae

Investigations of CRISPRs-linked genes can not only promote our understanding of the scattered distribution of CRISPR subtypes, but also facilitates study of the mechanisms by which the CRISPR-Cas subtypes might work. Here, we used subtype III-A in Mycobacteriaceae as an example. In our dataset, there are 308 strains of Mycobacteriaceae, 169 of them possess an identified subtype III-A system. After chi-square tests, 4 gene families were retrieved that positively associated with subtype III-A. The associated genes were subsequently annotated by EggNOG and the COG numbers were used as indexes for constructing a protein-protein association network via the STRING web tool (Figure 4.4b). As shown in Figure 4.4b, STRING constructs protein networks using eight different types of association. Node COG1353 in the Cas cluster was annotated as Cas10. Cas10 works as a nuclease and is clustered with other Cas proteins like Csm3 (COG1337) in the type III interference machinery (called Csm complex) (Niewoehner et al., 2017). In the STRING network, it connected with node COG2206 and node COG0664 in relationship of text mining and gene fusions. COG2206 is annotated as cyclic di-GMP phosphodiesterase class II (or its inactivated variant) in HD-GYP domain while COG0664 is annotated as complexes with cyclic AMP (cAMP)-activated global transcriptional regulator CRP. They work together in bacterial signalling pathways and both connect with COG1353 (Cas10).

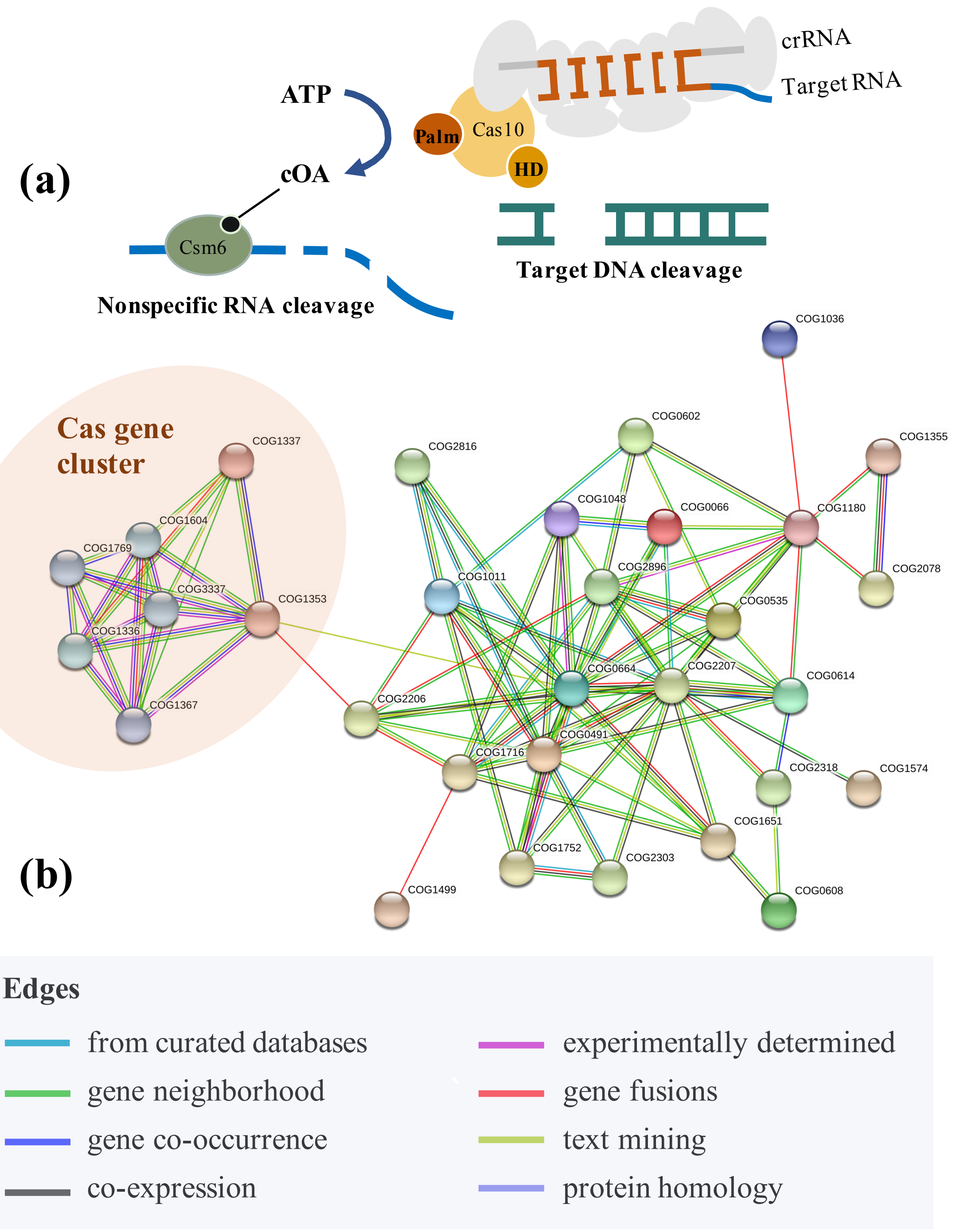


Figure 4.4 Interference mechanism and protein-protein interaction network related to CRISPR-Cas type III-A. **(a)** Type III-A system can abolish both DNA and RNA sequences. The interference is encoded by Csm complex that contains multiple Cas proteins and can be divided to three parts. Firstly, target RNA transcript is bind with crRNA and cleaved by crRNA-combined complex. Secondly, exogenous dsDNA is destroyed by HD domain of Cas10. Thirdly, Cas10 mediates the production of messager cOA from ATP. cOA then activates Csm6 and regulates a global RNA cleavage. Figure adapted from (Rouillon et al., 2019). **(b)** Protein-protein association network of genes in Mycobacteriaceae that associated with type III-A. Nodes represent proteins while edges represent protein-protein association. The detailed COG annotations were included Appendix C - Table S4.2.

4.5 Discussion

In this study, we analysed the distribution of different subtypes of CRISPR-Cas systems and attempted to elucidate genetic factors behind this pattern. We constructed two datasets: the large one is designed to investigate the broad distribution trend of CRISPR-Cas, which contains 12,184 completely sequenced prokaryotic genomes; the smaller one aims for exploring interactions between host genes and CRISPR-Cas, which is comprised of 1,824 species that retrieved from large dataset.

We observed that the distribution of CRISPR-Cas subtypes is not strictly related to phylogeny, which is consistent with a previous study (Makarova et al., 2020). The same study also proposed the most recent classification of CRISPR-Cas, which includes 6 types and 33 subtypes. However, considering the lag of software developments from up-to-the-minute discovery, our CRISPR-Cas subtype classification is still based on the old classification Makarova et al. (2015). Those *cas* gene clusters that could not be classified into known subtypes were grouped into type U. Despite the inevitable drawbacks from bioinformatics tools, we still identified three dominant CRISPR-Cas types (I, II, III), which account for 81.14% of all identified systems. Also, although diverse subtypes were detected in prokaryotes, I-B and I-C are dominant in Archaea and Bacteria, respectively. This result is in line with the study from Makarova et al. (2015).

To find the genes that co-occurred with CRISPR-Cas, we constructed 2x2 contingency tables for every gene family including the numbers of presence and absence of CRISPR-Cas and this gene family in our dataset. Then, genes that significantly associated with CRISPR loci and different types were functionally annotated and categorized to evaluate the existence of associated functional pathways. There are several possible mechanisms that affect the scattered distribution of CRISPR-Cas (Bernheim et al., 2019; Garrett et al., 2011). HGT of CRISPR-Cas has been reported to be common in prokaryotes (Chakraborty et al., 2010; Makarova et al., 2015). However, whether the transferred system will be fixed or lost by recipient is likely related to the host's genetic background, such as DSB repair systems RecBCD and AddAB. These systems have been reported that may involve in protospacer processing during adaptation stage (Bernheim et al., 2019; Radovčić et al., 2018). In addition, a gene that positively associates with CRISPR-Cas could have roles in operating or regulating CRISPR-Cas during defence (Radovčić et al., 2018), or confer fitness advantages to niches (Shehreen et al., 2019). By contrast, we found the number of genes that disassociated with CRISPR loci is higher than co-occurring genes and the coverage of functions is wider. This may indicate the influence of mechanisms shaping the distribution of CRISPR-Cas. This disassociation relationship can be elucidated from two sides. On one hand, a gene that significantly conflicts with CRISPR-Cas probably is a barrier for CRISPR-Cas retention. On the other hand, CRISPR-Cas system in host genome might constrain HGT of other genes. Many studies have reported the limitation effects from CRISPR-Cas to HGT (Marraffini and Sontheimer, 2008; Nozawa et al., 2011). This also has been supported by comparative analyses which revealed species in diverse environments tended to have inactivated or deleted CRISPR-Cas system. They suggested that species without CRISPR-Cas have more chance to obtain beneficial, adaptive genes and better adapt to stressed environments Zheng et al. (2020). However, in the gene-type co-occurrence study, negatively associated genes are quite different in three types, which may imply the different gene interactions behind different types. In addition, some of the KEGG categories like cancer or neurodegenerative diseases are not expected to be identified from prokaryotic proteins. These categories that significantly associated with CRISPR-Cas systems are likely to originate from very ancient history, or alternatively, have been incorrectly functional annotated. Bonferroni correction was applied for multiple tests in this study, which adjusted the p-value from 0.01 to 9.39e-09. Bonferroni correction was thought to be a very conservative adjustment, especially when applied in numerous tests simultaneously (Chen et al., 2017). Therefore, some CRISPR co-occurred genes are likely to be falsefully filtered out after correction. However, this study mainly concentrated on the function and distribution of genes that positively and negatively with CRISPR-Cas system. The current results of CRISPR-Cas mechanism analysis and subtype distribution would not be dramatically affected due to the conservativeness issue from Bonferroni correction. Also, due to the limitation of A4 paper frame, almost a half of CRISPR subtypes associated genes with lower p-value were hidden when analysing the effects of genetic background to subtype distribution (Figure 4.3). In further research, other less stringent adjustments like Benjamini-Hochberg adjustment could be considered when analysing all possible CRISPR-associated genes.

In order to further explore the relationships between CRISPR-Cas subtypes and linked genes, all associated genes were depicted using a phylogenetic tree associated with a presence-absence matrix. We could identify that the distributions of genes that were associated with many subtypes such as I-A, II-A and II-U, were compatible with archaeal and bacterial phyla. CRISPR immunity in some subtypes requires collaboration with nucleases from the host, especially during the expression and interference stage (Behler et al., 2018; Hille et al., 2018). In recipients that lack the required enzyme, CRISPR-Cas systems may not function properly. This may explain why CRISPR-Cas genes have been reported to potentially elicit dormancy or programmed cell death in response to immunity failure (Koonin and Zhang, 2017; Künne et al., 2016; Makarova et al., 2012). In addition, although some subtypes (like I-E) were found to be widespread across Bacteria, there were only a dozen linked genes, in agreement with their broad distribution. The remaining associated genes are only present in Actinobacteria, which probably indicates that the phyletic spread of CRISPR-Cas sometimes may be influenced by many genes (like Actinobacteria) but sometimes a few genes (like other species with type I-E). The specific mechanisms were not probed in this study, but our results could provide directions for further CRISPR-Cas-related research. Furthermore, considering negatively associated genes were only detected from type I-B and type I-F, our data is currently not enough to investigate the inhibition of different CRISPR-Cas subtypes to HGT or the effects from avoidant genes to the distribution of CRISPR-Cas loci, although it can suggest directions for further co-occurrence and avoidance studies.

In the last section, we aimed to expand the application of gene co-occurrence analyses into an investigation of CRISPR-Cas mechanisms. In our example, we applied a network approach on type III-A associated genes that are found in Mycobacteriaceae. Interactions between Cas10 (COG1353) and signal pathways

(COG2206: HD-GYP domain and COG0664: CRP) were observed. This result is consistent with a previous finding regarding Cas10 defence in CRISPR-Cas type III-A. Type III-A can target both DNA and RNA sequences. Upon CRISPR-Cas interference, the HD domain in Cas10 cleaves double-stranded DNA while the Palm domain mediates the production of cOA from ATP (Figure 4.4a). Following, the generated cOA serves as a second messenger, binding Cmr6 and activating nonspecific RNA degradation (Kazlauskiene et al., 2017; Rouillon et al., 2019). The positive regulation of type III-A by CRP signal pathway also has been reported by Agari et al. (2010). However, Cmr6 was not identified that associated with type III-A in our analysis. This is because the Cmr6 gene family is only present in 10 species and were ignored in the significance test.

In conclusion, we applied gene co-occurrence studies on three levels of CRISPR-Cas systems to investigate the interactions between genetic background and CRISPR-Cas loci. The spread of CRISPR-Cas subtypes is compatible with phylogeny and sometimes can be determined only by a dozen closely associated genes. The wide range of repellent genes may also indicate inhibition of HGT by CRISPR-Cas. Meanwhile, we applied a network approach on associated genes to verify our gene association analysis and detected a significant interaction between type III-A immunity and signal transduction pathway. Overall, our study demonstrates the usefulness of gene-gene association studies for understanding the function of CRISPR-Cas systems which can point the direction for further *in vivo* research.

Chapter 5.

Discussion, Future Work and Conclusion

5.1 General Discussion

Network-based models provide fresh angles of identifying and analysing introgressive descent and expand the evolutionary thinking (Bapteste et al., 2013). In a sequence similarity network (SSN), genes, genomes or organisms are represented as nodes and are connected by edges if they show significant homologous relationships. Hybrid nodes in the SSN enable its ability to identify gene remodelling events. Composite genes could be detected in non-transitive triplets, in which composite genes show significant similarity to component genes while there is no overlap between components (Corel et al., 2016; Watson et al., 2019). Horizontal gene transfer (HGT) events between species through mobile genetic elements (MGEs) can also show by SSNs (Fondi and Fani, 2010). Also, SSNs are flexible and can be adapted to different types of research. Apart from showing similarity between entities, nodes in a network can be coloured with additional information, such as modularity, taxonomic ranks and functional categories, whereas edges can be directed to show divergence or convergence, or be weighted based on similarity to parental/offspring genes. The characters of network allow scientists to navigate and manipulate according to their research objective.

To explore the potential of network approaches to analyse introgression of large-scale data, this thesis covers two datasets: one containing more than 1 million amino acid sequences to investigate gene fusion across three domains of life and MGEs; the other containing 12,184 complete nucleotide genomes to investigate the evolutionary history of prokaryote CRISPR-Cas systems.

In the first study, through identifying significantly similar sequences and constructing SSNs, a total of 221,045 composite genes were detected. These composite genes accounted for 18.57% of the complete dataset, connecting with 603,604 component genes. Although composite genes were widely distributed across all cellular organisms of our dataset, the proportion of composite genes within a species varied. 31.6% of genes in *Homo sapiens* were identified to be composites, which is the largest percentage among species across the three domains of life, followed by *Volvox carteri* f. *nagariensis* (29.77%). In addition, the average percentage of composite genes in eukaryotes is higher than in Archaea and Bacteria. To verify this result, we functionally annotated all composite and non-composite genes, and grouped them according to their origin and involved COG categories. Odd Ratio (OR) tests were performed to assess the association between a gene's origin and the likelihood of it being composite in each COG category. We concluded that the composite genes were more likely to have originated from eukaryotes than from prokaryotes, which is consistent with the initial result. The role gene fusion played in the origin of eukaryotes has been stated in many previous studies (Brennan et al., 2008; Leonard and Richards, 2012; Liu et al., 2009; Rogers et al., 2009) and a large number of composite genes were also identified in eukaryotes by Jachiet et al. (2013).

In the second study, we focused on the evolution of the CRISPR-Cas system, which is an important adaptive immune system in prokaryotes. Diverse studies investigating CRISPR-Cas systems including the evolution, mechanisms and applications have flourished in the past decades (Cong et al., 2013; Hille et al., 2018; Makarova et al., 2020). Here, to investigate possible HGT events in CRISPR-Cas systems, we constructed SSNs in which nodes represented CRISPR loci while edges represented evolutionary relationships such as sharing significantly similar spacers. Networks do not replace other evolutionary approaches such as tree-based models. Instead, a combined approach can promote our understanding of evolutionary biology (Corel et al., 2016). Therefore, we mapped the network results onto a phylogenetic tree. From both models, we found spacer sharing was rare between distant species but common between close species, whereas repeats were shared universally. This result is compatible with the hypothesized "pan-immune model" in prokaryotic population (Bernheim and Sorek, 2020). In this model, all immune systems are not required to be encoded by individual species but could be shared through HGT within a community. However, we only detected a few shared spacers across distant species, which suggests that the time or distance scale of pan-immunity is limited. That is to say, either CRISPR-Cas systems could not be transferred between distant species through genealogical and geological barriers, or transferred CRISPR-Cas could not be maintained. Although HGT of CRISPR-Cas

(Faure et al., 2019; Makarova et al., 2013, 2015) has been identified in different species, a study by Gophna et al. (2015) suggested that the effect of HGT on CRISPR-Cas was not significant in long time-scales, which is compatible with our latter hypothesis.

To depict diverse evolutionary processes such as integration, deletion, and recombination in CRISPR-Cas system, we used repeat-mutation patterns and comparative genomic analyse. Network thinking can also contribute to this task. Lam & Ye (2019) presented a directed compressed spacer network (graph) that connected spacers based on their order in CRISPR loci. In a compressed spacer network, conservative and uninterrupted spacers are united in one single node and edges with arrows indicate the direction and order of spacers. The main structure is stabilized by core spacers that shared between species. Rapid spacer loss were captured through identifying triangular motifs in the network. Multiple gains on the leading sides were also noticed through a radial pattern on one side of the structure. Lam and Ye's reported diverse spacer dynamics are consistent with our findings; however they did not identify ectopic spacer integrations because spacer integration time was not concluded in their work. Similarly, Kupczok et al. (2015) also identified CRISPR-Cas evolutionary patterns from graphs. In this work, they focused on detecting order divergence events (ODEs) from the comparison of two arrays that contained similar spacers. They concluded that adaptation and deletion contributed more to CRISPR evolution compared to recombination. This result is consistent with our comparative analysis but leaves us the question: which evolutionary process causes the irregular patterns of mutation in repeats? Although ectopic spacer integration has been reported in subtype II-A (McGinn and Marraffini, 2016), the irregular mutations in our results were widely distributed across all 11 subtypes. The underlying reason for this remains to be elucidated.

Another interesting finding are the highly conserved spacers at the end of CRISPR loci, which has been termed "trailer end clonality" (Weinberger et al., 2012). This sequence pattern has been found in many studies (Mick et al., 2013) and a metagenomic stability study found a conserved spacer that lasted for 5 years (Sun et al., 2016). The order of spacers has been reported to be critical to an array (McGinn and Marraffini, 2016). Spacers that are newly integrated (located at the leader end) possess higher resistance against phages by expressing more crRNAs than spacers in the middle (McGinn and Marraffini, 2016; Nickel et al., 2013; Richter et al., 2012b). This observation inspired our hypothesis about intention of priming. CRISPR-Cas can integrate more than one copy of the same spacer when under attack of a known invader, which is called priming. However, when encountering new MGEs, novel spacers will be integrated into the leading side again. Therefore, to maintain the robust immune response against familiar invaders, the lower number of transcribed crRNAs from a single downstream spacer can subsequently be made up by multiple copies of the same spacer in one array. Nonetheless, this hypothesis cannot explain the trailer-end clonality because we only identified one sole spacer conserved at the trailer end. Some other hypotheses about this pattern were stated. Diverse spacers encoded at the leading side are involved in constant expansion and contraction during evolution. Comparatively, spacers at the trailer end are conserved across a population in order to limit diversity of a CRISPR loci, which could possibly contribute to the maintenance of CRISPR-Cas (Haerter et al., 2011). Additionally, through a model simulation with metagenomic analysis, Weinberger et al. 2012) proposed that the clonality might be mediated by rapid selective sweeps of robust immune lineages.

