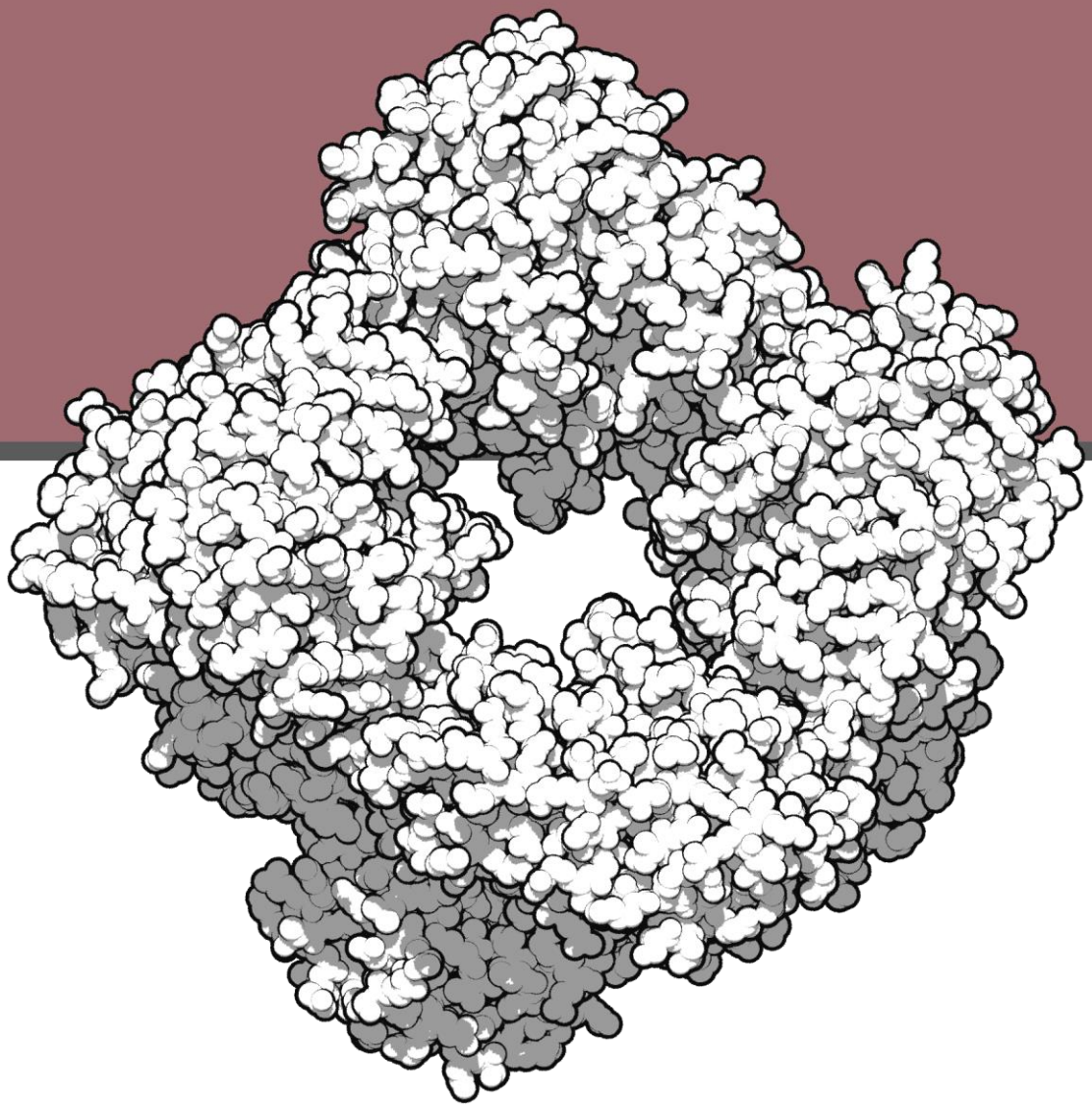
Analysis of **novel and unexplored** groups of prokaryotic **Reverse Transcriptases**

MARIO RODRÍGUEZ MESTRE



Máster Universitario en Biotecnología
Curso 2019/2020

UNIVERSIDAD DE GRANADA

eez Estación Experimental del Zaidín

CSIC Consejo Superior de Investigaciones Científicas

Cover figure: 3D structure of a Group II intron Reverse Transcriptase (PDB entry: 5HHL) from Eubacterium Rectale processed with QuteMol software

Analysis of novel and unexplored groups of prokaryotic Reverse Transcriptases

Mario Rodríguez Mestre[1]

Tutorizado por Francisco Martínez Abarca Pastor[1]

Memoria presentada por el graduado en Bioquímica

D. Mario Rodríguez Mestre como aspirante al título de

Máster en Biotecnología por la Universidad de Granada.

Fdo: Mario Rodríguez Mestre

VºBº Director:

Dr. Francisco Martínez-Abarca Pastor

Investigador Científico

[1]Estructura, Dinámica y Función de Genomas de Rizobacterias. Grupo de Ecología Genética de la Rizosfera. Departamento de Microbiología del Suelo y Sistemas, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas, C/ Profesor Albareda 1, 18008 Granada, Spain

Resumen

Pese a que las Reverso Transcriptasas (RTs) virales y eucariotas han sido ampliamente estudiadas, se conoce mucho menos acerca de las presentes en genomas procariotas. En la actualidad, solo se han caracterizado en profundidad tres tipos de retroelementos en bacterias: los intrones del Grupo II, los retroelementos generadores de diversidad (DGRs) y los retrones, además de otros grupos de RTs que han sido identificados pero aún no se han caracterizado. Estudios recientes han despertado interés en las Reverso Transcriptasas procariotas, ya que se han encontrado asociadas a sistemas CRISPR/Cas, y se ha hipotetizado que otras RTs no caracterizadas pueden haber sido domesticadas para llevar a cabo otras funciones de utilidad celular distintas a su papel en la movilidad, como en el caso de las asociadas a sistemas CRISPR. En este estudio, se ha llevado a cabo un profundo desarrollo y análisis bioinformático con el objetivo de revelar nuevas proteínas asociadas de RTs desconocidas e intentar vislumbrar cuál puede ser su papel en el funcionamiento de las células bacterianas.

Palabras clave: Retroelementos, Procariotas, Análisis Bioinformático, Filogenia, Genética

Abstract

Although viral and eukaryotic Reverse Transcriptases (RTs) have been extensively studied, very little is known about prokaryotic RTs. Until now, just three groups of bacterial retroelements have been deeply characterized: Group-II introns, Diversity Generating Retroelements (DGRs) and retrons. Other groups of RTs have been identified but remain uncharacterized. Recent studies have raised interest in prokaryotic RTs because some of them have been found to be associated with CRISPR/Cas systems. It has been hypothesized that non-characterized RTs could have been domensticated in order to perform useful cellular function, as happens with those associated with CRISPR. In this study, we develop a profound bioinformatics analysis with the objective of reveal novel associations of unknown RTs and try to shed light into their role in bacterial cellular processes.

Keywords: Retroelements, Prokaryotes, Bioinformatic Analysis, Phylogeny, Genetics

Index

1. Introduction

1.1 Reverse Transcriptases

Discovered in 1970 by Temin & Baltimore in tumor viruses (Baltimore 1970; Temin and Mizutani 1970; Coffin and Fan 2016), Reverse Transcriptases (RTs, also known as "RNA-dependent DNA polymerases") are enzymes capable of polymerizing cDNA from RNA. Even though they are present in a wide variety of genomes, RTs are a key component of mobile genetic retroelements (i.e. genetic elements containing an RT coding gene) and retrotransposons, which are retroelements capable of invading genomes and propagate through an intermediary RNA phase and the ulterior insertion of the resulting cDNA on novel localizations (Finnegan 2012).