The spread of CRISPR-Cas is non-uniform across the prokaryotic phylogeny and the reasons underlying this have not been described systematically before (Makarova et al., 2020). In my thesis, I utilized an association study on exploring the co-occurrence and avoidance between protein-coding genes and the CRISPR-Cas system. I identified a series of genes that co-occurred and disassociated with CRISPR-Cas types and found that these genes were mostly functional in different metabolic pathways. Through mapping the CRISPR-Cas subtypes and co-occurring genes to the phylogeny, we observed that the distribution of subtypes was diverse. This result is consistent with previous reports (Bernheim et al., 2019; Makarova et al., 2011). Specifically, subtype spreading is compatible with the distribution of positively associated genes. The number of highly co-occurring genes range from a dozen to hundreds, which indicates the influence of the genetic background to distribution of CRISPR-Cas. Considering that the HGT of CRISPR-Cas loci has been identified in our previous findings and many other studies (Faure et al., 2019; Shmakov et al., 2015), we conjectured that the distribution is influenced by the fitness trade-off of HGT and the CRISPR-Cas system. On one hand, the CRISPR-Cas system itself in prokaryotes may inhibit HGT (Bikard et al., 2012; Van Houte et al., 2016). It was proposed that species may deactivate CRISPR-Cas to promote HGT of beneficial genes in stressful environments, which serves as a "bet-hedging" strategy (Jiang et al., 2013; Zheng et al., 2020). Alternatively, the genetic background of the host may influence the fixation of the transferred system. This is consistent with the recent finding of an association between double-stranded DNA break (DSB) repairing systems and CRISPR-Cas subtypes (Bernheim et al., 2017, 2019). Moreover, diverse mutation rates in bacterial genomes also affect the distribution of CRISPR-Cas (Chevallereau et al., 2020). They indicated that bacteria with mutated surface genes, especially phage receptors, contributed to a fitness advantage and resulted in reducing of CRISPR-Cas evolution. Aside from the influences from internal factors of prokaryotes, the distribution of CRISPR-Cas could be affected from external factors of two levels: interaction between bacteria and phage, and ecological conditions. First, co-evolution studies have revealed the dynamics of CRISPR-Cas evolution to cope with constant phage infections (Common et al., 2019; Hampton et al., 2019; Paez-Espino et al., 2015). However, many CRISPR-Cas losses were observed in host-phage coevolution at a system level rather than a spacer level (Jiang et al., 2013; Weissman et al., 2018; Westra and Levin, 2020). A recent finding also revealed that a prophage in bacterial genome, deriving from temperate phage infection, could trigger CRISPR-Cas self-targeting. This could lead to loss of the CRISPR-Cas system (Rollie et al., 2020) through cytotoxic effects (Vercoe et al., 2013). Second, prokaryotic niches may also contribute to the sparse distribution of CRISPR-Cas system. Metagenomic experiments have revealed high prevalence of CRISPR-Cas in high-temperature environments (Anderson et al., 2011) and marine sponge-associated microbes (Horn et al., 2016) but shrunk popularity in groundwater filtrates (Burstein et al., 2016). Also, bacteria may favour different immune systems under different environmental conditions. For example, it was found that *Psedomonas aeruginosa* preferred CRISPR-Cas in nutrient deficient conditions while preferring surface mutations in nutrient excess conditions (Westra et al., 2015). Overall, the evolution of CRISPR-

Cas is affected by multifaceted processes, and therefore, good analyses may require thinking from the perspective of interdisciplinary methodology that combine bioinformatics, mathematical models with metagenomics (Westra and Levin, 2020).

5.2 Future Work

Regarding the study of composite genes, the general distributions and functional annotations of these results have been analysed in detail. A further step could focus on specific genes such as antimicrobial resistance (AMR) genes. In addition, corresponding RNAseq data could be included to map identified composite genes precisely, for example, by confirming the expression of these genes. This could also promote the understanding of the mechanisms behind composite genes generation. Furthermore, through EggNOG functional annotations, in 23 COG categories, only composite genes within the RNA processing and modification (A) as well as extracellular structures (W) categories show no preference between eukaryotes and prokaryotes. Genes belonging to these groups could be further investigated, in which the convergent or divergent evolutionary patterns could be mapped in the context of phylogeny.

Concerning CRISPR-Cas evolution, complicated evolution has been identified in this thesis. Inspired by the thought of identifying evolutionary patterns through mining different motifs in spacer networks (Kupczok et al., 2015; Lam and Ye, 2019), further research could focus for developing a pipeline of detecting insertion, deletion and recombination using network-based models. However, several potential problems may hinder such identification. First, large populations of long CRISPR arrays may include too many complicated patterns to analyse, whereas a small dataset may miss important evolutionary processes because of limited comparisons between CRISPR loci. Therefore, it is important to determine the proper threshold of dataset scale. Second, one motif could result from different kinds of evolutionary processes. For example, a triangle in a spacer network could be generated from middle insertion (or recombination, or HGT of a single spacer), or resulted from middle deletion. To comprehensively identify these process, a combination of network thinking, phylogenetic approaches and mathematical models could be considered. Third, recombination and duplication can disrupt the conservative order of core spacers, which could result in chaos within the network.

A large number of genes were identified that co-occurred with different CRISPR-Cas subtypes. Even though we employed four programmes to identify CRISPR loci, the results of Cas protein cluster detection are only based on one program CRISPRCasFinder, which classify subtypes according to the Makarova et al. (2015) method rather than the latest classification (Makarova et al., 2020). In future work, more Cas proteins and subtypes could be detected through new programs, such as CRISPRCasIdentifier (Padilha et al., 2020) and CRISPRCasTyper (Russel et al., 2020). In this study, the phylogenetic tree with heatmap was depicted to show relationships between the genetic background and distribution of CRISPR-Cas. However, the full heatmap is too long to clearly present in one A4 paper. In the future work, a bipartite network of subtypes and the associate genes could be involved in portraying the association. Furthermore, the role of avoidant genes in the distribution of CRISPR-Cas could also be further identified through network-based and tree-based approaches.

5.3 Conclusion

In this study, the power and application of network-based approaches is shown through investigations into composite genes and the CRISPR-Cas system. We constructed SSNs to identify composite genes in cellular organisms and find CRISPR sharing across prokaryotes. More specifically, we demonstrated that although composite genes were pervasive in prokaryotes and eukaryotes, the likelihood of composite genes deriving from eukaryotes was significantly higher than from prokaryotes. Separately, we found that CRISPR-Cas loci are involved in rapid and complex evolution. We identified and visualised integration, deletion and recombination through tracking mutations in repeats and comparative analysis of clustered CRISPR loci. Quite a number of irregular patterns might be produced by other spacer integration mechanisms except polarized adaptation and priming. Also, HGT in CRISPR-Cas was analysed through phylogenetic and network approaches. Distant sharing might be restricted but CRISPR-Cas sharing between close species is quite common, which potentially results in its unbalanced distribution across prokaryotes. However, we found that even though the spread of CRISPR-Cas is scattered, the subtype distribution is compatible with co-occurring genes. This suggests that species that encode CRISPR-Cas are affected by their genetic background. The number of closely related genes vary from a dozen to hundreds depending on the subtype. A protein-protein network was also constructed containing genes that co-occurred with CRISPR-Cas subtype II-A. Genes that connected with this Cas gene cluster can mediate RNA cleavage during interference, which have been proved in experiments. This suggested the potential application of co-occurrence and networks in exploring gene functions and mechanisms.

References

Abby, S.S., Néron, B., Ménager, H., Touchon, M., and Rocha, E.P.C. (2014). MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. PLoS One *9*, e110726.

Abudayyeh, O.O., Gootenberg, J.S., Konermann, S., Joung, J., Slaymaker, I.M., Cox, D.B.T., Shmakov, S., Makarova, K.S., Semenova, E., and Minakhin, L. (2016). C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science (80-.). *353*.

Adli, M. (2018). The CRISPR tool kit for genome editing and beyond. Nat. Commun. *9*.

Agari, Y., Sakamoto, K., Tamakoshi, M., Oshima, T., Kuramitsu, S., and Shinkai, A. (2010). Transcription profile of Thermus thermophilus CRISPR systems after phage infection. J. Mol. Biol. *395*, 270–281.

Al-Aamri, A., Taha, K., Al-Hammadi, Y., Maalouf, M., and Homouz, Di. (2019). Analyzing a co-occurrence gene-interaction network to identify disease-gene association. BMC Bioinformatics *20*, 1–15.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Alvarez-Ponce, D., Lopez, P., Bapteste, E., and McInerney, J.O. (2013). Gene similarity networks provide tools for understanding eukaryote origins and evolution. Proc. Natl. Acad. Sci. *110*, E1594–E1603.

AM, M., Hyland, E.M., Cormican, P., Moran, R.J., Webb, A.E., Lee, K.D., Hernandez, J., Prado-Martinez, J., Creevey, C.J., and Aspden, J.L. (2019). Gene Fusions derived by transcriptional readthrough are Driven by Segmental Duplication in Human. Genome Biol. Evol.

Amitai, G., and Sorek, R. (2016). CRISPR-Cas adaptation: Insights into the mechanism of action. Nat. Rev. Microbiol. *14*, 67–76.

Andam, C.P., Fournier, G.P., and Gogarten, J.P. (2011). Multilevel populations and the evolution of antibiotic resistance through horizontal gene transfer. FEMS Microbiol. Rev. *35*, 756–767.

Anderson, R.E., Brazelton, W.J., and Baross, J.A. (2011). Using CRISPRs as ametagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. FEMS Microbiol. Ecol. *77*, 120–133.

Arroyo, A.S., Iannes, R., Bapteste, E., and Ruiz-Trillo, I. (2020). Gene Similarity Networks Unveil a Potential Novel Unicellular Group Closely Related to Animals from the Tara Oceans Expedition. Genome Biol. Evol. *12*, 1664–1678.

Babushok, D. V., Ohshima, K., Ostertag, E.M., Chen, X., Wang, Y., Mandal, P.K., Okada, N., Abrams, C.S., and Kazazian, H.H. (2007). A novel testis ubiquitin-binding protein gene arose by exon shuffling in hominoids. Genome Res. *17*, 1129–1138.

Bah, T. (2007). Inkscape: guide to a vector drawing program (prentice hall press).

Bapteste, E. (2014). The origins of microbial adaptations: how introgressive descent, egalitarian evolutionary transitions and expanded kin selection shape the network of life. Front. Microbiol. *5*, 83.

Bapteste, E., Lopez, P., Bouchard, F., Baquero, F., McInerney, J.O., and Burian, R.M. (2012). Evolutionary analyses of non-genealogical bonds produced by introgressive descent. Proc. Natl. Acad. Sci. *109*, 18266–18272.

Bapteste, E., van Iersel, L., Janke, A., Kelchner, S., Kelk, S., McInerney, J.O., Morrison, D.A., Nakhleh, L., Steel, M., Stougie, L., et al. (2013). Networks: Expanding evolutionary thinking. Trends Genet. *29*, 439–441.

Barbour, A.G., Dai, Q., Restrepo, B.I., Stoenner, H.G., and Frank, S.A. (2006). Pathogen escape from host immunity by a genome program for antigenic variation. Proc. Natl. Acad. Sci. *103*, 18290–18295.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science (80-.). *315*, 1709–1712.

Basgall, E.M., Goetting, S.C., Goeckel, M.E., Giersch, R.M., Roggenkamp, E., Schrock, M.N., Halloran, M., and Finnigan, G.C. (2018). Gene drive inhibition by the anti-CRISPR proteins AcrIIA2 and AcrIIA4 in Saccharomyces cerevisiae. Microbiology *164*, 464.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In Third International AAAI Conference on Weblogs and Social Media, p.

Behler, J., Sharma, K., Reimann, V., Wilde, A., Urlaub, H., and Hess, W.R. (2018). The host-encoded RNase e endonuclease as the crRNA maturation enzyme in a CRISPR-Cas subtype III-Bv system. Nat. Microbiol. *3*, 367–377.

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

Bernheim, A., Calvo-Villamañán, A., Basier, C., Cui, L., Rocha, E.P.C., Touchon, M., and Bikard, D. (2017). Inhibition of NHEJ repair by type II-A CRISPR-Cas systems in bacteria. Nat. Commun. *8*, 1–9.

Bernheim, A., Bikard, D., Touchon, M., and Rocha, E.P.C. (2019). A matter of background : DNA repair pathways as a possible cause for the sparse distribution of CRISPR-Cas systems in bacteria.

Bhoobalan-Chitty, Y., Johansen, T.B., Di Cianni, N., and Peng, X. (2019). Inhibition of Type III CRISPR-Cas Immunity by an Archaeal Virus-Encoded Anti-CRISPR Protein. Cell *179*, 448-458.e11.

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

Biswas, A., Fineran, P.C., and Brown, C.M. (2014). Accurate computational prediction of the transcribed strand of CRISPR non-coding RNAs. Bioinformatics

30, 1805–1813.

Biswas, A., Staals, R.H.J., Morales, S.E., Fineran, P.C., and Brown, C.M. (2016). CRISPRDetect: A flexible algorithm to define CRISPR arrays. BMC Genomics *17*, 1–14.

Bland, J.M., and Altman, D.G. (1995). Multiple significance tests: the Bonferroni method. Bmj *310*, 170.

Bland, C., Ramsey, T.L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N.C., and Hugenholtz, P. (2007). CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics *8*, 1–8.

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

Bondy-Denomy, J., Garcia, B., Strum, S., Du, M., Rollins, M.F., Hidalgo-Reyes, Y., Wiedenheft, B., Maxwell, K.L., and Davidson, A.R. (2015). Multiple mechanisms for CRISPR-Cas inhibition by anti-CRISPR proteins. Nature.

Borges, A.L., Davidson, A.R., and Bondy-Denomy, J. (2017). The Discovery, Mechanisms, and Evolutionary Impact of Anti-CRISPRs. Annu. Rev. Virol. *4*, 37–59.

Brennan, G., Kozyrev, Y., and Hu, S.-L. (2008). TRIMCyp expression in Old World primates Macaca nemestrina and Macaca fascicularis. Proc. Natl. Acad. Sci. *105*, 3569–3574.

Broughton, J.P., Deng, X., Yu, G., Fasching, C.L., Servellita, V., Singh, J., Miao, X., Streithorst, J.A., Granados, A., Sotomayor-Gonzalez, A., et al. (2020). CRISPR–Cas12-based detection of SARS-CoV-2. Nat. Biotechnol. *38*, 870–874.

Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L., Dickman, M.J., Makarova, K.S., Koonin, E. V, and Van Der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. Science (80-.). *321*, 960–964.

Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59.

Burstein, D., Sun, C.L., Brown, C.T., Sharon, I., Anantharaman, K., Probst, A.J., Thomas, B.C., and Banfield, J.F. (2016). Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nat. Commun. *7*, 1–8.

Chaconas, G., and Kobryn, K. (2010). Structure, function, and evolution of linear replicons in Borrelia. Annu. Rev. Microbiol. *64*, 185–202.

Chakraborty, S., Snijders, A.P., Chakravorty, R., Ahmed, M., Tarek, A.M., and Hossain, M.A. (2010). Comparative network clustering of direct repeats (DRs) and cas genes confirms the possibility of the horizontal transfer of CRISPR locus among bacteria. Mol. Phylogenet. Evol. *56*, 878–887.

Charlebois, R.L., and Doolittle, W.F. (2004). Computing prokaryotic gene ubiquity: Rescuing the core from extinction. Genome Res. *14*, 2469–2477.

Chau, M. (2011). Visualizing web search results using glyphs: Design and evaluation of a flower metaphor. ACM Trans. Manag. Inf. Syst. *2*, 1–27.

Chen, S.Y., Feng, Z., and Yi, X. (2017). A general introduction to adjustment for multiple comparisons. J. Thorac. Dis. *9*, 1725–1729.

Cheng, S., Karkar, S., Bapteste, E., Yee, N., Falkowski, P., and Bhattacharya, D. (2014). Sequence similarity network reveals the imprints of major diversification events in the evolution of microbial life. Front. Ecol. Evol. *2*, 1–13.

Chevallereau, A., Meaden, S., Houte, S. Van, Westra, E.R., Rollie, C., and Rollie, C. (2020). The effect of bacterial mutation rate on the evolution of CRISPR-Cas adaptive immunity.

Chothia, C., Gough, J., Vogel, C., and Teichmann, S.A. (2003). Evolution of the protein repertoire. Science (80-.). *300*, 1701–1703.

Ciccarelli, F.D., Doerks, T., Mering, C. von, Creevey, C.J., Snel, B., and Bork, P. (2006). Toward Automatic Reconstruction of a Highly Resolved Tree of Life. Science (80-.). *311*, 1283–1288.

Coleman, O., Hogan, R., McGoldrick, N., Rudden, N., and McInerney, J. (2015). Evolution by Pervasive Gene Fusion in Antibiotic Resistance and Antibiotic Synthesizing Genes. Computation *3*, 114–127.

Common, J., Morley, D., Westra, E.R., and Houte, S. Van (2019). CRISPR-Cas immunity leads to a coevolutionary arms race between Streptococcus thermophilus and lytic phage.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., and Marraffini, L.A. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science (80-.). *339*, 819–823.

Consortium, G.O. (2004). The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. *32*, D258–D261.

Corel, E., Lopez, P., Méheust, R., and Bapteste, E. (2016). Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. Trends Microbiol. *24*, 224–237.

Corel, E., Méheust, R., Watson, A.K., Mcinerney, J.O., Lopez, P., and Bapteste, E. (2018). Bipartite network analysis of gene sharings in the microbial world. Mol. Biol. Evol. *35*, 899–913.

Couvin, D., Bernheim, A., Toffano-Nioche, C., Touchon, M., Michalik, J., Néron, B., Rocha, E.P.C., Vergnaud, G., Gautheret, D., and Pourcel, C. (2018). CRISPRCasFinder, an update of CRISRFinder, includes a portable version, enhanced performance and integrates search for Cas proteins. Nucleic Acids Res. *46*, W246–W251.

Creevey, C.J., Doerks, T., Fitzpatrick, D.A., Raes, J., and Bork, P. (2011). Universally distributed single-copy genes indicate a constant rate of horizontal transfer. PLoS One *6*, e22099.

Darwin, C. (1859). The origin of species. 6th (John Murray, London).

Deichelbohrer, I., Alonso, J.C., Lüder, G., and Trautner, T.A. (1985). Plasmid transduction by Bacillus subtilis bacteriophage SPP1: effects of DNA homology between plasmid and bacteriophage. J. Bacteriol. *162*, 1238–1243.

Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., Eckert, M.R., Vogel, J., and Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature *471*, 602–607.

DeSalle, R., and Riley, M. (2020). Should Networks Supplant Tree Building? Microorganisms *8*, 1179.

Deveau, H., Barrangou, R., Garneau, J.E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P., and Moineau, S. (2008). Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J. Bacteriol. *190*, 1390–1400.

Dion, M.B., Labrie, S.J., Shah, S.A., and Moineau, S. (2018). CRISPRStudio: A User-Friendly Software for Rapid CRISPR Array Visualization. Viruses *10*, 1–11.

Dong, D., Guo, M., Wang, S., Zhu, Y., Wang, S., Xiong, Z., Yang, J., Xu, Z., and Huang, Z. (2017). Structural basis of CRISPR–SpyCas9 inhibition by an anti-CRISPR protein. Nature *546*, 436–439.

Van Dongen, S.M. (2000). Graph clustering by flow simulation.

Dupuis, M.-È., Villion, M., Magadán, A.H., and Moineau, S. (2013a). CRISPR-Cas and restriction–modification systems are compatible and increase phage resistance. Nat. Commun. *4*, 1–7.

Dupuis, M.È., Villion, M., Magadán, A.H., and Moineau, S. (2013b). CRISPR-Cas and restriction-modification systems are compatible and increase phage resistance. Nat. Commun. *4*, 1–7.

East-Seletsky, A., O'Connell, M.R., Burstein, D., Knott, G.J., and Doudna, J.A. (2017). RNA targeting by functionally orthogonal type VI-A CRISPR-Cas enzymes. Mol. Cell *66*, 373–383.

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

Edgar, R.C. (2007). PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics *8*, 1–6.

Elmore, J.R., Yokooji, Y., Sato, T., Olson, S., Glover Claiborne VC, I.I.I., Graveley, B.R., Atomi, H., Terns, R.M., and Terns, M.P. (2013). Programmable plasmid interference by the CRISPR-Cas system in Thermococcus kodakarensis. RNA Biol. *10*, 828–840.

Elmore, J.R., Sheppard, N.F., Ramia, N., Deighan, T., Li, H., Terns, R.M., and Terns, M.P. (2016). Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR–Cas system. Genes Dev. *30*, 447–459.

Eme, L., Gentekaki, E., Curtis, B., Archibald, J.M., and Roger, A.J. (2017). Lateral Gene Transfer in the Adaptation of the Anaerobic Parasite Blastocystis to the Gut. Curr. Biol. *27*, 807–820.

Enright, A.J., Iliopoulos, I., Kyrpides, N.C., and Ouzounis, C.A. (1999). Protein interaction maps for complete genomes based on gene fusion events. Nature *402*, 86.

Van Etten, J., and Bhattacharya, D. (2020). Horizontal Gene Transfer in Eukaryotes: Not if, but How Much? Trends Genet. 1–11.

Fagerlund, R.D., Wilkinson, M.E., Klykov, O., Barendregt, A., Pearce, F.G., Kieper, S.N., Maxwell, H.W.R., Capolupo, A., Heck, A.J.R., and Krause, K.L. (2017). Spacer capture and integration by a type IF Cas1–Cas2-3 CRISPR adaptation complex. Proc. Natl. Acad. Sci. *114*, E5122–E5128.

Faure, G., Shmakov, S.A., Yan, W.X., Cheng, D.R., Scott, D.A., Peters, J.E., Makarova, K.S., and Koonin, E. V. (2019). CRISPR–Cas in mobile genetic elements: counter-defence and beyond. Nat. Rev. Microbiol. *17*.

Fineran, P.C., and Charpentier, E. (2012). Memory of viral infections by CRISPR-Cas adaptive immune systems: acquisition of new information. Virology *434*, 202–209.

Fineran, P.C., Gerritzen, M.J.H., Suárez-Diez, M., Künne, T., Boekhorst, J., van Hijum, S.A.F.T., Staals, R.H.J., and Brouns, S.J.J. (2014). Degenerate target sites mediate rapid primed CRISPR adaptation. Proc. Natl. Acad. Sci. *111*, E1629–E1638.

Fondi, M., and Fani, R. (2010). The horizontal flow of the plasmid resistome: Clues from inter-generic similarity networks. Environ. Microbiol. *12*, 3228–3242.

Forster, P., Forster, L., Renfrew, C., and Forster, M. (2020). Phylogenetic network analysis of SARS-CoV-2 genomes. Proc. Natl. Acad. Sci. U. S. A. *117*, 9241–9243.

Freeman, T.C., Horsewell, S., Patir, A., Harling-Lee, J., Regan, T., Shih, B.B., Prendergast, J., Hume, D.A., and Angus, T. (2020). Graphia: A platform for the graph-based visualisation and analysis of complex data. BioRxiv 2020.09.02.279349.

Fruchterman, T.M.J., and Reingold, E.M. (1991). Graph drawing by force-directed placement. Softw. Pract. Exp. *21*, 1129–1164.

Garrett, R.A., Vestergaard, G., and Shah, S.A. (2011). Archaeal CRISPR-based immune systems: Exchangeable functional modules. Trends Microbiol. *19*, 549–556.

Gemski, P., Lazere, J.R., Casey, T., and Wohlhieter, J.A. (1980). Presence of a virulence-associated plasmid in Yersinia pseudotuberculosis. Infect. Immun. *28*, 1044–1047.

Giribet, G. (2005). TNT: tree analysis using new technology.

Gleditzsch, D., Pausch, P., Müller-Esparza, H., Özcan, A., Guo, X., Bange, G., and Randau, L. (2019). PAM identification by CRISPR-Cas effector complexes: diversified mechanisms and structures. RNA Biol. *16*, 504–517.

Godde, J.S., and Bickerton, A. (2006). The repetitive DNA elements called CRISPRs and their associated genes: Evidence of horizontal transfer among prokaryotes. J. Mol. Evol. *62*, 718–729.

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

Goren, M.G., Doron, S., Globus, R., Amitai, G., Sorek, R., and Qimron, U. (2016). Repeat Size Determination by Two Molecular Rulers in the Type I-E CRISPR Array. Cell Rep. *16*, 2811–2818.

Graham, L.A., Lougheed, S.C., Ewart, K.V., and Davies, P.L. (2008). Lateral transfer of a lectin-like antifreeze protein gene in fishes. PLoS One *3*, e2616.

Griffith, B.Y.F. (1928). The Significance of pneumococcal types. Occurrence of a Variety of Serological Types in the Sputum from an individual case of pneumonia. J. Hyg. (Lond). *27*, 113–159.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics *8*, 1–10.

Grüschow, S., Athukoralage, J.S., Graham, S., Hoogeboom, T., and White, M.F. (2019). Cyclic oligoadenylate signalling mediates Mycobacterium tuberculosis CRISPR defence. Nucleic Acids Res. *47*, 9259–9270.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics *32*, 2847–2849.

Gudbjartsson, D.F., Helgason, A., Jonsson, H., Magnusson, O.T., Melsted, P., Norddahl, G.L., Saemundsdottir, J., Sigurdsson, A., Sulem, P., Agustsdottir, A.B., et al. (2020). Spread of SARS-CoV-2 in the Icelandic Population. N. Engl. J. Med. *382*, 2302–2315.

Guo, T.W., Bartesaghi, A., Yang, H., Falconieri, V., Rao, P., Merk, A., Eng, E.T., Raczkowski, A.M., Fox, T., Earl, L.A., et al. (2017). Cryo-EM Structures Reveal Mechanism and Inhibition of DNA Targeting by a CRISPR-Cas Surveillance Complex. Cell *171*, 414-426.e12.

Hacker, J., Blum-Oehler, G., Mühldorfer, I., and Tschäpe, H. (1997). Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol. Microbiol. *23*, 1089–1097.

Haerter, J.O., Trusina, A., and Sneppen, K. (2011). Targeted Bacterial Immunity Buffers Phage Diversity. J. Virol. *85*, 10554–10560.

Haft, D.H., Selengut, J., Mongodin, E.F., and Nelson, K.E. (2005). A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput Biol *1*, e60.

Haggerty, L.S., Jachiet, P.A., Hanage, W.P., Fitzpatrick, D.A., Lopez, P., O'Connell, M.J., Pisani, D., Wilkinson, M., Bapteste, E., and McInerney, J.O. (2014). A pluralistic account of homology: Adapting the models to the data. Mol. Biol. Evol. *31*, 501–516.

Halary, S., McInerney, J.O., Lopez, P., and Bapteste, E. (2013). EGN: a wizard for construction of gene and genome similarity networks. BMC Evol. Biol. *13*, 1–9.

Hampton, H.G., Patterson, A.G., Chang, J.T., Taylor, C., and Fineran, P.C. (2019). GalK limits type IF CRISPR-Cas expression in a CRP-dependent manner. FEMS Microbiol. Lett. *366*, fnz137.

Hatoum-Aslan, A., Maniv, I., Samai, P., and Marraffini, L.A. (2014). Genetic characterization of antiplasmid immunity through a type III-A CRISPR-cas system. J. Bacteriol. *196*, 310–317.

Haurwitz, R.E., Jinek, M., Wiedenheft, B., Zhou, K., and Doudna, J.A. (2010). Sequence-and structure-specific RNA processing by a CRISPR endonuclease. Science (80-.). *329*, 1355–1358.

Hayes, R.P., Xiao, Y., Ding, F., Van Erp, P.B.G., Rajashankar, K., Bailey, S., Wiedenheft, B., and Ke, A. (2016). Structural basis for promiscuous PAM recognition in type I–E Cascade from E. coli. Nature *530*, 499–503.

Heidelberg, J.F., Nelson, W.C., Schoenfeld, T., and Bhaya, D. (2009). Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes. PLoS One *4*, e4169.

Heler, R., Samai, P., Modell, J.W., Weiner, C., Goldberg, G.W., Bikard, D., and Marraffini, L.A. (2015). Cas9 specifies functional viral targets during CRISPR–Cas adaptation. Nature *519*, 199–202.

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

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

Horn, H., Slaby, B.M., Jahn, M.T., Bayer, K., Moitinho-Silva, L., Förster, F., Abdelmohsen, U.R., and Hentschel, U. (2016). An Enrichment of CRISPR and other defense-related features in marine sponge-associated microbial metagenomes. Front. Microbiol. *7*, 1–15.

Horvath, P., and Barrangou, R. (2010). CRISPR/Cas, the immune system of Bacteria and Archaea. Science (80-.). *327*, 167–170.

Hotopp, J.C.D., Clark, M.E., Oliveira, D.C.S.G., Foster, J.M., Fischer, P., Torres, M.C.M., Giebel, J.D., Kumar, N., Ishmael, N., and Wang, S. (2007). Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science (80-.). *317*, 1753–1756.

Van Houte, S., Ekroth, A.K.E., Broniewski, J.M., Chabas, H., Ashby, B., Bondy-Denomy, J., Gandon, S., Boots, M., Paterson, S., Buckling, A., et al. (2016). The diversity-generating benefits of a prokaryotic adaptive immune system. Nature *532*, 385–388.

Hsu, P.D., Lander, E.S., and Zhang, F. (2014). Development and applications of CRISPR-Cas9 for genome engineering. Cell *157*, 1262–1278.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M.C., Rattei, T., Mende, D.R., Sunagawa, S., and Kuhn, M. (2015). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. *44*, D286–D293.

Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol. Biol. Evol. *33*, 1635–1638.

Hwang, S., and Maxwell, K.L. (2019). Meet the Anti-CRISPRs: Widespread Protein

Inhibitors of CRISPR-Cas Systems. Cris. J. *2*, 23–30.

Hyatt, D., Chen, G.L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*.

Hynes, A.P., Rousseau, G.M., Lemay, M.L., Horvath, P., Romero, D.A., Fremaux, C., and Moineau, S. (2017). An anti-CRISPR from a virulent streptococcal phage inhibits Streptococcus pyogenes Cas9. Nat. Microbiol. *2*, 1374–1380.

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

Ivančić-Baće, I., Cass, S.D., Wearne, S.J., and Bolt, E.L. (2015). Different genome stability proteins underpin primed and naive adaptation in E. coli CRISPR-Cas immunity. Nucleic Acids Res. *43*, 10821–10830.

Jachiet, P.A., Pogorelcnik, R., Berry, A., Lopez, P., and Bapteste, E. (2013). MosaicFinder: Identification of fused gene families in sequence similarity networks. Bioinformatics *29*, 837–844.

Jachiet, P.A., Colson, P., Lopez, P., and Bapteste, E. (2014). Extensive gene remodeling in the viral world: New evidence for nongradual evolution in the mobilome network. Genome Biol. Evol. *6*, 2195–2205.

Jackson, R.N., Golden, S.M., van Erp, P.B.G., Carter, J., Westra, E.R., Brouns, S.J.J., van der Oost, J., Terwilliger, T.C., Read, R.J., and Wiedenheft, B. (2014). Crystal structure of the CRISPR RNA–guided surveillance complex from Escherichia coli. Science (80-.). *345*, 1473–1479.

Jackson, S.A., McKenzie, R.E., Fagerlund, R.D., Kieper, S.N., Fineran, P.C., and Brouns, S.J.J. (2017). CRISPR-Cas: Adapting to change. Science (80-.). *356*.

Jansen, R., Embden, J.D.A. van, Gaastra, W., and Schouls, L.M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. Mol. Microbiol. *43*, 1565–1575.

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

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. Science (80-.). *337*, 816–821.

Jinek, M., Jiang, F., Taylor, D.W., Sternberg, S.H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., and Lin, S. (2014). Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. Science (80-.). *343*.

Jordan, W.C., and Turnquist, M.A. (1983). A stochastic, dynamic network model for railroad car distribution. Transp. Sci. *17*, 123–145.

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. Genome Res. *20*, 1313–1326.

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., and Tanabe, M. (2019). New approach for understanding genome variations in KEGG. Nucleic Acids Res. *47*, D590–D595.

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

Katoh, K., Kuma, K., Toh, H., and Miyata, T. (2005). MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. *33*, 511–518.

Kay, E., Vogel, T.M., Bertolla, F., Nalin, R., and Simonet, P. (2002). In situ transfer of antibiotic resistance genes from transgenic (transplastomic) tobacco plants to bacteria. Appl. Environ. Microbiol. *68*, 3345–3351.

Kazlauskiene, M., Kostiuk, G., Venclovas, Č., Tamulaitis, G., and Siksnys, V. (2017). A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. Science (80-.). *357*, 605–609.

Keeling, P.J., and Palmer, J.D. (2008). Horizontal gene transfer in eukaryotic evolution. Nat. Rev. Genet. *9*, 605–618.

Kieper, S.N., Almendros, C., Behler, J., McKenzie, R.E., Nobrega, F.L., Haagsma, A.C., Vink, J.N.A., Hess, W.R., and Brouns, S.J.J. (2018). Cas4 facilitates PAM-compatible spacer selection during CRISPR adaptation. Cell Rep. *22*, 3377–3384.

Kieper, S.N., Almendros, C., and Brouns, S.J.J. (2019). Conserved motifs in the CRISPR leader sequence control spacer acquisition levels in Type ID CRISPR-Cas systems. FEMS Microbiol. Lett. *366*, fnz129.

Kim, J., and Yi, G.-S. (2012). PKMiner: a database for exploring type II polyketide synthases. BMC Microbiol. *12*, 169.

Kim, P.J., and Price, N.D. (2011). Genetic co-occurrence network across sequenced microbes. PLoS Comput. Biol. *7*.

Klovdahl, A.S. (1985). Social networks and the spread of infectious diseases: The AIDS example. Soc. Sci. Med. *21*, 1203–1216.

Koonin, E. V., and Krupovic, M. (2015). Evolution of adaptive immunity from transposable elements combined with innate immune systems. Nat. Rev. Genet. *16*, 184–192.

Koonin, E. V., and Makarova, K.S. (2018). Discovery of Oligonucleotide Signaling Mediated by CRISPR-Associated Polymerases Solves Two Puzzles but Leaves an Enigma. ACS Chem. Biol. *13*, 309–312.

Koonin, E. V., and Zhang, F. (2017). Coupling immunity and programmed cell suicide in prokaryotes: Life-or-death choices. BioEssays *39*, 1–9.

Koonin, E. V, and Makarova, K.S. (2013). CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. RNA Biol. *10*, 679–686.

Koonin, E. V, and Makarova, K.S. (2019). Origins and evolution of CRISPR-Cas systems.

Koonin, E. V., Makarova, K.S., and Aravind, L. (2001). Horizontal gene transfer in prokaryotes: Quantification and classification. Annu. Rev. Microbiol. *55*, 709–742.

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

Koonin, E. V, Makarova, K., Grishin, N. V, and Wolf, Y.I. (2006). A putative RNA-interference-based immune system in prokaryotes: the epitome of prokaryotic genomic diversity. In SYMPOSIA-SOCIETY FOR GENERAL MICROBIOLOGY, (Cambridge; Cambridge University Press; 1999), p. 39.

Krupovic, M., Makarova, K.S., Forterre, P., Prangishvili, D., and Koonin, E. V. (2014). Casposons: A new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. BMC Biol. *12*, 1–12.

Krupovic, M., Shmakov, S., Makarova, K.S., Forterre, P., and Koonin, E. V (2016). Recent mobility of casposons, self-synthesizing transposons at the origin of the CRISPR-Cas immunity. Genome Biol. Evol. *8*, 375–386.

Kunin, V., Sorek, R., and Hugenholtz, P. (2007). Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biol. *8*.

Künne, T., Kieper, S.N., Bannenberg, J.W., Vogel, A.I.M., Miellet, W.R., Klein, M., Depken, M., Suarez-Diez, M., and Brouns, S.J.J. (2016). Cas3-derived target DNA degradation fragments fuel primed CRISPR adaptation. Mol. Cell *63*, 852–864.

Kupczok, A., Landan, G., and Dagan, T. (2015). The Contribution of Genetic Recombination to CRISPR Array Evolution. Genome Biol. Evol. *7*, 1925–1939.

Lam, T.J., and Ye, Y. (2019). Long reads reveal the diversification and dynamics of CRISPR reservoir in microbiomes. BMC Genomics *20*, 1–12.

Latysheva, N.S., and Babu, M.M. (2016). Discovering and understanding oncogenic gene fusions through data intensive computational approaches. Nucleic Acids Res. *44*, 4487–4503.

Leonard, G., and Richards, T.A. (2012). Genome-scale comparative analysis of gene fusions, gene fissions, and the fungal tree of life. Proc. Natl. Acad. Sci. *109*, 21402–21407.

Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics *23*, 127–128.

Levy, A., Goren, M.G., Yosef, I., Auster, O., Manor, M., Amitai, G., Edgar, R., Qimron, U., and Sorek, R. (2015). CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature *520*, 505–510.

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

Linder, C.R., Moret, B.M.E., Nakhleh, L., and Warnow, T. (2004). Network (reticulate) evolution: biology, models, and algorithms. In The Ninth Pacific Symposium on Biocomputing (PSB), p.

Liu, L., Li, X., Wang, J., Wang, M., Chen, P., Yin, M., Li, J., Sheng, G., and Wang, Y. (2017). Two distant catalytic sites are responsible for C2c2 RNase activities. Cell

168, 121–134.

Liu, S.-L., Zhuang, Y., Zhang, P., and Adams, K.L. (2009). Comparative analysis of structural diversity and sequence evolution in plant mitochondrial genes transferred to the nucleus. Mol. Biol. Evol. *26*, 875–891.

Long, M. (2000). A new function evolved from gene fusion. Genome Res. *10*, 1655–1657.

Long, M., Betrán, E., Thornton, K., and Wang, W. (2003). The origin of new genes: glimpses from the young and old. Nat. Rev. Genet. *4*, 865–875.

Long, M., Vankuren, N.W., Chen, S., and Vibranovski, M.D. (2013). New gene evolution: Little did we know. Annu. Rev. Genet. *47*, 307–333.

Lopez-Sanchez, M., Sauvage, E., Da Cunha, V., Clermont, D., Ratsima Hariniaina, E., Gonzalez-Zorn, B., Poyart, C., Rosinski-Chupin, I., and Glaser, P. (2012). The highly dynamic CRISPR1 system of Streptococcus agalactiae controls the diversity of its mobilome. Mol. Microbiol. *85*, 1057–1071.

Majewski, J., and Cohan, F.M. (1999). DNA sequence similarity requirements for interspecific recombination in Bacillus. Genetics *153*, 1525–1533.

Makarova, K.S., Haft, D.H., Barrangou, R., Brouns, S.J.J., Charpentier, E., Horvath, P., Moineau, S., Mojica, F.J.M., Wolf, Y.I., Yakunin, A.F., et al. (2011). Evolution and classification of the CRISPR-Cas systems. Nat. Rev. Microbiol. *9*, 467–477.

Makarova, K.S., Anantharaman, V., Aravind, L., and Koonin, E. V. (2012). Live virus-free or die: coupling of antivirus immunity and programmed suicide or dormancy in prokaryotes. Biol. Direct *7*, 1–10.

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

Makarova, K.S., Anantharaman, V., Grishin, N. V., Koonin, E. V., and Aravind, L. (2014). CARF and WYL domains: Ligand-binding regulators of prokaryotic defense systems. Front. Genet. *5*, 1–9.

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J.J., Charpentier, E., Haft, D.H., et al. (2015). An updated evolutionary classification of CRISPR-Cas systems. Nat. Rev. Microbiol. *13*, 722–736.

Makarova, K.S., Wolf, Y.I., Iranzo, J., Shmakov, S.A., Alkhnbashi, O.S., Brouns, S.J.J., Charpentier, E., Cheng, D., Haft, D.H., and Horvath, P. (2019). Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. Nat. Rev. Microbiol. 1–17.

Makarova, K.S., Timinskas, A., Wolf, Y.I., Gussow, A.B., Siksnys, V., Venclovas, Č., and Koonin, E. V. (2020). Evolutionary and functional classification of the CARF domain superfamily, key sensors in prokaryotic antivirus defense. Nucleic Acids Res. *48*, 8828–8847.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013a). RNA-guided human genome engineering via Cas9. Science (80-.). *339*, 823–826.

Mali, P., Esvelt, K.M., and Church, G.M. (2013b). Cas9 as a versatile tool for engineering biology. Nat. Methods *10*, 957–963.

Maniv, I., Jiang, W., Bikard, D., and Marraffini, L.A. (2016). Impact of different target sequences on type III CRISPR-Cas immunity. J. Bacteriol. *198*, 941–950.

Marraffini, L.A. (2015). CRISPR-Cas immunity in prokaryotes. Nature *526*, 55–61.

Marraffini, L.A., and Sontheimer, E.J. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. Science (80-.). *322*, 1843–1845.

Marraffini, L.A., and Sontheimer, E.J. (2010). CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat. Rev. Genet. *11*, 181–190.

McGinn, J., and Marraffini, L.A. (2016). CRISPR-Cas systems optimize their immune response by specifying the site of spacer integration. Mol. Cell *64*, 616–623.

McGinn, J., and Marraffini, L.A. (2018). Molecular mechanisms of CRISPR–Cas spacer acquisition. Nat. Rev. Microbiol. *17*, 7–12.

McInerney, J.O., Pisani, D., Bapteste, E., and O'Connell, M.J. (2011). The public goods hypothesis for the evolution of life on Earth. Biol. Direct *6*, 41.

McInerney, J.O., O'Connell, M.J., and Pisani, D. (2014). The hybrid nature of the Eukaryota and a consilient view of life on Earth. Nat. Rev. Microbiol. *12*, 449–455.

McMahon, S.A., Zhu, W., Graham, S., Rambo, R., White, M.F., and Gloster, T.M. (2020). Structure and mechanism of a Type III CRISPR defence DNA nuclease activated by cyclic oligoadenylate. Nat. Commun. *11*.

Méheust, R., Zelzion, E., Bhattacharya, D., Lopez, P., and Bapteste, E. (2016). Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. Proc. Natl. Acad. Sci. *113*, 3579–3584.

Méheust, R., Watson, A.K., Lapointe, F.J., Papke, R.T., Lopez, P., and Bapteste, E. (2018). Hundreds of novel composite genes and chimeric genes with bacterial origins contributed to haloarchaeal evolution. Genome Biol. *19*, 1–12.

Mick, E., Stern, A., and Sorek, R. (2013). Holding a grudge. RNA Biol. *10*, 900–906.

Mir, A., Edraki, A., Lee, J., and Sontheimer, E.J. (2018). Type II-C CRISPR-Cas9 biology, mechanism, and application. ACS Chem. Biol. *13*, 357–365.

Modell, J.W., Jiang, W., and Marraffini, L.A. (2017). CRISPR–Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. Nature *544*, 101–104.

Mohanraju, P., Makarova, K.S., Zetsche, B., Zhang, F., Koonin, E. V., and Van Der Oost, J. (2016). Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science (80-.). *353*.

Mojica, F.J.M., Ferrer, C., Juez, G., and Rodriguez-Valera, F. (1995). Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning. Mol. Microbiol. *17*, 85–93.

Mojica, F.J.M., García-Martínez, J., and Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J. Mol. Evol. *60*, 174–182.

Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J., and Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology *155*, 733–740.

Mulepati, S., Héroux, A., and Bailey, S. (2014). Crystal structure of a CRISPR RNA–guided surveillance complex bound to a ssDNA target. Science (80-.). *345*, 1479–1484.

Nam, K.H., Huang, Q., and Ke, A. (2012). Nucleic acid binding surface and dimer interface revealed by CRISPR-associated CasB protein structures. FEBS Lett. *586*, 3956–3961.