Reverse transcription is a crucial factor in the replication of *Retroviridae, Metaviridae, Pseudoviridae, Hepadnaviridae, and Caulimoviridae*, and it has been found that RTs encoded in their genomes are phylogenetically related to those found in mobile genetic elements of prokaryotes and eukaryotes (Menéndez-Arias, Sebastián-Martín, and Álvarez 2017).

In eukaryotes, retroelements are abundant and comprise a vast extension of their genome, acting as a critical factor in their evolution and expression. In *Homo sapiens*, for example, they constitute approximately 45% of the total genetic material and 90% of all transposable elements (Bannert and Kurth 2004). Nonetheless, retroelements in prokaryotes don't imply such an impact when it comes to the proportion, representing less than 1% (Simon and Zimmerly 2008). Eukaryotic retroelements can be divided into two major groups based on the presence/absence of long terminal repeats (LTRs). The first group, containing LTRs, is represented by LTR retrotransposons, tyrosine recombinase retrotransposons, and endogenous retroviruses. The second, termed non-LTR retroelements, comprise long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), Penelope-Like Elements (PLEs), Telomerases (Bernardes and Blasco 2013) and processed pseudogenes (Gogvadze and Buzdin 2009).

Prokaryotic RTs were not discovered until 1989 when they were found to be part of retroelements known as retrons (Lampson, Inouye, and Inouye 1989; Lim and Maas 1989;

Inouye 2017). Later on, it has been demonstrated that RTs are not only present on retrons, but also in the so-called DGRs ("Diversity-Generating Retroelements")(Doulatov et al. 2004), Abi systems ("Abortive phage Infection")(Fortier, Bouchard, and Moineau 2005; Odegrip, Nilsson, and Haggård-Ljungquist 2006; Durmaz and Klaenhammer 2007) and Group II introns, which are the most abundant (~50% of described RTs), the better characterized and the only ones with demonstrated autonomous mobility (Toro and Nisa-Martínez 2014; Michel and Ferat 1995; Dai and Zimmerly 2003; Toro 2003; Lambowitz and Zimmerly 2004). In addition, putative RTs phylogenetically related to those encoded by mobile Group II introns and Retrons have been found to be associated with CRISPR-Cas systems, adjacent or fused at the C-terminus to Cas1 or PriS (Kojima and Kanehisa 2008; Silas et al. 2016; Silas et al. 2017; Toro, Martínez-Abarca, and González-Delgado 2017; Mohr et al. 2018; Toro et al. 2018; Schmidt, Cherepkova, and Platt 2018; Toro et al. 2019).

Reverse Transcriptases share a typical domain organization that involves seven peptide regions (domains 1-7) identified by the reconstruction of RT phylogenies and a comparison of the genetic organization of the various RT (**see Figure 1**) (Xiong and Eickbush 1990). In addition, it has been found that non-LTR RTs share a subdomain 0 that can be considered an N-terminal extension of the RT domain (Malik, Burke, and Eickbush 1999).

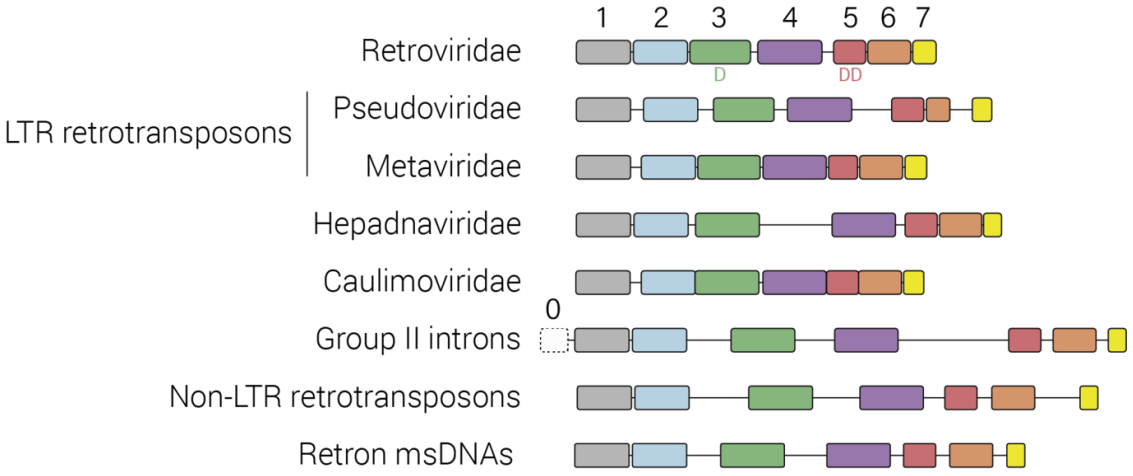


Figure 1. *Domain organization of the different DNA polymerase domains of RTs. Position of conserved peptidic motifs (1-7, defined by (Xiong and Eickbush 1990)) are indicated by colored squares. Catalytic aspartic residues (D) are indicated in motifs 3 and 5. Domain 0 found in Group II introns is indicated in dotted lines. Adapted from (Menéndez-Arias, Sebastián-Martín, and Álvarez 2017)*

The most exhaustive phylogenetic analysis of prokaryotic Reverse Transcriptases revealed that most of the RTs are group-II introns (47%), Retron/retron-like sequences (25%), and DGRs (12%). The remaining 16% clustered into distinct groups including RTs previously reported being linked to type III CRISPR-Cas systems or other uncharacterized RTs, such as G2L, Abi-like or UG (Unknowns) groups (**see Figure 2**). Further information regarding the different groups of RTs is described in the next sections.