Naser, I. Bin, Hoque, M.M., Nahid, M.A., Tareq, T.M., Rocky, M.K., and Faruque, S.M. (2017). Analysis of the CRISPR-Cas system in bacteriophages active on epidemic strains of Vibrio cholerae in Bangladesh. Sci. Rep. *7*, 1–10.

Nelson-Sathi, S., Dagan, T., Landan, G., Janssen, A., Steel, M., McInerney, J.O., Deppenmeier, U., and Martin, W.F. (2012). Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. Proc. Natl. Acad. Sci. *109*, 20537–20542.

Nelson-Sathi, S., Sousa, F.L., Roettger, M., Lozada-Chávez, N., Thiergart, T., Janssen, A., Bryant, D., Landan, G., Schönheit, P., Siebers, B., et al. (2015). Origins of major archaeal clades correspond to gene acquisitions from bacteria. Nature *517*, 77–80.

Newire, E., Aydin, A., Juma, S., Enne, V.I., and Roberts, A.P. (2020). Identification of a Type IV-A CRISPR-Cas System Located Exclusively on IncHI1B/IncFIB Plasmids in Enterobacteriaceae. Front. Microbiol. *11*, 1–11.

Nguyen, L.-T., Schmidt, H.A., Von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268–274.

Nicholson, T.J., Jackson, S.A., Croft, B.I., Staals, R.H.J., Fineran, P.C., and Brown, C.M. (2019). Bioinformatic evidence of widespread priming in type I and II CRISPR-Cas systems. RNA Biol. *16*, 566–576.

Nickel, L., Weidenbach, K., Jäger, D., Backofen, R., Lange, S.J., Heidrich, N., and Schmitz, R.A. (2013). Two CRISPR-Cas systems in Methanosarcina mazei strain Gö1 display common processing features despite belonging to different types I and III. RNA Biol. *10*, 779–791.

Niewoehner, O., Garcia-Doval, C., Rostøl, J.T., Berk, C., Schwede, F., Bigler, L., Hall, J., Marraffini, L.A., and Jinek, M. (2017). Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers. Nature *548*, 543–548.

Nozawa, T., Furukawa, N., Aikawa, C., Watanabe, T., Haobam, B., Kurokawa, K., Maruyama, F., and Nakagawa, I. (2011). CRISPR inhibition of prophage acquisition in streptococcus pyogenes. PLoS One *6*.

Nuñez, J.K., Lee, A.S.Y., Engelman, A., and Doudna, J.A. (2015). Integrasemediated spacer acquisition during CRISPR-Cas adaptive immunity. Nature *519*, 193–198.

Nuñez, J.K., Bai, L., Harrington, L.B., Hinder, T.L., and Doudna, J.A. (2016). CRISPR immunological memory requires a host factor for specificity. Mol. Cell *62*, 824–833.

Nussenzweig, P.M., McGinn, J., and Marraffini, L.A. (2019). Cas9 Cleavage of Viral Genomes Primes the Acquisition of New Immunological Memories. Cell Host Microbe *26*, 515-526.e6.

O'HARA, R.J. (1997). Population thinking and tree thinking in systematics. Zool. Scr. *26*, 323–329.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733–D745.

Oakley, T.H. (2017). Furcation and fusion: The phylogenetics of evolutionary novelty. Dev. Biol. *431*, 69–76.

Ocaña-Pallarès, E., Najle, S.R., Scazzocchio, C., and Ruiz-Trillo, I. (2019). Reticulate evolution in eukaryotes: Origin and evolution of the nitrate assimilation pathway. PLoS Genet. *15*, e1007986.

Ochman, H., Lawrence, J.G., and Groisman, E.A. (2000). Lateral gene transfer and the nature of bacterial innovation. Nature *405*, 299–304.

Oliveira, P.H., Touchon, M., and Rocha, E.P.C. (2014). The interplay of restriction-modification systems with mobile genetic elements and their prokaryotic hosts. Nucleic Acids Res. *42*, 10618–10631.

Van Der Oost, J., Westra, E.R., Jackson, R.N., and Wiedenheft, B. (2014). Unravelling the structural and mechanistic basis of CRISPR-Cas systems. Nat. Rev. Microbiol. *12*, 479–492.

Osawa, T., Inanaga, H., Sato, C., and Numata, T. (2015). Crystal structure of the CRISPR-Cas RNA silencing Cmr complex bound to a target analog. Mol. Cell *58*, 418–430.

Özcan, A., Pausch, P., Linden, A., Wulf, A., Schühle, K., Heider, J., Urlaub, H., Heimerl, T., Bange, G., and Randau, L. (2019). Type IV CRISPR RNA processing and effector complex formation in Aromatoleum aromaticum. Nat. Microbiol. *4*, 89–96.

Padilha, V.A., Alkhnbashi, O.S., Shah, S.A., de Carvalho, A.C.P.L.F., and Backofen, R. (2020). CRISPRcasIdentifier: Machine learning for accurate identification and classification of CRISPR-Cas systems. Gigascience *9*, 1–12.

Paez-Espino, D., Sharon, I., Morovic, W., Stahl, B., Thomas, B.C., Barrangou, R., and Banfielda, J.F. (2015). CRISPR immunity drives rapid phage genome evolution in streptococcus thermophilus. MBio *6*, 1–9.

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

Pasek, S., Risler, J.-L., and Brézellec, P. (2006). Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. Bioinformatics *22*, 1418–1423.

Pathmanathan, J.S., Lopez, P., Lapointe, F.-J., and Bapteste, E. (2017). CompositeSearch: a generalized network approach for composite gene families detection. Mol. Biol. Evol. *35*, 252–255.

Pawluk, A., Davidson, A.R., and Maxwell, K.L. (2018). Anti-CRISPR: Discovery, mechanism and function. Nat. Rev. Microbiol. *16*, 12–17.

Peng, W., Feng, M., Feng, X., Liang, Y.X., and She, Q. (2015). An archaeal CRISPR type III-B system exhibiting distinctive RNA targeting features and mediating dual RNA and DNA interference. Nucleic Acids Res. *43*, 406–417.

Peng, X., Mayo-Muñoz, D., Bhoobalan-Chitty, Y., and Martínez-Álvarez, L. (2020). Anti-CRISPR Proteins in Archaea. Trends Microbiol. *28*, 913–921.

Peters, J.E., Makarova, K.S., Shmakov, S., and Koonin, E. V. (2017). Recruitment of CRISPR-Cas systems by Tn7-like transposons. Proc. Natl. Acad. Sci. U. S. A. *114*, E7358–E7366.

Plagens, A., Tripp, V., Daume, M., Sharma, K., Klingl, A., Hrle, A., Conti, E., Urlaub, H., and Randau, L. (2014). In vitro assembly and activity of an archaeal CRISPR-Cas type IA Cascade interference complex. Nucleic Acids Res. *42*, 5125–5138.

Polz, M.F., Alm, E.J., and Hanage, W.P. (2013). Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends Genet. *29*, 170–175.

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

Pourcel, C., Touchon, M., Villeriot, N., Vernadet, J.P., Couvin, D., Toffano-Nioche, C., and Vergnaud, G. (2020). CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. Nucleic Acids Res. *48*, D535–D544.

Pradhan, M.P., Nagulapalli, K., and Palakal, M.J. (2012). Cliques for the identification of gene signatures for colorectal cancer across population. BMC Syst. Biol. *6*, S17.

Proulx, S.R., Promislow, D.E.L., and Phillips, P.C. (2005). Network thinking in ecology and evolution. Trends Ecol. Evol. *20*, 345–353.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2006). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. *35*, D61–D65.

Przybilski, R., Richter, C., Gristwood, T., Clulow, J.S., Vercoe, R.B., and Fineran, P.C. (2011). Csy4 is responsible for CRISPR RNA processing in Pectobacterium atrosepticum. RNA Biol. *8*, 517–528.

Pyenson, N.C., Gayvert, K., Varble, A., Elemento, O., and Marraffini, L.A. (2017). Broad Targeting Specificity during Bacterial Type III CRISPR-Cas Immunity Constrains Viral Escape. Cell Host Microbe *22*, 343-353.e3.

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

Rath, D., Amlinger, L., Rath, A., and Lundgren, M. (2015). The CRISPR-Cas immune system: Biology, mechanisms and applications. Biochimie *117*, 119–128.

Raymond, J.A., and Morgan-Kiss, R. (2017). Multiple ice-binding proteins of probable prokaryotic origin in an Antarctic lake alga, Chlamydomonas sp. ICE-MDV (Chlorophyceae). J. Phycol. *53*, 848–854.

Redding, S., Sternberg, S.H., Marshall, M., Gibb, B., Bhat, P., Guegler, C.K., Wiedenheft, B., Doudna, J.A., and Greene, E.C. (2015). Surveillance and processing of foreign DNA by the Escherichia coli CRISPR-Cas system. Cell *163*, 854–865.

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

Richter, C., Dy, R.L., McKenzie, R.E., Watson, B.N.J., Taylor, C., Chang, J.T., McNeil, M.B., Staals, R.H.J., and Fineran, P.C. (2014). Priming in the Type IF CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. Nucleic Acids Res. *42*, 8516–8526.

Richter, H., Zoephel, J., Schermuly, J., Maticzka, D., Backofen, R., and Randau, L. (2012b). Characterization of CRISPR RNA processing in Clostridium thermocellum and Methanococcus maripaludis. Nucleic Acids Res. *40*, 9887–9896.

Rivera, M.C., and Lake, J.A. (2004). The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature *431*, 152–155.

Rogers, R.L., Bedford, T., and Hartl, D.L. (2009). Formation and longevity of chimeric and duplicate genes in Drosophila melanogaster. Genetics *181*, 313–322.

Rollie, C., Chevallereau, A., Watson, B.N.J., Chyou, T., Fradet, O., McLeod, I., Fineran, P.C., Brown, C.M., Gandon, S., and Westra, E.R. (2020). Targeting of temperate phages drives loss of type I CRISPR–Cas systems. Nature *578*, 149–153.

Rollins, M.F., Chowdhury, S., Carter, J., Golden, S.M., Wilkinson, R.A., Bondy-Denomy, J., Lander, G.C., and Wiedenheft, B. (2017). Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. Proc. Natl. Acad. Sci. *114*, E5113–E5121.

Rouillon, C., Athukoralage, J.S., Graham, S., Grüschow, S., and White, M.F. (2019). Investigation of the cyclic oligoadenylate signaling pathway of type III CRISPR systems. Methods Enzymol. *616*, 191–218.

Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S.A., and Sørensen, S.J. (2020). CRISPRCasTyper: An automated tool for the identification, annotation and classification of CRISPR-Cas loci. BioRxiv.

Sapranauskas, R., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., and Siksnys, V. (2011). The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. Nucleic Acids Res. *39*, 9275–9282.

Sashital, D.G., Jinek, M., and Doudna, J.A. (2011). An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. Nat. Struct. Mol. Biol. *18*, 680.

Schönknecht, G., Chen, W.H., Ternes, C.M., Barbier, G.G., Shrestha, R.P., Stanke, M., Bräutigam, A., Baker, B.J., Banfield, J.F., Garavito, R.M., et al. (2013). Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. Science (80-.). *339*, 1207–1210.

Sedgwick, P. (2012). Multiple significance tests: the Bonferroni correction. BMJ *344*, e509.

Seed, K.D. (2015). Battling phages: how bacteria defend against viral attack. PLoS Pathog *11*, e1004847.

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

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

Sefcikova, J., Roth, M., Yu, G., and Li, H. (2017). Cas6 processes tight and relaxed repeat RNA via multiple mechanisms: A hypothesis. Bioessays *39*, 1700019.

Segata, N., and Huttenhower, C. (2011). Toward an efficient method of identifying core genes for evolutionary and functional microbial phylogenies. PLoS One *6*.

Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B., van der Oost, J., Brouns, S.J.J., and Severinov, K. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc. Natl. Acad. Sci. *108*, 10098–10103.

Shah, S.A., Erdmann, S., Mojica, F.J.M., and Garrett, R.A. (2013). Protospacer recognition motifs: mixed identities and functional diversity. RNA Biol. *10*, 891–899.

Shah, S.A., Alkhnbashi, O.S., Behler, J., Han, W., She, Q., Hess, W.R., Garrett, R.A., and Backofen, R. (2019). Comprehensive search for accessory proteins encoded with archaeal and bacterial type III CRISPR-cas gene cassettes reveals 39 new cas gene families. RNA Biol. *16*, 530–542.

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

Shao, Y., and Li, H. (2013). Recognition and cleavage of a nonstructured CRISPR RNA by its processing endoribonuclease Cas6. Structure *21*, 385–393.

Shao, S., Zhang, W., Hu, H., Xue, B., Qin, J., Sun, C., Sun, Y., Wei, W., and Sun, Y. (2016). Long-term dual-color tracking of genomic loci by modified sgRNAs of the

CRISPR/Cas9 system. Nucleic Acids Res. *44*, e86–e86.

Shapiro, R.S., Chavez, A., Porter, C.B.M., Hamblin, M., Kaas, C.S., DiCarlo, J.E., Zeng, G., Xu, X., Revtovich, A. V., Kirienko, N. V., et al. (2017). A CRISPR-Cas9-based gene drive platform for genetic interaction analysis in Candida albicans. Nat. Microbiol. *3*, 73–82.

Shehreen, S., Chyou, T., Fineran, P.C., Brown, C.M., and 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.

Shin, J., Jiang, F., Liu, J.-J., Bray, N.L., Rauch, B.J., Baik, S.H., Nogales, E., Bondy-Denomy, J., Corn, J.E., and Doudna, J.A. (2017). Disabling Cas9 by an anti-CRISPR DNA mimic. Sci. Adv. *3*, e1701620.

Shmakov, S., Abudayyeh, O.O., Makarova, K.S., Wolf, Y.I., Gootenberg, J.S., Semenova, E., Minakhin, L., Joung, J., Konermann, S., and Severinov, K. (2015). Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. Mol. Cell *60*, 385–397.

Shmakov, S., Smargon, A., Scott, D., Cox, D., Pyzocha, N., Yan, W., Abudayyeh, O.O., Gootenberg, J.S., Makarova, K.S., Wolf, Y.I., et al. (2017a). Diversity and evolution of class 2 CRISPR-Cas systems. Nat. Rev. Microbiol. *15*, 169–182.

Shmakov, S., Smargon, A., Scott, D., Cox, D., and Pyzocha, N. (2018a). Diversity and evolution of class 2 CRISPR–Cas systems. *15*, 169–182.

Shmakov, S.A., Sitnik, V., Makarova, K.S., Wolf, Y.I., Severinov, K. V, and Koonin, E. V (2017b). The CRISPR spacer space is dominated by sequences from species-specific mobilomes. MBio *8*.

Shmakov, S.A., Koonin, E. V., Severinov, K. V., Wolf, Y.I., and Makarova, K.S. (2018b). Systematic prediction of genes functionally linked to CRISPR-Cas systems by gene neighborhood analysis. Proc. Natl. Acad. Sci. *115*, E5307–E5316.

Sibbald, S.J., Hopkins, J.F., Filloramo, G. V, and Archibald, J.M. (2019). Ubiquitin fusion proteins in algae: implications for cell biology and the spread of photosynthesis. BMC Genomics *20*, 38.

Silas, S., Makarova, K.S., Shmakov, S., Páez-Espino, D., Mohr, G., Liu, Y., Davison, M., Roux, S., Krishnamurthy, S.R., and Fu, B.X.H. (2017a). On the origin of reverse transcriptase-using CRISPR-Cas systems and their hyperdiverse, enigmatic spacer repertoires. MBio *8*.

Silas, S., Lucas-Elio, P., Jackson, S.A., Aroca-Crevillén, A., Hansen, L.L., Fineran, P.C., Fire, A.Z., and Sánchez-Amat, A. (2017b). Type III CRISPR-Cas systems can provide redundancy to counteract viral escape from type I systems. Elife *6*, e27601.

Smith, G.R. (2012). How RecBCD enzyme and Chi promote DNA break repair and recombination: a molecular biologist's view. Microbiol. Mol. Biol. Rev. *76*, 217–228.

Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., and Ideker, T. (2011). Cytoscape 2.8: New features for data integration and network visualization. Bioinformatics *27*, 431–432.

Sontheimer, E.J., and Davidson, A.R. (2017). Inhibition of CRISPR-Cas systems by mobile genetic elements. Curr. Opin. Microbiol. *37*, 120–127.

Van Soolingen, D., De Haas, P.E., Hermans, P.W., Groenen, P.M., and Van Embden, J.D. (1993). Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of Mycobacterium tuberculosis. J. Clin. Microbiol. *31*, 1987–1995.

Staals, R.H.J., and Brouns, S.J.J. (2013). Distribution and mechanism of the type I CRISPR-Cas systems. In CRISPR-Cas Systems, (Springer), pp. 145–169.

Staals, R.H.J., Zhu, Y., Taylor, D.W., Kornfeld, J.E., Sharma, K., Barendregt, A., Koehorst, J.J., Vlot, M., Neupane, N., and Varossieau, K. (2014). RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus thermophilus. Mol. Cell *56*, 518–530.

Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C., and Doudna, J.A. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature *507*, 62–67.

Strich, J.R., and Chertow, D.S. (2018). CRISPR-cas biology and its application to infectious diseases. J. Clin. Microbiol. *57*, 1–14.

Strzyz, P. (2020). CRISPR–Cas9 wins Nobel. Nat. Rev. Mol. Cell Biol. 1.

Sun, C.L., Barrangou, R., Thomas, B.C., Horvath, P., Fremaux, C., and Banfield, J.F. (2013). Phage mutations in response to CRISPR diversification in a bacterial population. Environ. Microbiol. *15*, 463–470.

Sun, C.L., Thomas, B.C., Barrangou, R., and Banfield, J.F. (2016). Metagenomic reconstructions of bacterial CRISPR loci constrain population histories. ISME J. *10*, 858–870.

Swarts, D.C., Mosterd, C., Van Passel, M.W.J., and Brouns, S.J.J. (2012). CRISPR interference directs strand specific spacer acquisition. PLoS One *7*, e35888.

Swarts, D.C., van der Oost, J., and Jinek, M. (2017). Structural basis for guide RNA processing and seed-dependent DNA targeting by CRISPR-Cas12a. Mol. Cell *66*, 221–233.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P., et al. (2015). STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. *43*, D447–D452.

Szollosi, G.J., Boussau, B., Abby, S.S., Tannier, E., and Daubin, V. (2012). Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. Proc. Natl. Acad. Sci. *109*, 17513–17518.

Szumilas, M. (2010). Explaining odds ratios. J. Can. Acad. Child Adolesc. Psychiatry *19*, 227.

Tatusov, R. L., Galperin, M.Y., Natale, D.A., Koonin, E.V. (2000). The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. *28*, 33–36.

Tatusov, R.L., Koonin, E. V, and Lipman, D.J. (1997). A genomic perspective on protein families. Science (80-.). *278*, 631–637.

Taylor, D.W., Zhu, Y., Staals, R.H.J., Kornfeld, J.E., Shinkai, A., van der Oost, J., Nogales, E., and Doudna, J.A. (2015). Structures of the CRISPR-Cmr complex reveal mode of RNA target positioning. Science (80-.). *348*, 581–585.

Trasanidou, D., Gerós, A.S., Mohanraju, P., Nieuwenweg, A.C., Nobrega, F.L., and Staals, R.H.J. (2019). Keeping crispr in check: Diverse mechanisms of phage-encoded anti-crisprs. FEMS Microbiol. Lett. *366*, 1–14.

Varble, A., Meaden, S., Barrangou, R., Westra, E.R., and Marraffini, L.A. (2019). Recombination between phages and CRISPR−cas loci facilitates horizontal gene transfer in staphylococci. Nat. Microbiol.

Vercoe, R.B., Chang, J.T., Dy, R.L., Taylor, C., Gristwood, T., Clulow, J.S., Richter, C., Przybilski, R., Pitman, A.R., and Fineran, P.C. (2013). Cytotoxic Chromosomal Targeting by CRISPR/Cas Systems Can Reshape Bacterial Genomes and Expel or Remodel Pathogenicity Islands. PLoS Genet. *9*.

Viswanath, B., Mislove, A., Cha, M., and Gummadi, K.P. (2009). On the evolution of user interaction in facebook. In Proceedings of the 2nd ACM Workshop on Online Social Networks, pp. 37–42.

Wang, R., and Li, H. (2012). The mysterious RAMP proteins and their roles in small RNA-based immunity. Protein Sci. *21*, 463–470.

Wang, R., Li, M., Gong, L., Hu, S., and Xiang, H. (2016a). DNA motifs determining the accuracy of repeat duplication during CRISPR adaptation in Haloarcula hispanica. Nucleic Acids Res. *44*, 4266–4277.

Wang, X., Yao, D., Xu, J.-G., Li, A.-R., Xu, J., Fu, P., Zhou, Y., and Zhu, Y. (2016b). Structural basis of Cas3 inhibition by the bacteriophage protein AcrF3. Nat. Struct. Mol. Biol. *23*, 868–870.

Watson, B.N.J., Staals, R.H.J., and Fineran, P.C. (2018). CRISPR-Cas-mediated phage resistance enhances horizontal gene transfer by transduction. MBio *9*.

Watson, B.N.J., Easingwood, R.A., Tong, B., Wolf, M., Salmond, G.P.C., Staals, R.H.J., Bostina, M., and Fineran, P.C. (2019). Different genetic and morphological outcomes for phages targeted by single or multiple CRISPR-Cas spacers.

Wei, W., Zhang, S., Fleming, J., Chen, Y., Li, Z., Fan, S., Liu, Y., Wang, W., Wang, T., Liu, Y., et al. (2019). Mycobacterium tuberculosis type III-A CRISPR/Cas system crRNA and its maturation have atypical features. FASEB J. *33*, 1496–1509.

Weinberger, A.D., Sun, C.L., Pluciński, M.M., Denef, V.J., Thomas, B.C., Horvath, P., Barrangou, R., Gilmore, M.S., Getz, W.M., and Banfield, J.F. (2012). Persisting viral sequences shape microbial CRISPR-based immunity. PLoS Comput Biol *8*, e1002475.

Weissman, J.L., Fagan, W.F., and Johnson, P.L.F. (2018). Selective maintenance of multiple CRISPR arrays across prokaryotes. Cris. J. *1*, 405–413.

Welch, R.A., Burland, V., Plunkett, G., Redford, P., Roesch, P., Rasko, D., Buckles, E.L., Liou, S.-R., Boutin, A., and Hackett, J. (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc. Natl. Acad. Sci. *99*, 17020–17024.

Westra, E., and Levin, B. (2020). How important is CRISPR-Cas for protecting natural populations of bacteria against infections by mobile genetic elements? 1–9.

Westra, E.R., Buckling, A., and Fineran, P.C. (2014). CRISPR-Cas systems: Beyond adaptive immunity. Nat. Rev. Microbiol. *12*, 317–326.

Westra, E.R., Van houte, S., Oyesiku-Blakemore, S., Makin, B., Broniewski, J.M., Best, A., Bondy-Denomy, J., Davidson, A., Boots, M., and Buckling, A. (2015). Parasite exposure drives selective evolution of constitutive versus inducible defense. Curr. Biol. *25*, 1043–1049.

Westra, E.R., Dowling, A.J., Broniewski, J.M., and Van Houte, S. (2016). Evolution and Ecology of CRISPR. Annu. Rev. Ecol. Evol. Syst. *47*, 307–331.

Wheatley, R.M., and Maclean, R.C. (2020). CRISPR-Cas systems restrict horizontal gene transfer in Pseudomonas aeruginosa 2 Running title: CRISPR-Cas systems in Pseudomonas aeruginosa 3 4. BioRxiv 2020.09.19.304717.

Whelan, F.J., Rusilowicz, M., and McInerney, J.O. (2020). Coinfinder: detecting significant associations and dissociations in pangenomes. Microb. Genomics *6*.

Wiedenheft, B., van Duijn, E., Bultema, J.B., Waghmare, S.P., Zhou, K., Barendregt, A., Westphal, W., Heck, A.J.R., Boekema, E.J., and Dickman, M.J. (2011). RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. Proc. Natl. Acad. Sci. *108*, 10092–10097.

Wigley, D.B. (2013). Bacterial DNA repair: recent insights into the mechanism of RecBCD, AddAB and AdnAB. Nat. Rev. Microbiol. *11*, 9–13.

Williams, T.A., Szöllősi, G.J., Spang, A., Foster, P.G., Heaps, S.E., Boussau, B., Ettema, T.J.G., and Embley, T.M. (2017). Integrative modeling of gene and genome evolution roots the archaeal tree of life. Proc. Natl. Acad. Sci. *114*, E4602–E4611.

Wright, A. V, and Doudna, J.A. (2016). Protecting genome integrity during CRISPR immune adaptation. Nat. Struct. Mol. Biol. *23*, 876.

Xiao, Y., Luo, M., Hayes, R.P., Kim, J., Ng, S., Ding, F., Liao, M., and Ke, A. (2017). Structure basis for directional R-loop formation and substrate handover mechanisms in type I CRISPR-Cas system. Cell *170*, 48–60.

Yamada, M., Watanabe, Y., Gootenberg, J.S., Hirano, H., Ran, F.A., Nakane, T., Ishitani, R., Zhang, F., Nishimasu, H., and Nureki, O. (2017). Crystal structure of the minimal Cas9 from Campylobacter jejuni reveals the molecular diversity in the CRISPR-Cas9 systems. Mol. Cell *65*, 1109–1121.

Yang, H., and Patel, D.J. (2017). Inhibition mechanism of an anti-CRISPR suppressor AcrIIA4 targeting SpyCas9. Mol. Cell *67*, 117–127.

Yarza, P., and Munoz, R. (2014). The all-species living tree project. In Methods in Microbiology, (Elsevier), pp. 45–59.

Yoganand, K.N.R., Sivathanu, R., Nimkar, S., and Anand, B. (2017). Asymmetric positioning of Cas1–2 complex and Integration Host Factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type IE system. Nucleic Acids Res. *45*, 367–381.

Yosef, I., Goren, M.G., and Qimron, U. (2012). Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Res. *40*, 5569–5576.

Zhang, J., Dean, A.M., Brunet, F., and Long, M. (2004). Evolving protein functional diversity in new genes of Drosophila. Proc. Natl. Acad. Sci. *101*, 16246–16250.

Zhang, J., Graham, S., Tello, A., Iu, H., and White, M.F. (2016). Multiple nucleic acid cleavage modes in divergent type III CRISPR systems. Nucleic Acids Res. *44*, 1789–1799.

Zhang, Y., Heidrich, N., Ampattu, B.J., Gunderson, C.W., Seifert, H.S., Schoen, C., Vogel, J., and Sontheimer, E.J. (2013). Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis. Mol. Cell *50*, 488–503.

Zheng, Z., Zhang, Y., Liu, Z., Dong, Z., Xie, C., Bravo, A., Soberón, M., Mahillon, J., Sun, M., and Peng, D. (2020). The CRISPR-Cas systems were selectively inactivated during evolution of Bacillus cereus group for adaptation to diverse environments. ISME J. 1–15.

Appendix A:

Supplementary Materials for Chapter 2

S1. Odds ratio (OR) test on GO annotations

To analysis the distribution of composite gene in eukaryotes and prokaryotes, apart from the COG annotations, we carried out the Odds ratio (OR) test on GO annotations as well. The results were collected from EggNOG output. Analysis was similar to COG analysis, for both eukaryotes and prokaryotes genes, counted the frequency of each function, OR value, upper and lower confidential interval (CI) values with conservative Bonferroni correction. Composite genes acted as cellular components and involved in biological process were statistical more likely to be from eukaryotes but genes in molecular function did not show much difference between eukaryotes and prokaryotes. The detailed information was showing in ST1.

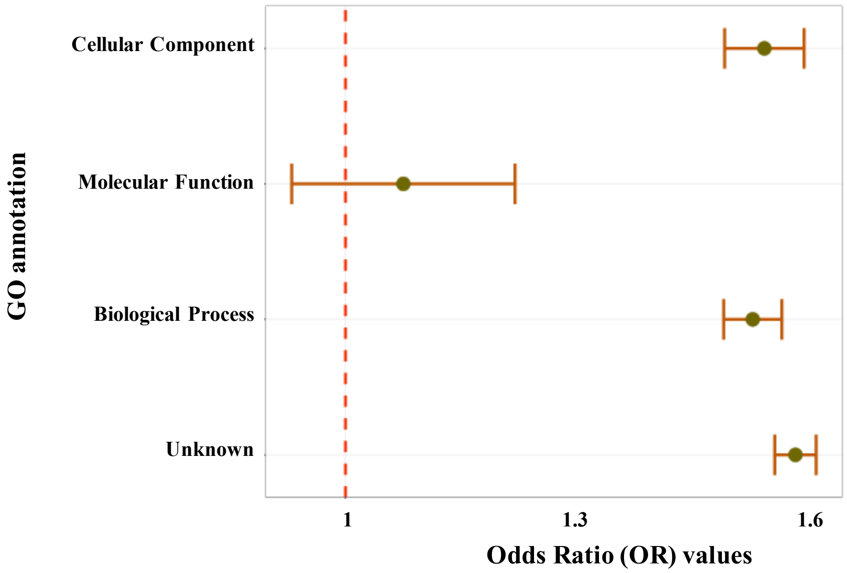


Figure S2.1. Numbers of OR, corrected upper 95% CI and lower 95% CI value across all GO annotations.

Table S2.1. Numbers of composite and non-composite genes from eukaryotes and prokaryotes in different functional categories, OR, corrected upper and lower 95% CI values.

		Composite		Non-composite			Corrected	Corrected
Functional annotations		Eukaryotes	Prokaryotes	Eukaryotes	Prokaryotes	OR	lower	upper
		(a)	(c)	(b)	(d)		95% CI	95% CI
COG	A	4633	27	11303	40	0.61	0.281	1.314
	W	1019	19	2174	44	1.09	0.461	2.556
	C	3408	5603	9049	19626	1.32	1.22	1.426
	L	4043	5159	8477	14734	1.36	1.261	1.472
	T	16220	3376	41878	13821	1.59	1.485	1.694
	J	4594	4213	12711	18724	1.61	1.49	1.731
	P	4674	4867	9861	16715	1.63	1.511	1.754
	E	4163	7247	7456	22038	1.7	1.578	1.826
	D	2620	638	6151	2620	1.75	1.499	2.041
	G	5291	4078	11275	15524	1.79	1.658	1.925
	Q	2905	1291	6917	5547	1.81	1.605	2.029
	N	89	535	304	3474	1.9	1.275	2.833
	B	2303	44	4606	169	1.92	1.132	3.259
	H	1954	3231	3633	11762	1.96	1.761	2.177
	I	4701	1919	10648	8966	2.06	1.876	2.268
	U	6011	814	16986	4758	2.07	1.824	2.346
	O	11432	2225	29759	12086	2.09	1.928	2.258
	F	1538	1777	3174	7748	2.11	1.863	2.396
	M	1916	3310	4139	15532	2.17	1.958	2.41
	V	826	1161	1906	6145	2.29	1.95	2.698
	K	8418	3372	22301	22575	2.53	2.358	2.709
	Z	4391	11	10486	91	3.46	1.29	9.304
	S	37613	9440	122130	125751	4.1	3.951	4.26
GO	Cellular Component	21063	9743	57403	41686	1.57	1.504	1.639
	Molecular Function	1832	1863	4469	4837	1.06	0.944	1.2
	Biological Process	60444	16205	139926	58163	1.55	1.503	1.6
	Unknown	58928	35835	283635	280001	1.62	1.588	1.66

S2. Sequence Similarity Network (SSN) of Composite genes in *Prochlorococcus marinus*

Network-based analyse has the potential to become one of best ways of showing the complexity and diversity of gene evolution. In a network, nodes (representing genes) are connected by edges when they show are potentially homologous. Previous studies have shown the important role of gene fusion in cyanobacteria, thus here we used composite genes in *Prochlorococcus marinus* as an example to show how potentially can network been utilized in investigating genomic evolutionary relationships (Méheust et al., 2016). Using our apporach, we detected 218 composite genes in *P. marinus*, which showed significant similarity with 4,505 components. In the network, there were 4,717 nodes attached by 137,681 edges that formed 171 connected components (CCs).

All nodes were coloured by the COG category which the corresponding gene belongs to. Interestingly, most CCs consisted of genes from a single category code. In most non-transitive triplets, composite genes shared the same COG category with its component genes even though there was no obvious homology between these component genes. In Figure S2, there were several composite genes which showed the same functional type with a small subset of its component genes. For instance, Node 1 shared a COG category (labelled in green) with 2 components rather than 105 component genes. Another interesting observation was that composite and components involved in non-transitive triplets belonging to different COG categories. For example, Node 2 from *P. marius* was involved in tRNA-methyltransferase function whereas Node 3 from *P. marius* included functions such as phosphoribosyl-AMP cyclohydrolase and phosphoribosyl-ATP diphosphatase HisIE. However, the gene Family 4 that connected these two families that play roles in generation of 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase and they were from *Galdieria sulphuraria* and *Cyanidioschyzon merolae*. These genes had quite different functions but formed composite gene origin from non-homologous species showed potential for further research.

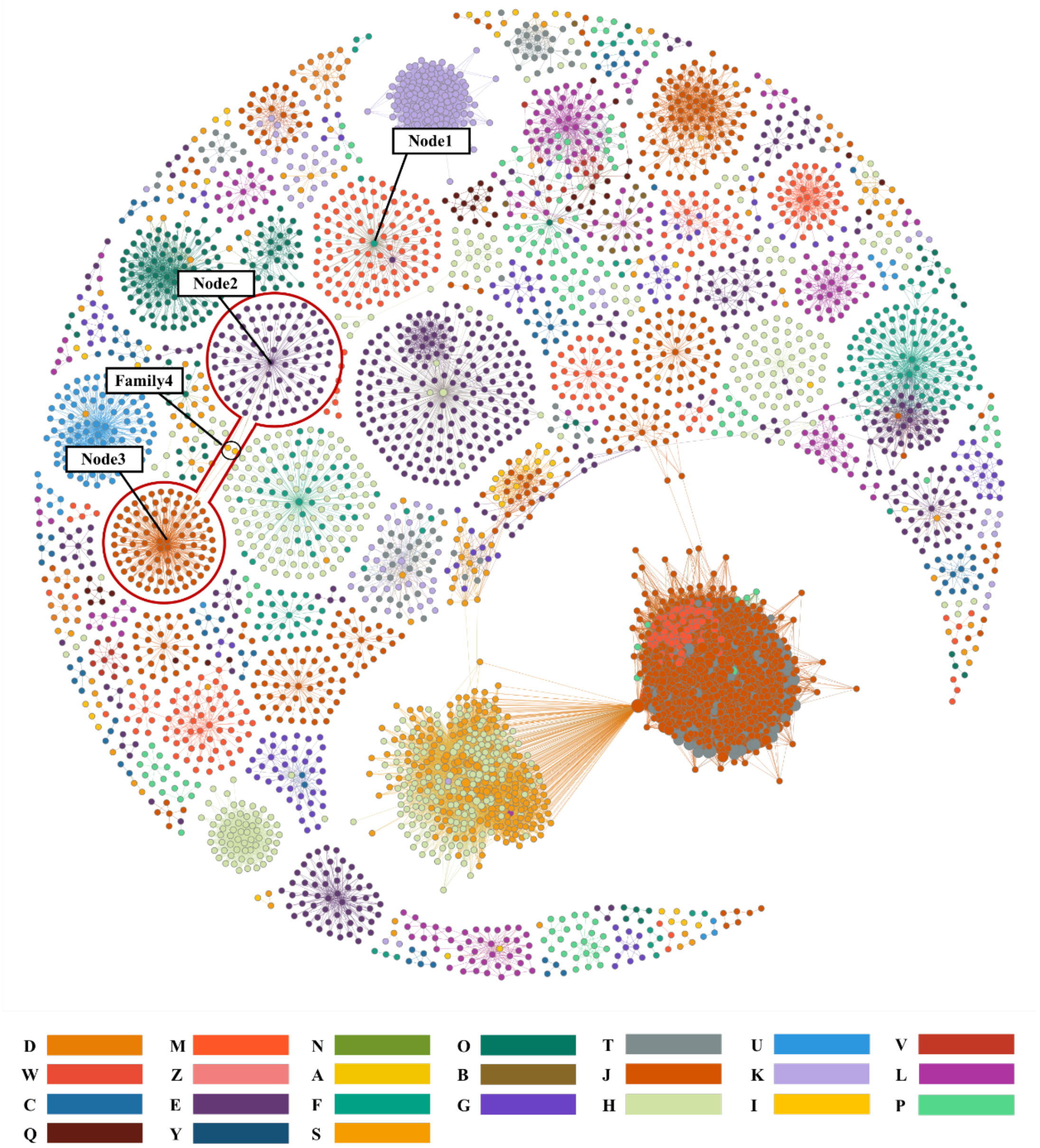


Figure S2. Example of sequence similarity network of composite and related component genes from *Prochlorococcus marinus*. The Fruchterman-Reingold algorithm was used to determine node layout. All nodes were coloured based on its COG category codes.

Reprint: Ou and McInerney (2019)

genes

MDPI

Article

Eukaryote Genes Are More Likely than Prokaryote Genes to Be Composites

Yaqing Ou [1,*] **and James O. McInerney** [1,2,*]

[1] Division of Evolution and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester M13 9PL, UK
[2] School of Life Sciences, University of Nottingham, Nottingham NG7 2UH, UK
* Correspondence: Yaqing.Ou@manchester.ac.uk (Y.O.); James.McInerney@nottingham.ac.uk (J.O.M.)

Received: 29 July 2019; Accepted: 23 August 2019; Published: 28 August 2019

Abstract: The formation of new genes by combining parts of existing genes is an important evolutionary process. Remodelled genes, which we call composites, have been investigated in many species, however, their distribution across all of life is still unknown. We set out to examine the extent to which genomes from cells and mobile genetic elements contain composite genes. We identify composite genes as those that show partial homology to at least two unrelated component genes. In order to identify composite and component genes, we constructed sequence similarity networks (SSNs) of more than one million genes from all three domains of life, as well as viruses and plasmids. We identified non-transitive triplets of nodes in this network and explored the homology relationships in these triplets to see if the middle nodes were indeed composite genes. In total, we identified 221,043 (18.57%) composites genes, which were distributed across all genomic and functional categories. In particular, the presence of composite genes is statistically more likely in eukaryotes than prokaryotes.

Keywords: composite genes; sequence similarity networks; odds ratio test

1. Introduction

Reticulation occurs when two or more evolutionary lineages merge, and consequently, reticulation cannot be visualised or analysed using tree-like models of evolution. We see reticulate events occurring during meiotic recombination, horizontal gene transfer (HGT, also known as lateral gene transfer) [1], exon shuffling [2], and hybrid speciation [3] for example. Merger events can be seen at multiple levels, such as genes, genomes, operons and gene clusters.

This paper focuses on the combination of genetic fragments from unrelated gene families to produce a single gene. This process of gene fusion occurs when parental (or component) genes merge to form a new gene called a composite (or fused) gene [2,4]. Because reticulate evolution cannot be adequately represented using tree-like representations, we constructed sequence similarity networks (SSNs, also known as protein/gene similarity networks) and visualised them using Gephi [5] and Cytoscape [6]. In these kinds of networks, gene, genome or species data can be used to detect recombination events. In the SSNs that we have constructed, genes or proteins are represented as nodes while inferences of homology between genes are represented by edges. Within the framework of the SSN, some special relationships, such as non-transitive triplets when two component genes have no overlap, can be identified as motifs in the network. SSNs have been used elsewhere in order to investigate the existence of composite genes [7,8]. In an analysis of 15 eukaryotic genomes Haggerty et al. [7] constructed a network that contained a giant connected component (GCC) where one quarter of all sequences were identified as composite genes and approximately 10% of sequences were identified as multi-composite genes (those formed from the union of two or more composite genes).

Moreover, Coleman et al. [8] used SSNs to explore 1642 antibiotic resistance genes derived from more than 100 species. They found 73 fused genes using the FusedTriplets software [9,10], which accounted for 4.43% of the total gene count. In addition, Jachiet et al. [10], using the MosaicFinder software, found gene fusions in both cellular organisms and mobile genetic elements (MGEs). In another analysis using the same kind of approach, viruses were suggested to consist of only 8–15% of composite genes, with this low number being attributed to the blurry boundaries between viral gene families [11]. In addition, gene fusion has been shown to have played an essential role in the evolution of the cellular life cycle, with composite gene formation seen in genes related to chromatin structure and nucleotide metabolism [12]. Also, Ocaña-Pallarès et al. [13] concluded that there was a significant role for gene fusion in the origin of eukaryotes, as evidenced by SSN built from eukaryotic EUKaryotic restricted Nitrate Reductase (EUKNR) and similar eukaryotic and prokaryotic sequences. The result indicated that EUKNR was formed by a fusion of eukaryotic sulfite oxidases (SUOX, N-terminal) and NADH (C-terminal) reductases. Therefore, while it is clear that gene fusion is a common feature of genes, a comprehensive comparison across a broad range of taxa and molecule types would provide more evidence for its frequency and impact.

In this paper, we describe an approach to identify composite genes using a dataset of 1875 completed genomes, comprising more than one million sequences, from all three domains of life as well as from MGEs. We tested whether the rate of gene remodelling has been uniform across all of life, and all cellular functional categories.

2. Materials and Methods

2.1. Dataset Construction and BLAST Analysis

A total of 1,190,265 protein sequences were collected from the RefSeq database at the National Centre for Biotechnology Information [14]. We manually selected taxa in order to maximise diversity, while also ensuring computational tractability. The final dataset covered 261 species from the main representative lineages, belonging to 36 eukaryotes (13 phyla, 21 classes), 56 archaea (4 phyla, 9 classes), 90 bacteria (25 phyla, 32 classes), 79 viruses and 1,614 plasmids. Homology between pairs of amino acid sequences was inferred using an all-versus-all protein BLAST (BLASTP version 2.4.0, NCBI, Bethesda, MD, United States), with an E-value cutoff of 1e-5, 5000 max target sequences, and soft masking parameter (the others by default) [15]. The dataset species information and download paths are available at https://github.com/JMcInerneyLab/CompositeGenes/blob/master/accession.txt.