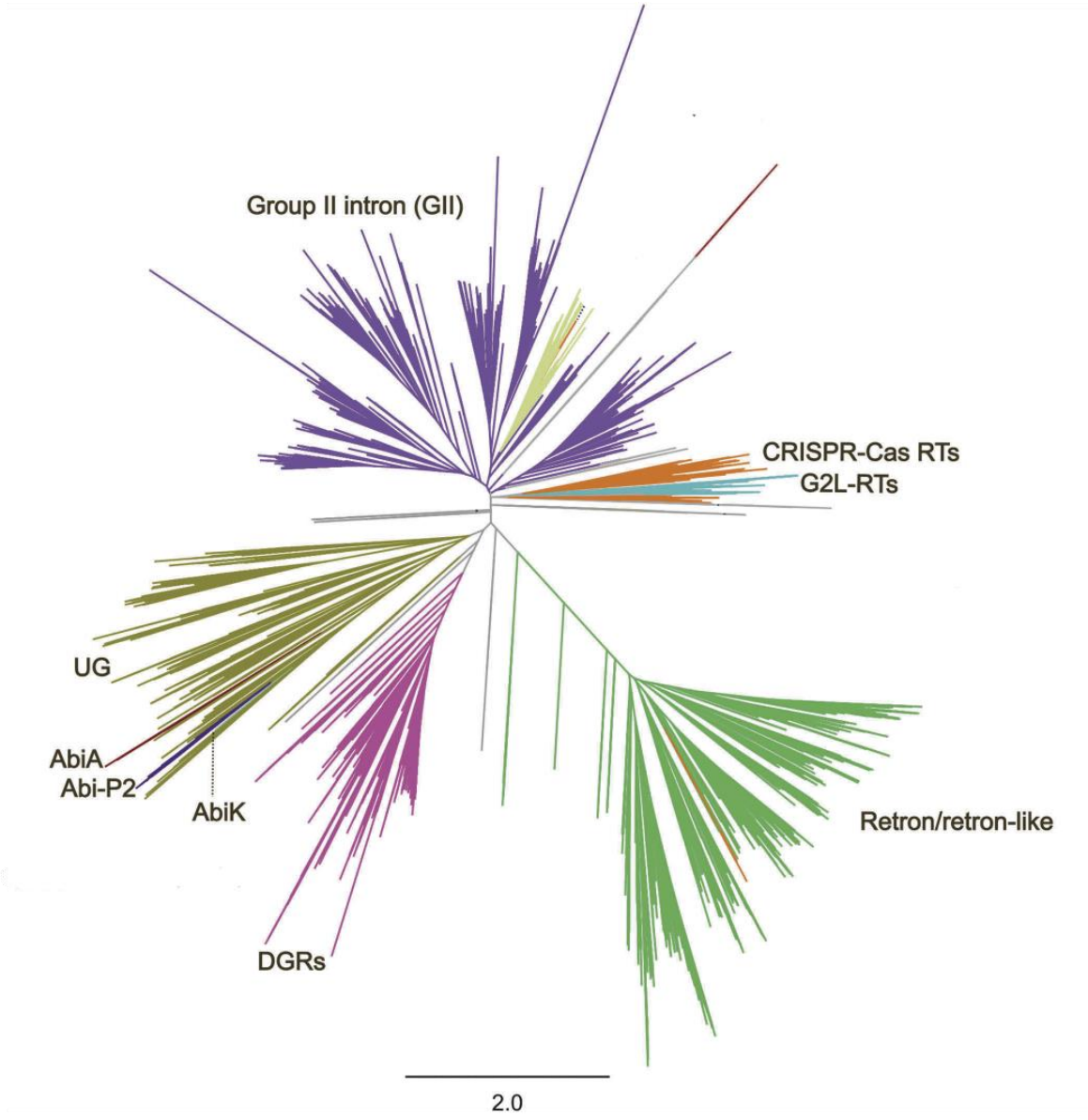


Figure 2. *Phylogeny of prokaryotic RTs. The unrooted tree was constructed from an alignment of 9,141 unique predicted RT protein sequences obtained with the FastTree program. The branches corresponding to group-II introns (GII), GII class F, Retron/retron-like, DGRs, CRISPR-Cas, G2L, Abi, and UG RTs are indicated and highlighted with distinct colors.*

1.1.1 Group-II Introns

Group II introns are the best characterized bacterial retroelement, the most abundant and the best understood. They were first identified in the mitochondrial and chloroplast genomes, and have subsequently been described in bacteria and archaea (Michel and Ferat 1995). They consist of a ~500bp to ~800bp sequence that acts as an autocatalytic, self-splicing intron RNA (ribozyme), and an encoded ~1.0 to ~1.5kb ORF that is translated into an intron-encoded protein (IEP) (Michel and Ferat 1995; Lambowitz and Zimmerly 2004; Toro, Jiménez-Zurdo, and García-Rodríguez 2007; Zimmerly and Wu 2015). RNA folds into a conserved three-dimensional structure organized into six double-helical domains, DI to DVI (San and Lambowitz 2002; Lambowitz and Zimmerly 2004). Most bacterial group II introns have an open reading frame (ORF) encoding an intron-encoded protein (IEP) in DIV. This IEP consists of an RT followed by a putative RNA-binding domain with RNA splicing or maturase activity (the X domain), and, in some intron lineages, a C-terminal DNA-binding and endonuclease domain (Lambowitz and Zimmerly 2004). Group II introns are absent from the nuclear genomes of eukaryotes but are thought to be the predecessor of the spliceosome and to have played an essential role in the evolution of several features of eukaryotic cell organization (Xiong and Eickbush 1990).

An essential property of group II introns in bacteria is that they behave primarily as retroelements rather than introns. Their selfish nature is evident in several ways. Firstly, the introns are generally excluded from housekeeping or conserved genes, suggesting that they inhibit gene expression in some way. Second, over half of intron copies in bacteria are truncated and often co-located with other mobile DNAs, suggesting a nomad rather than stable lifestyle (Zimmerly and Semper 2015). In contrast, group II introns from mitochondrial and chloroplast genomes are located in housekeeping genes, and many are nonmobile splicing units. Third, the distribution of group II introns across bacterial species and strains is patchy. For example, related strains of *Escherichia coli* can harbor from one to 15 copies of group II intron, and some genomes contain over 20 identical copies of an intron (Martínez-Abarca, Zekri, and Toro 1998).

The overall process of the mobility of group II introns occurs through the mechanism of target-primed reverse transcription (**see Figure 5**) (TPRT). In the TPRT mechanism participates a Ribonucleoprotein complex (Spliced RNA and Intron Encoded Protein)

which initiates the reverse splicing of the lariat RNA into a double-stranded DNA target, followed by cleavage of the bottom strand by the En domain of the IEP to form a primer, and finally reverse transcription of the inserted intron. The last steps of the integration are commonly carried out by host repair machinery that generates the double-stranded (Lambowitz and Zimmerly 2004).

Some IEPs lack the En domain and require an alternative primer for bottom strand synthesis. This has been shown to be a nascent DNA strand provided by a replication fork (Martínez-Abarca et al. 2004). Notably, group II introns are by far the most numerous of RT types in bacteria, consistent with their robust retromobility (Toro and Nisa-Martínez 2014).

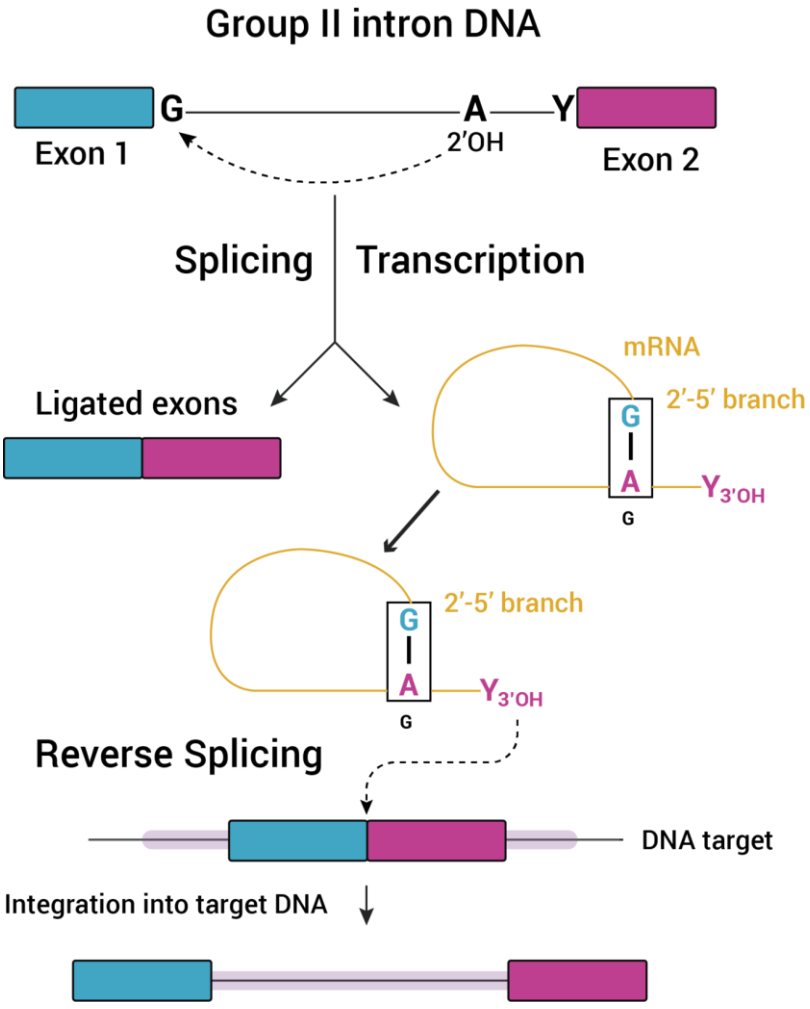


Figure 5. *Mechanism of retrohoming of Group II introns. Firstly, the Group II intron DNA is transcribed, a process in which the intron is spliced out and the two exons remain ligated. Secondly, the spliced RNA intron performs a nucleophilic attack on the DNA target, thus introducing the intron and separating the two exons.*

1.1.2 Retrons

Retrons are the fist prokaryotic retroelement to be characterized, and consist of ~2000 bp that encodes an RT-coding gene (named *ret*) and contiguous inverted repeat sequences (named *msr* and *msd*). The *ret*, *msr*, and *msd* genes are transcribed as a single mRNA which is folded into a specific secondary structure. Once translated, the RT binds to the inverted repeats present in the mRNA and starts reverse transcription of the RNA template-assisted by the 2'OH group present in a specific branching G residue that acts as a primer. Thereby, the resulting DNA, named *msDNA* (after *multicopy single-stranded DNA*), remains covalently attached to its RNA template by a 2'-5' phosphodiester bond and base-pairing attachment of their respective 3' ends **(see Figure 3)** (Doulatov et al. 2004; Lampson, Inouye, and Inouye 2005; Xie and Yang 2017).

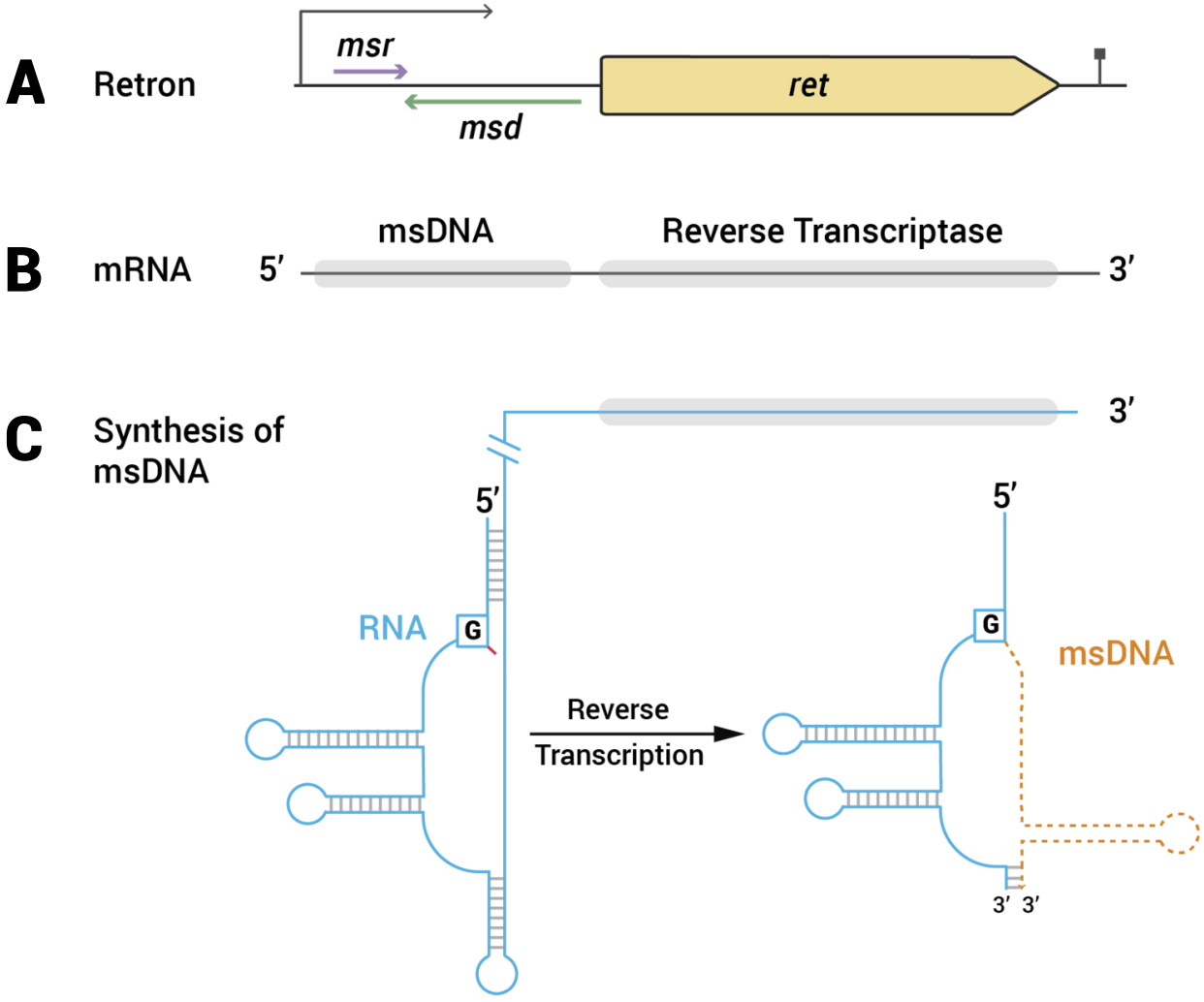


Figure 3. *Typical architecture of a Retron and synthesis of msDNA. (A) Typycal locus architecture of DGRs. (B) Transcribed mRNA of Retron. (C) Synthesis of msDNA. The G branching nucleotide is highlighted with a square, and the 2'OH is highlighted in red*

Despite a lot of effort done to elucidate the function of retrons and/or msDNA, its biological role remains still unclear, and until know, retrons has been described as non-essential, as their lack or presence in natural conditions do not seem to generate relevant phenotypic changes. However, recent advances have described two different msDNA generated by *Escherichia coli* retrons as substrates of Exonuclease VII (Jung et al. 2015). In this work, it has been hypothesized that high concentrations of msDNA could block the apoptotic-like cell death orchestrated by Exonuclease VII subunits. In addition, work linked retrons or msDNA to Dam methylase-like activity based on the similitudes between Dam methylase retron-EC67 (Hsu, Inouye, and Inouye 1990).

However, different hypotheses have been made in relation to other experiments performed. Maas et al. described that msDNA (from retrons Ec83 and Ec86) with non-perfectly matched base pairs in their stem-loop structures are mutagenic when present in high copy numbers by titrating out MutS, a protein involved in DNA repair (Maas et al. 1994; Maas et al. 1996). These authors hypothesize that, as RT alone is not mutagenic, it could be linked to starvation, and it may enhance mutation rates by uprising the concentrations of msDNA. In addition, RT may compete with MutS in binding to msDNA, thereby controlling the mutation rate. This property could increase the mutation rate and provide advantages in bacterial adaptation (Zimmerly and Wu 2015).

1.1.3 Diversity Generating Retroelements

DGRs are another type of retroelements that are present in bacteria, archaea, phages, and plasmids. The most studied DGR, discovered in *Bordetella* phage *BPP-1 (Doulatov et al. 2004)*, consist of four adjacents genes: the target gene, named ***mtd***, that contains a variable repeat (**VR**), the reverse transcriptase *brt*, an accessory variability determinant (***avd***) gene, and a template repeat (**TR**). The proteins encoded by *brt* and *avd* genes generate a highly variable cDNA from the TR that is then inserted within the target protein, enabling the introduction of manifold sequence mutations. This process, termed "mutagenic retrohoming," yields a VR that is distinct from the TR exclusively at adenine bases thanks to an A-to-N error-prone reverse transcription **(see Figure 4)** (Handa et al. 2018).

The biological role of DGRs has been widely studied based on different methodologies. So far, the most promising strategy comes from large-scale genomic analysis (Ye 2014; Paul et al. 2017; Wu et al. 2018; Sharifi and Ye 2019). but the majority of them remain still uncharacterized, and more biological experiments need to be done.