2.2. Composite Gene Identification

Using the BLAST results as input, we identified composite genes as motifs of triplets in the graph where there was a "non-transitive" relationship between three nodes [4]. Composite gene detection was carried out by the CompositeSearch program [16] when associated component genes have no overlap theoretically, with default identity cutoff of 30% and 20 amino acid overlaps to limit false negative error. The CompositeSearch output contains information on composite genes, component genes and the families to which they belong. This output was depicted, explored, and manipulated using Gephi (version 0.9.2, The Gephi Consortium, Paris, France) [5].

Because the proportion of composite gene from different domains might be affected by biased sequence database sampling, we randomly sampled 50,000 genes from archaea, bacteria, eukaryotes and plasmids respectively. These random samples were taken forward for analysis in the same way as the original data. The major difference between the sub-sampled datasets and the original data was that in the subsampled datasets, the number of genes from each of the four kinds of dataset was the same. We used CompositeSearch in order to construct an SSN from the BLASTP output of the subsampled datasets containing 200,000 genes. These SSNs were then used in order to identify composite genes. Sampling was repeated 100 times and the results were summarised graphically.

2.3. Functional Annotations

We used EggNOG (version 4.5.1, Computational Biology Group–EMBL, Heidelberg, Germany) [17] in order to assign gene functional categories. The analysis was carried out through the web interface using the DIAMOND [18] mapping mode. In the output, genes were assigned to different Orthologous Groups (OGs), and each OG had functional annotations that included Clusters of Orthologous Groups (COGs) functional categories: COG for universal Bacteria, EuKaryotic Orthologous Groups (KOGs) for Eukaryotes and arKOGs for Archaea [19]; Gene Ontology (GO) terms [20]; Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and SMART/Pfam protein domains. Both composite and non-composite genes were placed into at least one of 23 COG categories and at least one of four GO terms.

2.4. Statistical Analysis

In the EggNOG output, each gene has a detailed functional annotation and is associated with at least one general COG category code (A to Z apart from R and X). Because of recombination, the category code for a given gene could be single letter like 'A' or multiple letters such as 'ABC'. When counting the number of genes that possess a particular function, if a multiple letter category was selected, we counted this gene multiple times. For instance, if the most common COG category for a gene was 'ABC', and then this gene was counted three times as A, B and C.

To investigate the distribution of composite genes and non-composite genes among eukaryotes and prokaryotes, an odds ratio (OR) test [21] was carried out. OR tests are normally used to test the strength of the association between two events. Here, for each protein function, we used an OR test to test the association between gene origin and the likelihood of fusion Equation (1).

$$OR = \frac{a/c}{b/d} = \frac{ad}{bc} \tag{1}$$

where a is the number of composite eukaryote genes, b is the number of non-composite eukaryote genes, c is the number of composite prokaryote genes, d is the number of non-composite prokaryote genes. The 95% confidence intervals (CI) were calculated by

$$Upper\ 95\%\ CI = e\left[\ln(OR) + 1.96\sqrt{\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right)}\right]$$

$$Lower\ 95\%\ CI = e\left[\ln(OR) - 1.96\sqrt{\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right)}\right] \tag{2}$$

For all OR tests that were carried out on 24 COG categories, we used a conservative Bonferroni correction [22] to limit type I error. The critical level of significance was initially set as $\alpha = 5\%$, we corrected it as $\alpha/2N$, N is the number of performed tests, which in our case is 24. The new significance level is 0.1% and corresponding confidence coefficient of 99.9% is 3.09 standard deviations, using the standard normal distribution table. The corrected CI was calculated by

$$Upper\ 95\%\ CI = e\left[\ln(OR) + 3.09\sqrt{\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right)}\right]$$

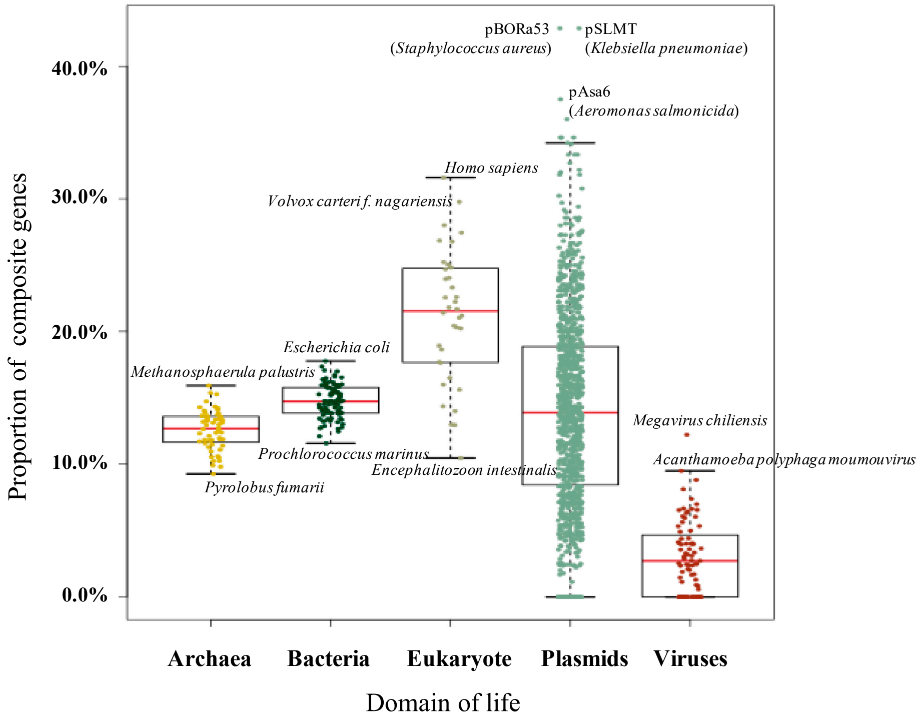
$$Lower\ 95\%\ CI = e\left[\ln(OR) - 3.09\sqrt{\left(\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}\right)}\right] \tag{3}$$

3. Results

3.1. Pervasive Existence of Composite Genes across All of Life

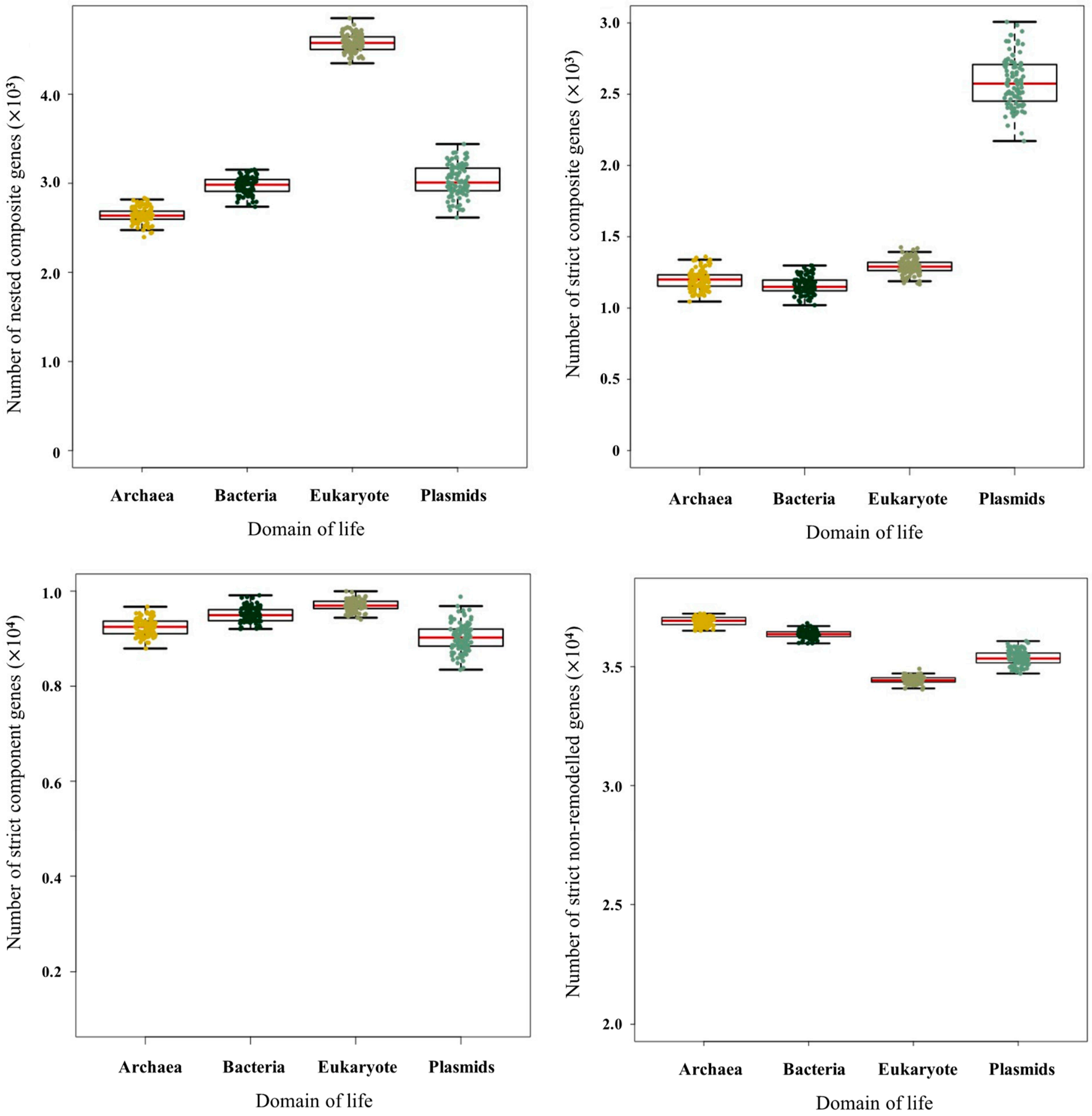
We assembled a dataset of 1,190,256 genes from 36 eukaryotes, 56 archaea, 90 bacteria, 79 viruses and 1614 plasmids from more than 60 taxonomic classes. Following an all-versus-all BLAST, a total of 540,325,758 significant hits were detected. Using CompositeSearch, an SSN containing 1,025,263 nodes and 109,650,422 edges was constructed. In this network, 221,043 composite genes (18.57% of the gene dataset, Figure 1a) were identified, linked to 603,604 component genes. Collectively, these genes were assigned to 360,981 gene families.

To gain a better understanding of those genes involved in non-homologous recombination, all genes were categorized into four groups: nested composite genes, strict composite genes, strict component genes, and non-remodelled genes (Figure 2). Nested composite genes have been formed by the merging of at least two sequences but are additionally involved in other non-transitive triplets as components; that is to say they themselves are composites but also form other composites. In contrast to nested composites, strict composite genes only act as composite genes in the network, similar to strict component genes. Non-remodelled genes do not show evidence of having participated in any recombination events. In our dataset, 181,157 genes as nested composite genes, 39,886 genes were identified as strict composite genes, 422,447 as strict component genes, and 546,775 as non-remodelled genes (Figure 3a).



(**a**)

Figure 1. *Cont.*



(**b**)

Figure 1. Pictures of composite and component genes among different domains. (**a**) Proportion of composite genes within different domains and mobile genetic elements. Dots represent individual genomes. (**b**) Numbers of nested composite, strict composite, strict component, and non-remodelled genes within different domains for each of the 100 replicates of equal sampling. All analyses were replicated 100 times and each replicate is represented by a dot.

Within 182 species across the three domains of life, remodelled composite genes were discovered in all species, indicating that gene fusion is, and has been, widespread across all life on Earth. Overall, 23.66% (205,913 composites identified from 870,120 eukaryotic and prokaryotic genes) of examined genes were identified as composite. However, there was a considerable amount of variation in the proportion of composite genes across species and molecule type. Table 1 presents the ten genomes with the highest and lowest rates of composite genes among eukaryotes and prokaryotes. Composite genes account for almost one third of the genomes of *Homo sapiens*, *Volvox carteri f. nagariensis* and *Aureococcus anophagefferens*.

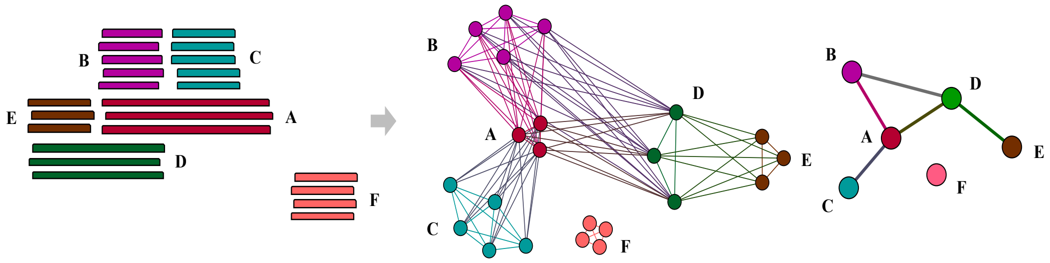
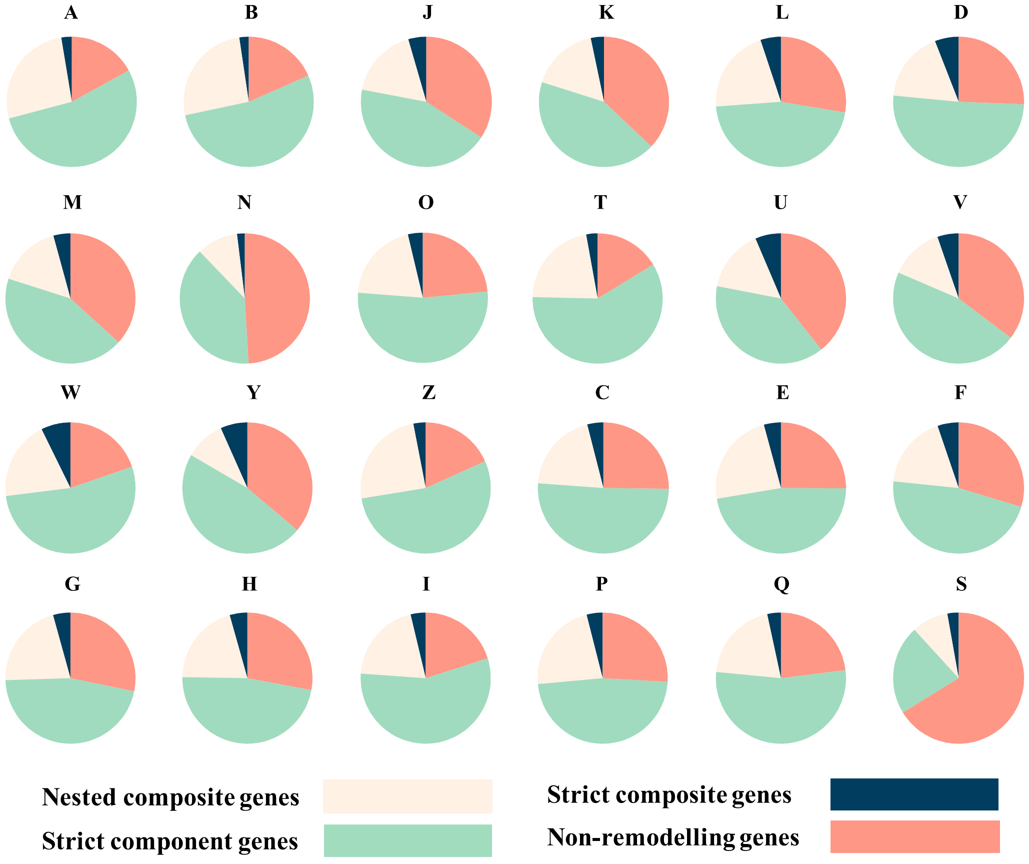
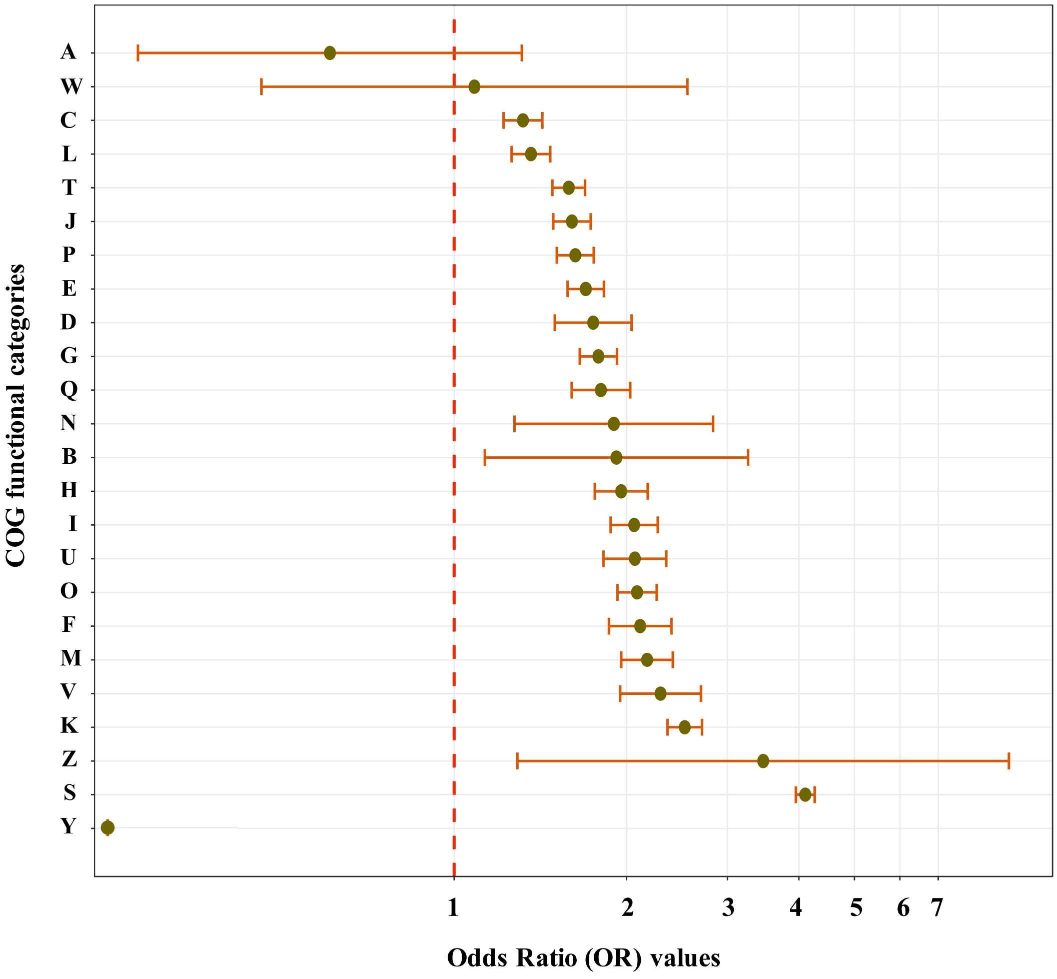


Figure 2. Sample patterns of nested composite, strict composite, strict component, and non-remodelled gene families. Genes in family A not only participate in the fusion of genes in family D as component genes but are also formed by genes from family B and C as composite gens; this is regarded as nested composite genes. In contrast, genes in family B, C and E belong to strict component families which only act as component genes in this network. Similarly, genes in family D as members of a strict composite family. In additional, family F is non-remodelled gene. Also, because there is no overlap between gene family A and C, gene family B and E so 'A-B-C' and 'B-D-E' can be regarded as non-transitive triplets.



(**a**)

Figure 3. *Cont.*



(**b**)

Figure 3. Function analysis of composite and non-composite genes. (**a**) Numbers of nested composite, strict composite, strict component, and non-remodelled genes across all COG categories. (**b**) Numbers of OR, upper 95% CI and lower 95% CI value (after Bonferroni correction) across all COG functions. The detailed numbers are shown in Table S1. There was not composite gene identified from prokaryote in category Y in this dataset, so OR test was not applied. Apart from A and W, which span 1.0, the odds of composite gene presence in all COG categories shows statistically significant tendency in eukaryotes. A: RNA processing and modification; B: chromatin structure and dynamics; C: energy production and conversion; D: cell cycle control and mitosis; E: amino acid metabolism and transport; F: nucleotide metabolism and transport; G: carbohydrate metabolism and transport; H: coenzyme metabolism; I: lipid metabolism; J: translation; K: transcription; L: replication and repair; M: cell wall/membrane/envelope biogenesis; N: Cell motility; O: post-translational modification: protein turnover, chaperone functions; P: Inorganic ion transport and metabolism; Q: secondary metabolites biosynthesis: transport and catabolism; T: signal transduction; U: intracellular trafficking and secretion; V: defence mechanisms; W: extracellular structures; Y: Nuclear structure; Z: cytoskeleton; S: function unknown.

As shown in Figure 1a, the proportion of composite genes often shows a wide distribution, depending on the classification of the genome in which the gene is found. Among cellular lifeforms, eukaryote genomes contain the highest proportion of composite genes on average (22.66%), followed by bacteria (14.76%) and then archaea (12.78%). However, the distributions are quite wide though prokaryote species manifested a narrower distribution of composite frequency when compared with eukaryotes. When considering mobile genetic elements, the average percentage of composite genes in plasmids (14.69%) is almost the same as bacteria but is noticeably higher than the average seen for virus genes (4.82%).

Table 1. The ten species that contain the highest (left) and lowest (right) proportions of composite genes. All species that contains more than 24% composite genes are from eukaryotes, whereas most species that contain less than 11% composite genes are from archaea (Crenarchaeota family, mostly).