To illustrate the process, the example of tropism switching in phage BPP-1 has been widely used. *Bordetella* (named after Jules Bordet) is a genus of gram-negative obligate parasite bacteria that can infect humans. *Bordetella* colonizes the ciliated respiratory mucosa, a surface designed to eliminate foreign particles, thereby making the adherence and persistence mechanisms of these bacteria crucial. They display a series of adhesins proteins at their cell surface in order to remain attached to the respiratory epithelium, and the composition of the cellular surface changes as its infectious cycle goes on (Mattoo et al. 2001). Phage BPP-1 adheres to *Bordetella* through Mtd, a protein located at the phage tail fibers. As the cellular surface changes its composition, the BPP-1 phage needs to change its tropism in order to adapt to those modifications, and BPP-1 phage is capable of doing this thanks to the DGR mechanism, which directs Mtd mutagenesis due to the mutagenic retrohoming process.

The BPP-1 TR sequence contains 23 aa's that can be mutagenized, which corresponds to 1013 aa' sequences in the Mtd protein (Wu et al. 2018). As a result, the BPP-1 tail fiber that mediates adsorption to bacterial host receptors is hypervariable, enabling tropism switching on host *Bordetella* species.

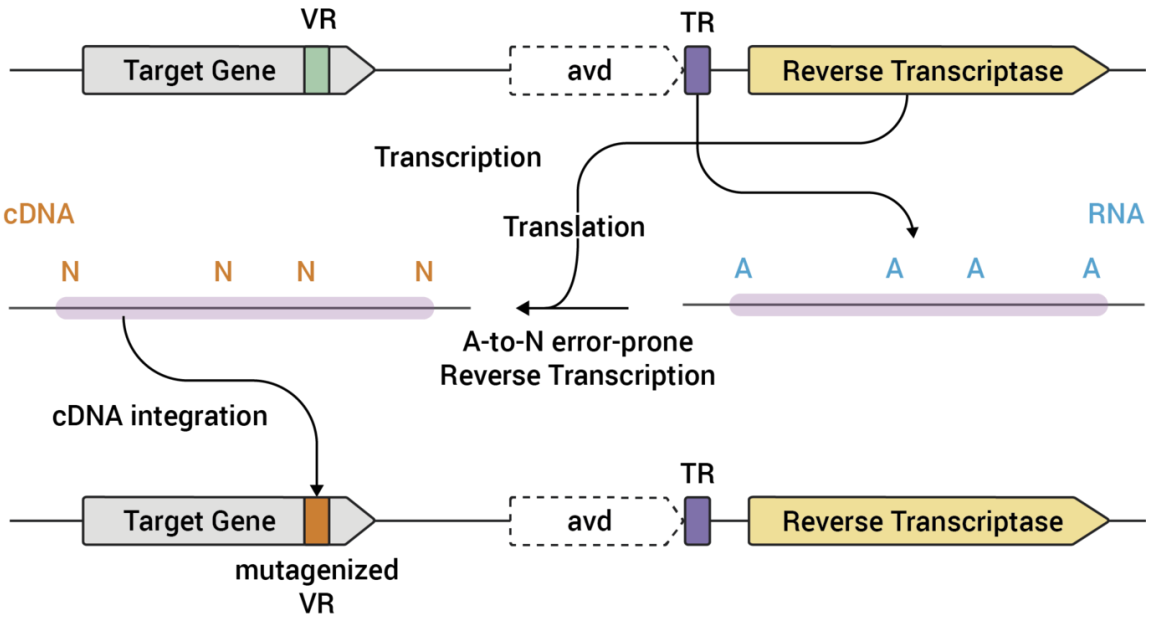


Figure 4. *Schematic process of mutagenic retrohoming performed by Diversity Generating Retroelements.*

Recent analyses shed more light into the DGRs biological role, and it has been discovered that DGRs can not only diversify C-type lectin fold domain-containing proteins but also

Ig-like domain-containing proteins. They also revealed that in some cases accessory proteins like HRDC (helicase and RNaseD C-terminal) are present (Paul et al. 2017; Wu et al. 2018). These analyses also revealed that prototypical DGR architecture is only a possibility among highly diverse structures, and sometimes DGR can lack the *avd* gene and have multiple target genes (Wu et al. 2018).

1.1.4 Abi systems

Abi RTs are another type of RTs involved in phage resistance. Until now, three different types of Abi-RT systems have been described: AbiK, AbiA, and Abi-P2. AbiA and AbiK mediate abortive infection (Abi), a process that occurs when phage infects bacteria, but the infection cycle and the multiplication of the virus are blocked. Another kind of Abi RT is Abi-P2, which is found in the P2 prophage of some *E. coli* strains. In this case, Abi-P2 mediates bacteriophage exclusion, which in contrast to Abi, prevents new phage infections of those cells that carry this system. While AbiK has been demonstrated to provide phage resistance on its own, AbiA and Abi-P2 are thought to require some accessory proteins or nucleic acids, as the genetic loci conferring resistance are about 3kb (Zimmerly and Wu 2015).

AbiK

The abiK gene encodes a ~600aa protein whose half N-terminal domain contains the prototypical RT motifs, and the C-terminus comprises a ~275aa region that contains the putative thumb domain of the polymerase, as well as an additional sequence with no identified domains. The C-terminal region and the RT motifs have an essential function because mutations disrupt or abolish abortive infection (Emond et al. 1997; Fortier, Bouchard, and Moineau 2005; Wang et al. 2011).

Heterologous expression of AbiK in *E. coli* revealed that it is a random and untemplated DNA polymerase DNA remains covalently attached to the protein (Wang et al. 2011). Its priming properties reside in the OH of a tyrosine residue, resembling RTs from hepadnavirus. However, it remains unclear whether AbiK exhibits actual RT activity under conditions not tested. It has been suggested that AbiK could block phage multiplication by interacting with the replication machinery (Zimmerly and Wu 2015).

AbiA and Abi-P2

The *abiA* gene encodes a protein with a similar size to AbiK (~600aa) with a C-terminal extension lacking detectable protein motifs. It also protects against the major classes of lactococcal phages and is similarly inferred to act at the stage of phage DNA replication, because DNA products do not accumulate in infected bacteria (Hill, Miller, and Klaenhammer 1990; Tangney and Fitzgerald 2002; Zimmerly and Wu 2015). The *abi-P2* gene is found in two P2 prophages in *E. coli* strains ECOR30 and ECOR58. The two *abi-P2* genes share ~75% identity and It has been demonstrated that *E. coli* strain BL21(DE3) overexpressing this gene (ORF570 of ECOR30) is resistant at levels of 10–7 relative to infection by T5 phage. Moreover, the expressed ORF570 was reported to have RT activity in an in vitro biochemical assay (Wang et al. 2011).

As it has been shown, AbiK, AbiA, and Abi-P2 posses similar properties regarding to resistance against bacteriophages. Although they can be mechanistically related, this cannot be automatically assumed because their RT phylogeny is distant and they differ in sequence motifs. To test this hypothesis, more studies are required with homologs of AbiK, AbiA, and Abi-P2 found in a vast range of species across multiple eubacterial phyla, including in clinical isolates. Despite having passed more than 15 years since the first description of Abi-RT systems, phage resistance conferred by RT-related proteins‘is not yet understood (Zimmerly and Wu 2015).

1.1.5 CRISPR-associated RTs

CRISPR/Cas (**C**lustered **R**epeated **I**nterspaced **S**hort **P**alindromic **R**epeats / **C**RISPR-**A**ssociated **P**roteins, respectively) systems are a naturally occurring type of prokaryotic defense systems that provide immunity against foreign genetic elements such as bacteriophages, transposable elements or conjugative plasmids. Its mechanism relies on the acquisition of a genetic memory from previous infections, through which CRISPR/Cas systems are capable of identify the foreign DNA or RNA sequences upon re-infection and mediate it clearing thanks to the action of RNA guided nucleases like Cas9 or Cas13 (Barrangou et al. 2007; Jackson et al. 2017; Hille et al. 2018).

These systems are characterized for possessing a CRISPR array composed of direct repeats (“DRs”) regularly interspaced by the so-called spacers, which represent past expositions to foreign nucleic acids. The difference between CRISPR/Cas and any other prokaryotic defense systems is its adaptative nature, which allows the host to change and

adapt to novel and constantly evolving biological threats. In this way, CRISPR/Cas-mediated defense is able to recognize, cleave and store foreign nucleic acids to prevent future invasions, which resembles adaptive immunity in eukaryotes (Jackson et al. 2017; Hille et al. 2018).

CRISPR/Cas loci possess a defined structure, in which an AT-rich leader located upstream of the CRISPR array contains the promoter sequence that allows its transcription. In the neighborhood of the array, it is common to find Cas proteins that perform the different actions needed for the immunity to be active. These actions are grouped into three different stages: adaptation, expression, and interference (**Figure 6**) (Jackson et al. 2017; Hille et al. 2018).

In the adaptive stage (i), the acquisition module constituted by Cas1 and Cas2 proteins (and other auxiliary Cas proteins) makes possible the processing and integration of foreign nucleic acids (termed protospacers) into the CRISPR array allowing the formation of a new spacer. In the expression stage (ii), the CRISPR array is transcribed and processed to form crRNAs (CRISPR RNAs), which associate to the effector complex and guide the recognition and cleavage of complementary nucleic acids during the interference stage (iii) **(see Figure 6)** (Jackson et al. 2017; Hille et al. 2018).

While the acquisition module is widely conserved among prokaryotic species, the effector modules are highly variable and define the differences of each type of CRISPR/Cas system. Based on effector proteins and other signatures, CRISPR/Cas systems can be classified into Class 1 systems (types I, III and IV) and Class 2 systems (types II, V and VI) (Makarova and Koonin 2015; Makarova, Wolf, and Koonin 2018).

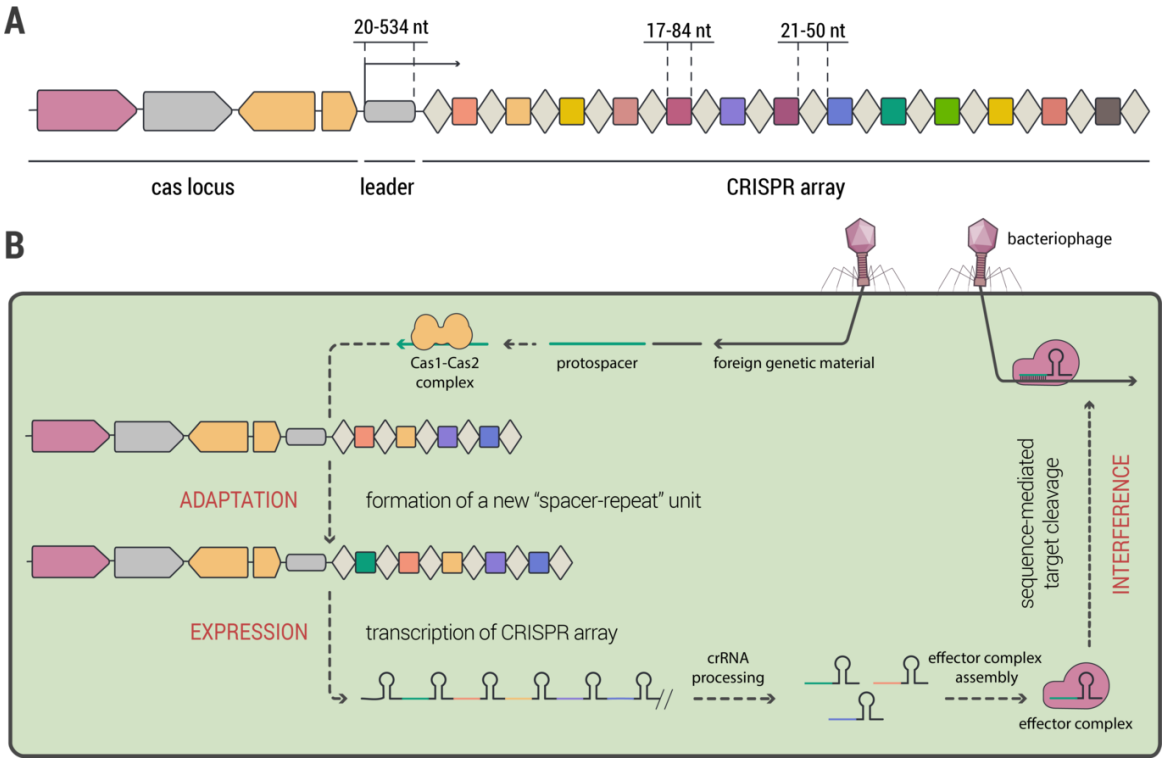


Figure 6. *(A) Typical architecture of a CRISPR/Cas locus. (B) Summary of the three steps of CRISPR/Cas immunity: Adaptation, Expression and Interference. Acquisition module is highlighted in yellow, and interference complex is highlighted in red.*

Recently, it has been found RTs can be associated with CRISPR/Cas systems, although its role is not yet fully understood. Nonetheless, It has been demonstrated that RT-containing systems are able of acquiring spacers from RNA molecules thanks to the reverse transcriptase activity provided by RTs (Silas et al. 2016; Silas et al. 2017; Mohr et al. 2018; Schmidt, Cherepkova, and Platt 2018; González-Delgado et al. 2019). The majority of the CRISPR-associated RTs derive from Group II introns and are commonly associated with Type III systems, but It has been recently shown that these proteins could be associated with Type I and VI and can also originate from Retron-like and Abi-like RTs (Toro et al. 2019).

The way in which RTs associated with CRISPR/Cas systems can be flexible and has varied during the different stages of its evolution (**see Figure 7**). RTs can be found fused to Cas1 or PriS and even merged to Cas6 and Cas1 at the same time (Toro et al. 2019). These associations represent a promising field of study that already has led to the development of a technique called Record-seq that allows recording transcriptional events (Schmidt, Cherepkova, and Platt 2018).