Species	Total Number of Genes	Number of Composite Genes	Proportion	Species	Total Number of Genes	Number of Composite Genes	Proportion
Homo sapiens	109,018	34,455	31.60%	*Fervidicoccus fontis*	1385	152	10.97%
Volvox carteri f. nagariensis	14,436	4298	29.77%	*Thermoproteus uzoniensis*	2112	224	10.61%
Aureococcus anophagefferens	11,520	3227	28.01%	*Nanoarchaeum equitans*	540	57	10.56%
Capsaspora owczarzaki	8792	2413	27.45%	*Staphylothermus marinus*	1598	168	10.51%
Chlorella variabilis	9780	2626	26.85%	*Encephalitozoon intestinalis*	1939	203	10.47%
Polysphondylium pallidum	12,367	3313	26.79%	*Ignisphaera aggregans*	1930	198	10.26%
Monosiga brevicollis	9203	2322	25.23%	*Methanopyrus kandleri*	1687	173	10.25%
Salpingoeca rosetta	11,731	2939	25.05%	*Pyrobaculum neutrophilum*	1966	195	9.92%
Allomyces macrogynus	19,446	4829	24.83%	*Hyperthermus butylicus*	1681	165	9.82%
Tetrahymena thermophila	10,626	2625	24.70%	*Pyrolobus fumarii*	1885	175	9.28%

To avoid the effects of unequal sampling in large dataset, we used a jackknife resampling approach in order to generate datasets of 50,000 sequences each from eukaryotes, archaea, bacteria and plasmids. With these uniformly-sized gene sets we used the same analysis methods as for the large dataset: sampling, identifying homologs and constructing SSNs. We then replicated this process 100 times. On average, across all replicates, 19,443 (9.72%) genes were identified as composite genes (Figure 1b), which is approximately half the percentage identified from the large dataset (18.57%). The difference indicates that the detection rate of composite genes is related to genomic sequence sampling size and therefore, the reporting of composite genes is always a lower bound for the actual percentage. The resampling procedure was designed to analyse composite gene distribution while attempting to normalise for the difference in data size for each of the four main classifications (eukaryote, bacteria, archaea and plasmids). Plasmids have the highest proportion of strict composite genes while eukaryotes have the largest proportion of nested composite genes (Figure 1b). Nonetheless, even though there is no obvious difference between eukaryotes and prokaryotes in terms of the number of nested composite genes, strict composite genes are approximately twice as likely in eukaryotes as in archaea and bacteria. Bacteria and archaea are quite similar, in terms of their proportions, for all four categories of remodelled and non-remodelled genes. Finally, strict component genes do not show much difference across any of our genome types though eukaryotes have the highest number of strict components but the lowest number non-remodelled genes.

3.2. Sequence Functional Annotations

The EggNOG mapper program [17] was used to assign functions to all sequences. For all results, COG and GO annotations were used to evaluate functional categories. First, composite genes were found to be widespread across all functional categories (Figure 3, Figure S1). Gene distributions show different patterns across different functions (Figure 3a, Table S1). Genes with unknown function (category S, 66.23% non-remodelled) are less likely to have been remodelled. The category of genes that have the second-lowest rate of remodelling is cell motility (N, 49.08% non-remodelled). Genes in RNA processing and modification (A, 26.52%) and dynamics (B, 25.96%) had the highest rate of nested composites. Conversely, genes involved in signal transduction (T, 59.05%) tend to have the highest proportion of strict component genes whereas genes involved in extracellular structures (W, 7.3%) are more likely to be strict composite.

We used an odds ratio (OR) test and Bonferroni correction (see Methods) on composite and non-composite genes from eukaryotes and prokaryotes in different functional categories in order to understand if genes from different classifications were more likely to be remodelled in one or the other. If the OR value and its upper and lower 95% CI value span 1, we take this as evidence that there is no significant difference in composite gene formation between eukaryotes and prokaryotes, and vice versa. If the OR number is greater than 1, this indicates a positive correlation between remodelling and being from a eukaryotic genome, while if the number is less than 1, it indicates an association between remodelling and being a prokaryote. From the results of these analyses, the frequency of composite genes in eukaryotes were found to be statistically higher than from that of prokaryotes for most kinds of gene (Figure 3b, Figure S1, Table S1). Some exceptions were found for genes in extracellular structures (category W) and RNA processing (category A) whose 95% CI was found to span 1 (Figure 3b). Therefore, across all the species examined, the odds of a gene being a composite if it is a eukaryote is statistically significant higher than if it is a prokaryote.

4. Discussion

Network models such as SSNs have been broadly employed in studies of evolutionary relationships [23] and gene sharing and recombination detection. We carried out a large-scale examination of more than one million genes across 1875 complete proteomes including archaea, bacteria, eukaryote, plasmids and viruses. The results suggest that composite genes exist in all organisms and across all kinds of genes.

Eukaryotes, are known to have originated from the merger of an archaeon and a bacterium [24]. On average, more than one fifth of eukaryote genes show evidence of remodelling by gene fusion and the probability of a gene in our dataset being composite if it is derived from a eukaryote genome are significantly higher than the probability if the genes comes from a prokaryote genome. What is not known at this stage is the process that has led to the change in frequency of gene remodelling. Candidates for the process include the combination of homologous recombination during meiosis, combined with the relatively lower level of horizontal gene transfer (HGT) in eukaryotes compared with prokaryotes. The lower level of HGT means that evolutionary innovation via HGT is more restricted in eukaryotes and this restriction, combined with the opportunities for illegitimate crossover events during meiosis could account for the elevated levels of remodelling. In other words, restricting HGT sets up a situation where composite gene formation is one of the main routes to evolutionary innovation. These findings are consistent with Jachiet et al. [10] who found that eukaryote sequence evolution was highly influenced by gene fusion.

Although evidence of remodelling is quite high in eukaryote genes, plasmid genes also show evidence for a large number of gene fusion events. The average percentage composite genes found in plasmid genomes in our dataset is 14.69%, which is almost as high as the percentage recorded for bacteria. In 2013, Jachiet et al. [10] mined a data set from three domains of life and MGEs, discovered 42% of composite genes were included at least one MGE gene as a component. Likewise, Halary et al. [25] found that the plasmids in Borrelia genes behaved like "private genetic goods" [26] and were much less likely to be involved in gene remodelling or sharing with other taxa. It has been suggested that this restriction in gene sharing contributed to the survival of Borrelia against the host immune environment [27,28]. The high level of remodelling seen in plasmid genes would suggest that MGEs act as a source for remodelling. Corel et al. also found that gene externalization (gene fusion between cellular organism and MGE) played an important role in microbial evolution [29].

In our dataset, compared to non-composite genes, fusion genes are more likely to be involved in chromatin structure and dynamics, extracellular RNA processing and modification, as well as cytoskeleton. It has already been shown for eukaryotes that composite genes have been foundational [12], particularly in photosynthetic lineages (such as ubiquitin-nickel superoxide dismutase fusion protein in algae [30]). Further, a recent published work by McCartney et al. suggested novel functional protein coding genes in human could emerge through transcription-derived gene fusion [31]. Novel composite genes also have been reported in the origin of haloarchaeal lineages contributed by bacteria, which is named as chimeric (ChiC) genes [32]. ChiC genes are more likely to be involved transport and metabolism whereas other composite genes more likely to be involved in replication, recombination and repair, both functions have high composite gene portion in my dataset. In additional, the research from Corel et al. also suggested that recent externalized genes in abundant in replication, recombination, and repair but hard to accumulate, which could be the result of transposon [29]. Moreover, composite genes in viruses tend to be found in nucleotide metabolism and transport, replication and repair, cell wall, membrane and envelope biogenesis as well as post-translational modification. This finding is consistent with Jachiet et al. [11].

5. Conclusions

In conclusion, we applied a network approach in order to investigate composite gene in species across all of life, although the results of this study really only provide a lower-bounds estimate of the extent of gene remodelling, we have been able to show that it is a pervasive and important element of evolutionary history.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4425/10/9/648/s1, Figure S1: Numbers of OR, corrected upper 95% CI and lower 95% CI value across all GO annotations, Table S1: Numbers of composite and non-composite genes from eukaryotes and prokaryotes in different functional categories, OR, corrected upper and lower 95% CI values.

Author Contributions: Conceptualization, Y.O. and J.O.M.; Data curation, Y.O.; Formal analysis, Y.O.; Funding acquisition, J.O.M.; Investigation, Y.O.; Methodology, Y.O. and J.O.M.; Project administration, J.O.M.; Supervision, J.O.M.; Writing—original draft, Y.O.; Writing—review & editing, Y.O. and J.O.M.

Funding: Y.O. is funded by The University of Manchester and Chinese Government Scholarship (201708060436).

Acknowledgments: We thank our lab colleagues Fiona Whelan, Maria Rosa Domingo Sananes, Martin Rusilowicz, Ignacio Riquelme Medina and Weihao Sun for insightful discussion, helps on coding and statistics. This research was supported in part through computational resources The Computational Shared Facility (CSF) by The University of Manchester.

Conflicts of Interest: The authors declare no conflict of interest.

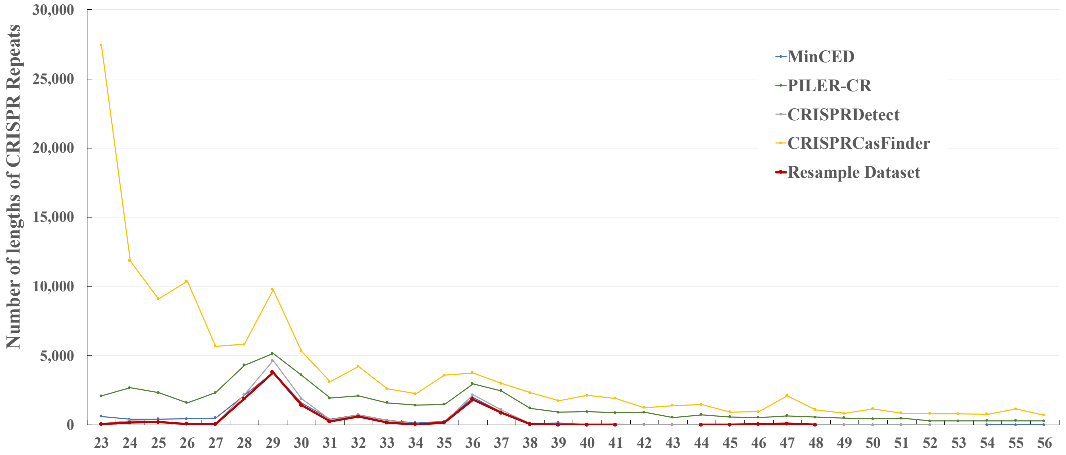
References

1. Nelson-Sathi, S.; Dagan, T.; Landan, G.; Janssen, A.; Steel, M.; McInerney, J.O.; Deppenmeier, U.; Martin, W.F. Acquisition of 1000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 20537–20542. [CrossRef]
2. Oakley, T.H. Furcation and fusion: The phylogenetics of evolutionary novelty. *Dev. Biol.* **2017**, *431*, 69–76. [CrossRef]
3. Linder, C.R.; Moret, B.M.E.; Nakhleh, L.; Warnow, T. Network (reticulate) evolution: Biology, models, and algorithms. In Proceedings of the Ninth Pacific Symposium on Biocomputing (PSB), Big Island, HI, USA, 6–10 January 2004.
4. Corel, E.; Lopez, P.; Méheust, R.; Bapteste, E. Network-Thinking: Graphs to Analyze Microbial Complexity and Evolution. *Trends Microbiol.* **2016**, *24*, 224–237. [CrossRef]
5. Bastian, M.; Heymann, S.; Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. In Proceedings of the Third International AAAI Conference on Weblogs and Social Media, San Jose, CA, USA, 17–20 May 2009.
6. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498–2504. [CrossRef]
7. Haggerty, L.S.; Jachiet, P.A.; Hanage, W.P.; Fitzpatrick, D.A.; Lopez, P.; O'Connell, M.J.; Pisani, D.; Wilkinson, M.; Bapteste, E.; McInerney, J.O. A pluralistic account of homology: Adapting the models to the data. *Mol. Biol. Evol.* **2014**, *31*, 501–516. [CrossRef]
8. Coleman, O.; Hogan, R.; McGoldrick, N.; Rudden, N.; McInerney, J. Evolution by Pervasive Gene Fusion in Antibiotic Resistance and Antibiotic Synthesizing Genes. *Computation* **2015**, *3*, 114–127. [CrossRef]
9. Enright, A.J.; Iliopoulos, I.; Kyrpides, N.C.; Ouzounis, C.A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **1999**, *402*, 86. [CrossRef]
10. Jachiet, P.A.; Pogorelcnik, R.; Berry, A.; Lopez, P.; Bapteste, E. MosaicFinder: Identification of fused gene families in sequence similarity networks. *Bioinformatics* **2013**, *29*, 837–844. [CrossRef]

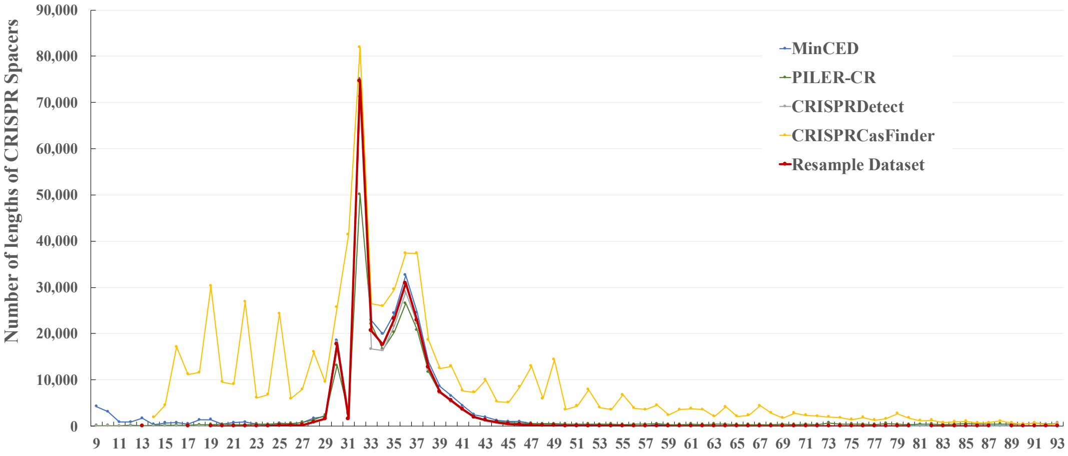
11. Jachiet, P.A.; Colson, P.; Lopez, P.; Bapteste, E. Extensive gene remodeling in the viral world: New evidence for nongradual evolution in the mobilome network. *Genome Biol. Evol.* **2014**, *6*, 2195–2205. [CrossRef]
12. Méheust, R.; Zelzion, E.; Bhattacharya, D.; Lopez, P.; Bapteste, E. Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 3579–3584. [CrossRef]
13. Ocaña-Pallarès, E.; Najle, S.R.; Scazzocchio, C.; Ruiz-Trillo, I. Reticulate evolution in eukaryotes: Origin and evolution of the nitrate assimilation pathway. *PLoS Genet.* **2019**, *15*, e1007986. [CrossRef]
14. Pruitt, K.D.; Tatusova, T.; Maglott, D.R. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **2006**, *35*, D61–D65. [CrossRef]
15. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef]
16. Pathmanathan, J.S.; Lopez, P.; Lapointe, F.-J.; Bapteste, E. CompositeSearch: A generalized network approach for composite gene families detection. *Mol. Biol. Evol.* **2017**, *35*, 252–255. [CrossRef]
17. Huerta-Cepas, J.; Szklarczyk, D.; Forslund, K.; Cook, H.; Heller, D.; Walter, M.C.; Rattei, T.; Mende, D.R.; Sunagawa, S.; Kuhn, M. eggNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **2015**, *44*, D286–D293. [CrossRef]
18. Buchfink, B.; Xie, C.; Huson, D.H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **2015**, *12*, 59. [CrossRef]
19. Tatusov, R.L.; Galperin, M.Y.; Natale, D.A.; Koonin, E.V. The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **2000**, *28*, 33–36. [CrossRef]
20. Consortium, G.O. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **2004**, *32*, D258–D261. [CrossRef]
21. Szumilas, M. Explaining odds ratios. *J. Can. Acad. child Adolesc. Psychiatry* **2010**, *19*, 227.
22. Sedgwick, P. Multiple significance tests: The Bonferroni correction. *BMJ* **2012**, *344*, e509. [CrossRef]
23. Alvarez-Ponce, D.; Lopez, P.; Bapteste, E.; McInerney, J.O. Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, E1594–E1603. [CrossRef] [PubMed]
24. McInerney, J.O.; O'Connell, M.J.; Pisani, D. The hybrid nature of the Eukaryota and a consilient view of life on Earth. *Nat. Rev. Microbiol.* **2014**, *12*, 449–455. [CrossRef] [PubMed]
25. Halary, S.; McInerney, J.O.; Lopez, P.; Bapteste, E. EGN: A wizard for construction of gene and genome similarity networks. *BMC Evol. Biol.* **2013**, *13*, 146. [CrossRef] [PubMed]
26. McInerney, J.O.; Pisani, D.; Bapteste, E.; O'Connell, M.J. The public goods hypothesis for the evolution of life on Earth. *Biol. Direct* **2011**, *6*, 41. [CrossRef] [PubMed]
27. Barbour, A.G.; Dai, Q.; Restrepo, B.I.; Stoenner, H.G.; Frank, S.A. Pathogen escape from host immunity by a genome program for antigenic variation. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 18290–18295. [CrossRef]
28. Chaconas, G.; Kobryn, K. Structure, function, and evolution of linear replicons in Borrelia. *Annu. Rev. Microbiol.* **2010**, *64*, 185–202. [CrossRef]
29. Corel, E.; Méheust, R.; Watson, A.K.; Mcinerney, J.O.; Lopez, P.; Bapteste, E. Bipartite network analysis of gene sharings in the microbial world. *Mol. Biol. Evol.* **2018**, *35*, 899–913. [CrossRef]
30. Sibbald, S.J.; Hopkins, J.F.; Filloramo, G.V.; Archibald, J.M. Ubiquitin fusion proteins in algae: Implications for cell biology and the spread of photosynthesis. *BMC Genomics* **2019**, *20*, 1–13. [CrossRef]
31. AM, M.; Hyland, E.M.; Cormican, P.; Moran, R.J.; Webb, A.E.; Lee, K.D.; Hernandez, J.; Prado-Martinez, J.; Creevey, C.J.; Aspden, J.L. Gene Fusions derived by transcriptional readthrough are Driven by Segmental Duplication in Human. *Genome Biol. Evol.* **2019**. [CrossRef]
32. Méheust, R.; Watson, A.K.; Lapointe, F.J.; Papke, R.T.; Lopez, P.; Bapteste, E. Hundreds of novel composite genes and chimeric genes with bacterial origins contributed to haloarchaeal evolution. *Genome Biol.* **2018**, *19*, 1–12. [CrossRef]

Appendix B:

Supplementary Materials for Chapter 3



(a)



(b)

Figure S3.1. Numbers of CRISPR repeats (a) and spacers (b) that identified by program MinCED, PILER-CR, CRISPRDetect, CRISPRCasFinder and in resample dataset.