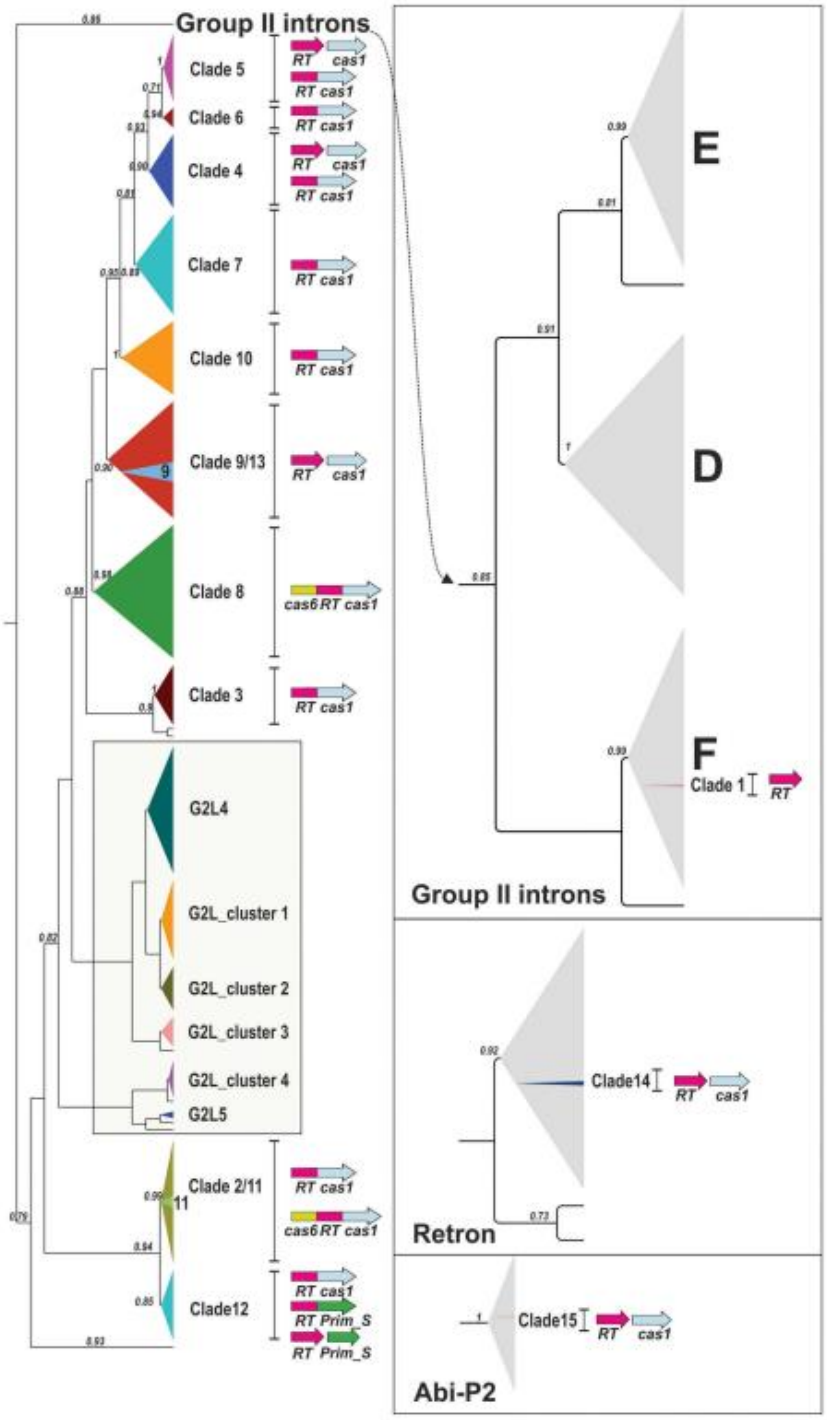


Figure 7. *Phylogeny of CRISPR-Cas encoded RTs. The identified lineages of CRISPR-Cas RTs, three evolving from group-II introns, one from Retron/retron-like and one from Abi-P2 RTs, are shown. The CRISPR-Cas RTs and neighboring group-II intron classes (F, D, and E); G2L; Retron and Abi-P2 clades are depicted schematically, with collapsed. For the CRISPR-Cas RT clades, the most common RT domains of gene organizations are indicated. Prim_S indicates an archaeo-eukaryotic primase AE_Prim_S-like domain*

1.1.6 Unknown Groups (UG)

Although some groups of prokaryotic RTs such as Group II intron or DGRs have been extensively studied, some others remain uncharacterized and very little is known about them. Nonetheless, It has been hypothesized that they could derive from known retroelements and, after being domesticated, remained by providing evolutionary advantages to the host. Previous studies have tried to shed light on these groups (Kojima and Kanehisa 2008; Zimmerly and Wu 2015), but they only characterized a few of them. Nonetheless, RTs have been described to play an essential role in bacterial genome evolution, and many species encode several RT genes within their genomes. In this way, the characterization of these groups could contribute to increasing significantly the understanding of prokaryotic forms of life (Zimmerly and Wu 2015). Some may hypothesize that they could represent selfish genetic elements as happens with eukaryotic retroelements, but Zimmerly & Wu pointed out that most of these putative RTs do not show signs of being mobile, as they do not accumulate to multiple copies in genomes, and If the RTs are, in fact, part of mobile DNAs, the mobility levels would have to be much lower than group II introns or most other mobile DNAs. some RTs remain uncharacterized or appear to have been domesticated, and It has been suggested that they could perform useful cellular functions.

Of these groups, some retron RTs have been found to possess a Peptidase domain at the C-terminus. Unknown groups (UG) 1 and 5 have an appended C-terminal nitrilase, and UG4 RTs have a pilus-related domain. In UG10, the protein posses a primase domain that could suggest concerted priming and reverse transcription reaction. Finally, UG3 and UG8 have been found adjacent to each other, suggesting that they operate together (Kojima and Kanehisa 2008; Zimmerly and Wu 2015).

1.2 Prediction of protein association

In the last years, techniques such as co-purification, co-immunoprecipitation, yeast two-hybrid, and protein microarrays have been developed to analyze protein-protein interactions (Shoemaker and Panchenko 2007). However, these methods can be highly expensive and time-consuming, and their results can entail many false positives/negatives. Besides, the potential for two proteins to interact is not only specified by the physical and structural properties of their structures, but it is also encoded at a genomic level. For example, genes that interact are can be usually co-expressed (both temporally and spatially), and thus the information of two proteins physically interacting is meaningless unless they are present in the same part of the cell at the same time (Vella et al. 2017).

For this reason, many computational approaches have been developed to predict protein-protein interactions (PPI). Next-generation sequencing has provided a large amount of genomic data that can be used for these analyses, and many of these computational approaches also take advantage of high-throughput experimental information such as gene-expression data, cellular locality, and molecular complex information (Vella et al. 2017).

Mainly, computational methods use the structural, genomic, and biological context of proteins and genes to predict protein interaction and functional connections between proteins. Some of these approaches use sequence data alone to predict interactions (Wang et al. 2017), while others combine multiple computational and experimental datasets intending to build accurate protein interaction maps (Szklarczyk et al. 2019). These hybrid computational approaches exploit both the genomic and biological context of genes and proteins to predict interactions.

In the next sections, we will describe different methods developed to explore the associations of protein pairs or putative functional systems.

1.2.1 Structural-based approaches

Unlike other methods, structural approaches are limited by the availability of three-dimensional protein data in databases such as the Protein Data Bank (PDB), and only a small proportion of protein sequences have accurate structures described. However, structural approaches allow for high-resolution analysis of protein interactions, as they

cannot only predict whether two proteins interact but also the physical properties of the given interaction (Hayashi et al. 2018).

Often, residues that form an interaction surface are tightly packed and tend to be hydrophobic, and analysis of these characteristics can be used to predict interaction sites with up to a 60% success rate. Further studies using a six-parameter analysis regarding solvation potential, residue interface potential, hydrophobicity, planarity, protrusion, and accessible surface area yielded accurate predictions for 66% of 59 structures (Murakami and Jones 2006).

Apart from these aspects, some other downstream analyses of predicted protein pairs require high expertise such as shape complementarity (primarily used in docking studies which focus on finding the best fit of the two interacting proteins) and electrostatic complementarity between interfaces (Fernández-Recio, Totrov, and Abagyan 2004).

1.2.2 Evolutionary Context Approaches

1.2.2.1 Co-localization

One of the most used methods to predict interactions between proteins comes from the study of the genomic context of protein-coding genes. This approach is based on the hypothesis that proteins that are functionally or physically associated are located together to each other in the genome. This is very frequent in bacterial and archaeal operons, where genes that act as a functional system are transcribed from the same promoter in a polycistronic way. Thus, it is common to find proteins involved in the same function to be encoded in the same mRNA. In addition, operons encoding for co-regulated genes are usually conserved **(see Figure 8A)** (Lemay et al. 2012).