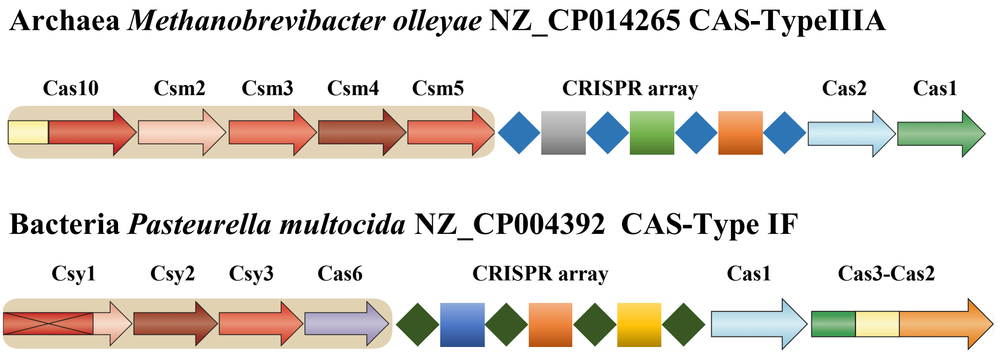
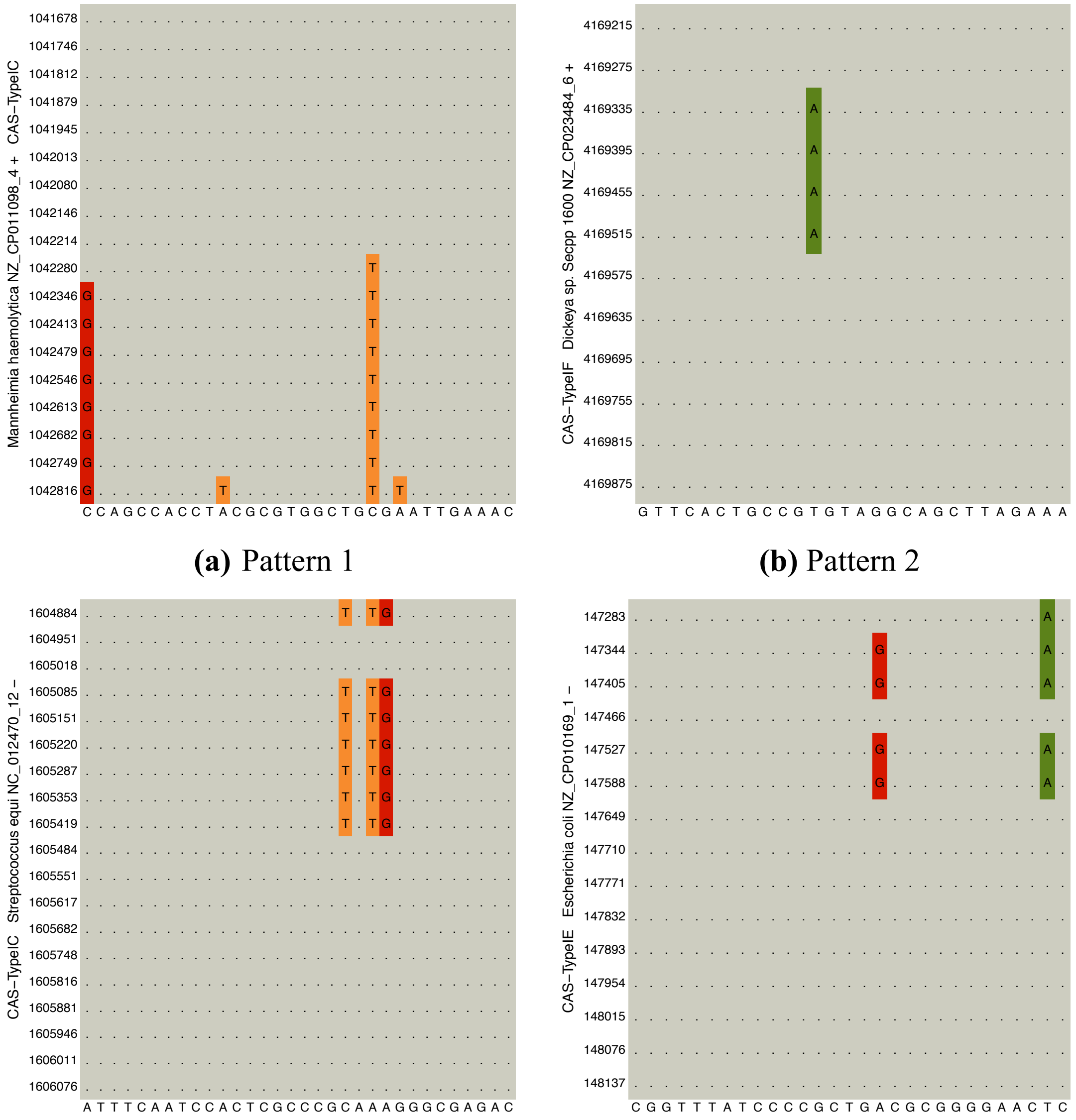

Archaea *Methanobrevibacter olleyae* NZ_CP014265 CAS-TypeIIIA
Cas10
Csm2
Csm3
Csm4
Csm5
CRISPR array
Cas2
Cas1
Bacteria *Pasteurella multocida* NZ_CP004392 CAS-Type IF
Csy1
Csy2
Csy3
Cas6
CRISPR array
Cas1
Cas3-Cas2

Figure S3.2. Sample of CRISPR array that situated within *cas* gene clusters. The *cas* gene modes adapted from Makarova et al. (2020).

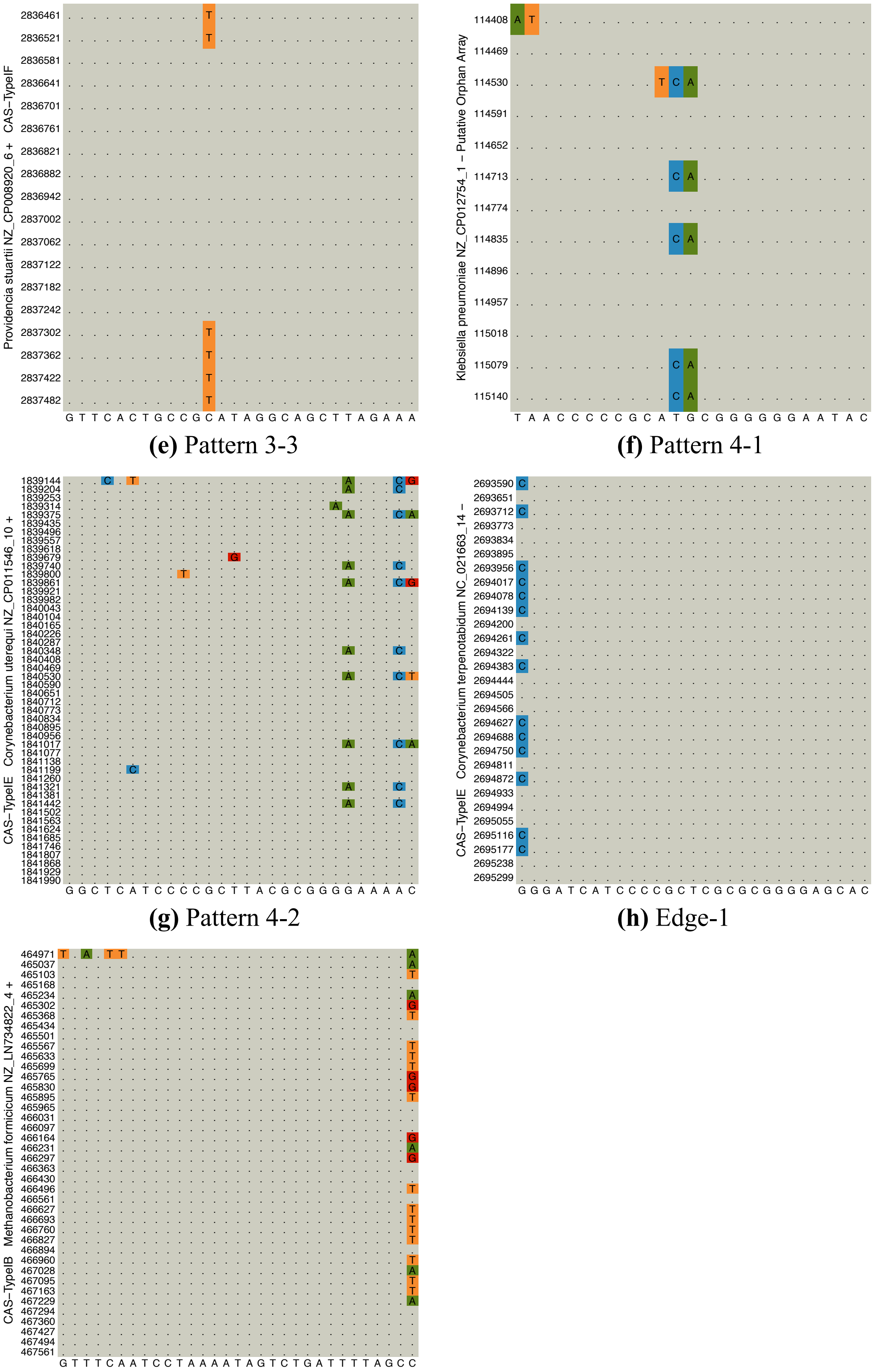


(a) Pattern 1

(b) Pattern 2

(c) Pattern 3-1

(d) Pattern 3-2



(e) Pattern 3-3

(f) Pattern 4-1

(g) Pattern 4-2

(h) Edge-1

(i) Edge-2

Figure S3.3 Full-size pictures of example repeat patterns. (a): Pattern 1; **(b)**: Pattern 2; **(c-e)**: Pattern 3; **(f-g)**: Pattern 4; **(h-i)**: Edge.

Table S3.1. Summary of species in CRISPR-Cas identification dataset.

Domain	Taxonomic Order	Species Number	Domain	Taxonomic Order	Species Number
Archaea	Acidilobales	3	Bacteria	Fibrobacterales	2
	Archaeoglobales	8		Fimbriimonadales	1
	Candidatus Nitrosocaldales	1		Flavobacteriales	270
	Desulfurococcales	12		Frankiales	5
	Halobacteriales	21		Fusobacteriales	49
	Haloferacales	15		Gemmatimonadales	3
	Methanobacteriales	24		Geodermatophilales	3
	Methanocellales	2		Gloeobacterales	2
	Methanococcales	19		Glycomycetales	1
	Methanomassiliicoccales	4		Halanaerobiales	7
	Methanomicrobiales	10		Holosporales	3
	Methanosarcinales	33		Hydrogenophilales	1
	Natrialbales	14		Ignavibacteriales	2
	Nitrosopumilales	3		Immundisolibacterales	1
	Nitrososphaerales	1		Kineosporiales	1
	Sulfolobales	41		Kiritimatiellales	1
	Thermococcales	39		Kosmotogales	3
	Thermoplasmatales	7		Lactobacillales	1054
	Thermoproteales	10		Legionellales	126
	NO RANK	10		Limnochordales	1
Bacteria	Acholeplasmatales	12		Magnetococcales	1
	Acidaminococcales	3		Marinilabiliales	4
	Acidiferrobacterales	3		Mariprofundales	2
	Acidimicrobiales	1		Methylacidiphilales	2
	Acidithiobacillales	7		Methylococcales	9
	Acidobacteriales	7		Micrococcales	157
	Acidothermales	1		Micromonosporales	19
	Actinomycetales	32		Mycoplasmatales	181
	Actinopolysporales	1		Myxococcales	22
	Aeromonadales	59		Nakamurellales	2
	Alteromonadales	148		Natranaerobiales	1
	Anaerolineales	4		Nautiliales	2
	Aquificales	14		Neisseriales	169
	Ardenticatenales	1		Nevskiales	2
	Bacillales	1349		Nitrosomonadales	27

Bacteria	Bacteriovoracales	4	Bacteria	Nitrospirales	9
	Bacteroidales	90		Nostocales	25
	Bacteroidetes Order II. Incertae sedis	12		Oceanospirillales	64
	Bdellovibrionales	6		Opitutales	3
	Bifidobacteriales	130		Orbales	2
	Brachyspirales	9		Oscillatoriales	14
	Bradymonadales	1		Parachlamydiales	4
	Bryobacterales	1		Parvularculales	1
	Burkholderiales	1002		Pasteurellales	206
	Caldilineales	1		Pelagibacterales	4
	Caldisericales	1		Petrotogales	4
	Calditrichales	1		Phycisphaerales	1
	Campylobacterales	406		Planctomycetales	13
	Candidatus Babeliales	1		Pleurocapsales	3
	Candidatus Brocadiales	1		Propionibacteriales	74
	Candidatus Nanopelagicales	16		Pseudomonadales	688
	Cardiobacteriales	2		Pseudonocardiales	35
	Catenulisporales	1		Puniceicoccales	1
	Caulobacterales	24		Rhizobiales	401
	Cellvibrionales	16		Rhodobacterales	137
	Chitinophagales	10		Rhodocyclales	20
	Chlamydiales	155		Rhodospirillales	89
	Chlorobiales	15		Rickettsiales	110
	Chloroflexales	5		Rubrobacterales	3
	Chromatiales	30		Salinisphaerales	1
	Chroococcales	13		Saprospirales	2
	Chroococcidiopsidales	1		Sedimentisphaerales	2
	Chrysiogenales	1		Selenomonadales	9
	Clostridiales	278		Solirubrobacterales	1
	Coprothermobacterales	1		Sphaerobacterales	1
	Coriobacteriales	10		Sphingobacteriales	18
	Corynebacteriales	588		Sphingomonadales	100
	Cytophagales	46		Spirochaetales	62
	Deferribacterales	4		Streptomycetales	138
	Dehalococcoidales	24		Streptosporangiales	11
	Deinococcales	19		Synechococcales	63
	Desulfarculales	1		Synergistales	5
	Desulfobacterales	10		Syntrophobacterales	4

Bacteria	Desulfovibrionales	23	Bacteria	Thermales	16
	Desulfurellales	2		Thermoanaerobacterales	36
	Desulfurobacteriales	2		Thermodesulfobacteriales	4
	Desulfuromonadales	19		Thermomicrobiales	1
	Dictyoglomales	2		Thermotogales	29
	Eggerthellales	11		Thiotrichales	99
	Elusimicrobiales	1		Tissierellales	11
	Endomicrobiales	2		Veillonellales	13
	Enterobacterales	2270		Verrucomicrobiales	22
	Entomoplasmatales	48		Vibrionales	186
	Erysipelotrichales	14		Xanthomonadales	226
	Euzebyales	1		NO RANK	104

Table S3.2. Score scheme of CRISPRDetect. Adapted from Biswas et al. (2016).

No.	Elements	Score	Calculation
1.	**Presence of either *cas1* or *cas2* genes in the genome is awarded**.	+1, or 0	This method is only applied when an annotation file (NCBI gbk or gbff file) is used as input. The annotation files are searched (term based) to create a list of all cas genes present in the genome. The scoring system awards the quality score with '+1' when annotation of either cas1 or cas2 genes are present in the input file.
2.	**Match to known repeat using a set of reference repeats from high confidence arrays**	+3, or 0	26 experimentally were used to verify representative repeats as reference and increased the set of known repeats by allowing up to 7 base mismatches. This extended set of around 400 repeat was used to predict a higher confidence set. Arrays were predicted then those with greater than 7 repeats and scores > 4 were used to predict a set of likely repeats. This file was converted in to a BLAST database and potential repeat searched against that with blastn-short which is optimised for short sequences. When a match is found, the array quality score is awarded '+3'. This file or the score can be modified in the commandline version.

3.	**Repeat has at least 23 bases and ATTGAAA (N) at the end**.	+3, or 0	Another feature adapted from the CRISPRDirection algorithm is the presence of motif ATTGAAA(N) at the 3' end of repeats. We observed that, this motif is an accurate indicator of the direction of transcription. In that paper we also observed that all the potential repeats that are >=23nt long containing this motif were genuine CRISPRs. Hence, we used this information to contribute to the quality score, and the quality score is awarded with '+3' when the repeats are >=23nt long and contains ATTGAAA(N) at the 3' end.
4.	**Overall repeat identity within an array**	0 to 1	The overall repeats identity score (S) is calculated using the following method **S= (average % identity of the repeats - 80)/20** The maximum possible positive score can be 1 (when all repeats are identical). However, the score will be negative, when the overall repeat identity is <80%.
5.	**The repeats in the array do not form one sequence similarity cluster**.	-1.5, or 0	The repeats are clustered using CD-HIT-EST if they form more than one cluster the quality score is penalized by '-1.5'.
6.	**Scoring the repeat lengths**	range -3 to +1.	In this method, frequent repeat length distribution was used. The relative score (S) for a repeat of length (L) is determined using the following rules: **S= 0.25 + L/H** [where, L>=23 and L =< 47; H is the most abundant repeat length for Bacteria or Archaea] **S= -0.25*(23 - L)** [where, L <23] **S= -0.25*(L - 47)** [where, L >47] The maximum negative score limit is set to -3, and maximum positive score limit is +1.

7.	**Scoring the spacer lengths**.	range -3 to +3	In this method, each spacer of an array is independently scored, and counted towards a final spacer length score. The individual spacer length score (S) for a spacer with length (L) within the range 28-48 are awarded a positive score using the formula: **S = 0.01 + N/H** [where, 27< L =<48] N= Total number of spacers of this length; H= Most abundant spacer length for Bacteria or Archaea Any spacer length outside this range is penalised by the following rule: **S=-0.10* (28 - L)** [where, L<28] **S=-0.10* (L -48)** [where, L>48] Finally, an average spacer score for the current array is calculated using ***Average score=Sum_of_scores/no_of_spacers*** The maximum negative score limit is -3 and maximum positive score limit is +1.
8.	**Overall spacer identity**	-3 to +1	In this method we test the sequence (dis)similarity among all the spacers. If the spacers are all near identical it is more likely to be a direct repeat, possibly a tandem repeat rather than a CRISPR array. If the spacers belong to a total number of clusters (C) with identity >=80%, the spacer identity score (S) for an array with number of spacers (N) is calculated using the following rule: **S= -3** [where, C =< integer (N/2);] **S= 0.20*C** [where, C > integer (N/2);] The positive score limit is +1.
9.	**Scoring total number of identical repeats**	0 to 1	Since longer arrays, and those with a greater number of identical repeats are more likely to be a true CRISPR, this scoring method uses both. If an array contains 'P' identical repeats out of the 'N' total number of repeats, then the score (S) is calculated using the following rule: **S= log (N) - log (N-P)** [where, P=Identical repeats, N= total number of repeats] The maximum positive score limit is +1.

Table S3.2. Evidence level rating system of CRISPRCasFinder (Couvin et al., 2018)**.**

Level	Requirement
1	CRISPR-like arrays having 3 spacers or less.
2	CRISPR arrays having an entropy-based conservation (EBcons) of repeats lower than 70.
3	CRISPR arrays having a EBcons of repeats greater or equal to 70, and a spacer conservation (BioPerl's overall percentage identity) greater than 8%,
4	CRISPR arrays having a EBcons of repeats greater or equal to 70, and a spacer conservation (BioPerl's overall percentage identity) lower than 8%.

Appendix C:

Supplementary Materials for Chapter 4

Table S4.1. The number of genes that co-occur or disassociate with different CRISPR-Cas subtypes. Gene families that show significant associations from chi-square tests were counted and then classified into positive and negative related groups by OR value.

CRISPR-Cas subtype	Number of genes that positively associated CRISPR-Cas	Number of genes that negatively associated CRISPR-Cas
Type I-A	784	0
Type I-B	1321	395
Type I-C	62	0
Type I-D	36	0
Type I-E	1641	0
Type I-F	1344	4
Type I-U	65	0
Type II-A	1204	0
Type II-B	88	0
Type II-U	1576	0
Type III-A	276	0
Type III-B	78	0
Type III-U	122	0
Type U	69	0

Table S4.2. Annotations of COG number in type III-A protein association network by STRING.

Node Colour	COG	Annotation
●	COG0066	*3-isopropylmalate dehydratase small subunit*
●	COG0491	*Glyoxylase or a related metal-dependent hydrolase, beta-lactamase superfamily II*
●	COG0535	*Radical SAM superfamily enzyme, MoaA/NifB/PqqE/SkfB family*
●	COG0602	*Organic radical activating enzyme*
●	COG0608	*Single-stranded DNA-specific exonuclease, DHH superfamily, may be involved in Archaeal DNA replication intiation*
●	COG0614	*ABC-type Fe3+-hydroxamate transport system, periplasmic component*
●	**COG0664**	***cAMP-binding domain of CRP or a regulatory subunit of cAMP-dependent protein kinases***
●	COG1011	*FMN phosphatase YigB, HAD superfamily*
●	COG1036	*Archaeal flavoprotein*
●	COG1048	*Aconitase A*
●	COG1180	*Pyruvate-formate lyase-activating enzyme*
●	**COG1337**	***CRISPR/Cas system CSM-associated protein Csm3, group 7 of RAMP superfamily***
●	**COG1353**	***CRISPR/Cas system-associated protein Cas10, large subunit of type III CRISPR-Cas systems, contains HD superfamily nuclease domain***
●	COG1355	*Predicted class III extradiol dioxygenase, MEMO1 family*
●	COG1499	*NMD protein affecting ribosome stability and mRNA decay*
●	COG1574	*Predicted amidohydrolase YtcJ*
●	COG1651	*Protein-disulfide isomerase*
●	COG1716	*Forkhead associated (FHA) domain, binds pSer, pThr, pTyr*
●	COG1752	*Predicted acylesterase/phospholipase RssA, containd patatin domain*
●	COG2078	*Uncharacterized conserved protein, AMMECR1 domain*
●	**COG2206**	***HD-GYP domain, c-di-GMP phosphodiesterase class II (or its inactivated variant)***

●	COG2207	*AraC-type DNA-binding domain and AraC-containing proteins*
●	COG2303	*Choline dehydrogenase or related flavoprotein*
●	COG2318	*Uncharacterized damage-inducible protein DinB (forms a four-helix bundle)*
●	COG2816	*NADH pyrophosphatase NudC, Nudix superfamily*
●	COG2896	*Molybdenum cofactor biosynthesis enzyme MoaA*
●	COG3861	*Stress response protein YsnF (function unknown)*
●	COG5317	*Uncharacterized protein*
●	NOG11049	*non supervised orthologous group*
●	NOG257039	*RAMP superfamily*
●	NOG299149	*Hypothetical protein DUF2513)*
●	NOG32667	*Protein of unknown function (DUF3108)*
●	NOG48198	*Domain of unknown function (DUF4148)*
●	NOG52950	*non supervised orthologous group*
●	NOG56581	*non supervised orthologous group*
●	NOG76658	*non supervised orthologous group*
●	NOG89043	*non supervised orthologous group*
●	arCOG01916	*RHH-fold DNA-binding ptotein*
●	arCOG03712	*HEPN domain containing protein*
●	COG1336	*CRISPR/Cas system CMR subunit Cmr4, Cas7 group, RAMP superfamily*
●	COG1604	*CRISPR/Cas system CMR subunit Cmr6, Cas7 group, RAMP superfamily*
●	COG3337	*CRISPR/Cas system CMR-associated protein Cmr5, small subunit*
●	COG1769	*CRISPR/Cas system CMR-associated protein Cmr3, group 5 of RAMP superfamily*
●	COG1367	*CRISPR/Cas system CMR-associated protein Cmr1, group 7 of RAMP superfamily*