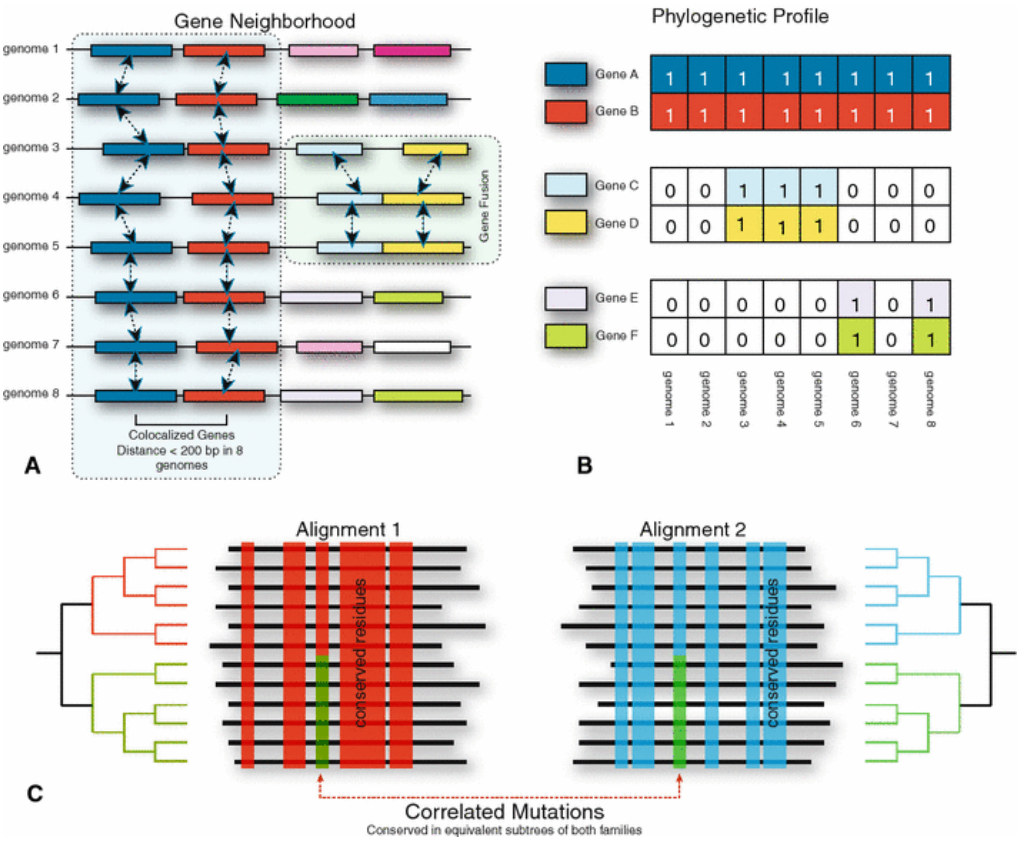


Figure 8*. Overview of evolutionary context approaches. (**A**) Gene neighborhood approach shows co-located genes (blue and red). Gene fusion is represented for two proteins (pale blue and yellow). (**B**) Phylogenetic profile approach (**C**) Mirrortree/matrixmatchmaker/i2h approach. Conserved regions are highlighted in red and blue. Correlated mutations (shown in green) are present in two identical sub-trees for each family (Shaktay et al. 2015)*

1.2.2.2 Phylogenetic Profiles

Another method derived from genomic context analysis is the co-occurrence of pairs of genes across multiple genomes. Two or more genes adjacent across many different genomes represent an evolutionary scenario that suggests that the products of these genes can be functionally or structurally related. In this case, the stringency is lower than previously described methods such as gene co-localization, where genes need to co-occur and be adjacent (Pellegrini 2012).

This type of phylogenetic analysis has been termed *phylogenetic profiling.* Typically, it is represented as a binary matrix of the presence/absence of a gene across multiple

genomes that needs to be scanned to find genes that exhibit similar patterns. Pairs of genes detected in this manner are candidates for physical interaction or functional association. In this way, a pair of genes with similar profiles across many of bacterial, archaeal and eukaryotic genomes is much more likely to interact than genes found to co-occur in a small number of closely related species **(see Figure 8B)** (Niu et al. 2017).

This method has been used not only to infer physical interaction but also to predict the cellular localization of gene products. However, evolutionary processes such as lineage-specific gene loss, horizontal gene transfer, or non-orthologous gene-displacement can make orthology assignment across genomes complicated and lead to misconceptions. Also, phylogenetic profiling between two proteins can sometimes produce false correlations, although an increasing number of completely sequenced genomes the accuracy of these predictions is expected to improve over time (Škunca and Dessimoz 2015).

1.2.2.3 Gene Fusion or Rosetta Stone

Evolutionary context approaches also include the study of gene fusion across different species. This method is complementary to co-localization and phylogenetic profiles and uses both gene location and phylogenetic analysis to infer function or interaction (Date 2008).

A gene fusion event represents the union of ORFs into a single multi-functional gene and can be considered the ultimate form of gene co-localization (**Figure 8A**): interacting genes are not just adjacent but are also physically merged into a single entity. Some authors hypothesize that the driving force behind these events is to lower the regulation load of multiple interacting gene products(Enright et al. 1999; Skrabanek et al. 2008). In this way, gene fusion analysis provides a useful way to detect functional and physical interactions between proteins.

Nonetheless, predictions based on this approach can be challenging. The most significant barrier is the presence of promiscuous domains that are extremely abounding in eukaryotic and prokaryotic organisms (such as helix-turn-helix (HTH) and DnaJ) (Skrabanek et al. 2008).

1.2.2.4 Protein coevolution

During the course of evolution, if one protein undergoes a mutation, its interacting partner may mirror this event by a compensatory mutation to maintain this binding or functional association by natural selection. It is widely accepted that interacting protein undergoes a coevolutionary process, and that correlated mutations can be used to detect a coevolutionary signal (Yin and Yau 2017). The detection of correlated mutations can not only be used to predict protein-protein interactions but also has the potential to identify specific residues involved in the interaction sites (Pazos and Valencia 2002). In this way, it has previously been shown that a mutation in the sequence of one protein in a pair of interacting proteins is frequently mirrored by a compensatory mutation in its interacting partner (Melamed et al. 2015).

Based on this approach, several methods have been developed, such as mirrortree (Ochoa and Pazos 2010), matrixmatchmaker (Rodionov et al. 2011) or *in-silico* two-hybrid (Pazos and Valencia 2002). Mirrortree evaluates the similarity of phylogenetic trees that represent the putative evolutionary histories of the corresponding protein families. Similar protein trees imply similar evolutionary histories. In its purest form, *mirrortree*-related approaches represent a phylogenetic tree of a family of orthologs as a distance matrix and quantify the similarity between two trees as the Pearson's linear correlation between the sets of values of their corresponding matrices (Pazos et al. 1997; Pazos and Valencia 2001; Pazos et al. 2005; Pazos and Valencia 2008; Ochoa and Pazos 2010; Ochoa and Pazos 2014; Ochoa et al. 2015). MatrixMatchMaker method uses the distance matrices of the protein families as input. Instead of using statistical correlation as mirrortree, it searches for pairs of submatrices that are similar (one being a scaled version of the other) within a tolerance (Bezginov et al. 2013; Rodionov et al. 2011; Tillier and Charlebois 2009). The in-silico two-hybrid method extends this approach by searching for such mutations across different protein or domain families. Prediction of protein-protein interactions using this approach is achieved by taking pairs of protein family alignments, and a correlation function is applied to detect residues that are correlated both within and across families **(see Figure 8C)** (Pazos and Valencia 2002).

1.2.3 Domain-based approaches

Domain-based prediction is based on the principle that for two query protein A and B, along with the knowledge of interacting proteins C and D, if a domain in protein A

resembles the domain in protein C, and a domain in protein B resembles a domain in protein D, there is a possibility that A and B could interact (Dong and Provart 2018).

1.2.4 Function-based approaches

It is widely accepted that interacting proteins tend to have similar rates of gene expression, and some authors used gene coexpression as an indirect way to infer PPIs. Some other studies have revealed that there is no correlation between gene expression profiles and PPI associations, so these type of results cannot be used as a primary source of information but are an excellent metric to confirm previously predicted interactions (Vella et al. 2017).

In other hand, Gene Function Annotations can show whether two proteins act in the same biological process, thus revealing a possible link between those proteins that are more likely to interact than two proteins involved in different biological processes. Biological function information can be accessed from some annotation systems, such as GO (http://geneontology.org/), KEGG (Kanehisa et al. 2012) or EC (Li et al. 2018) which provide information of co-localization and co-functioning in shared cellular process implicit to PPIs (Dong and Provart 2018).

Finally, some authors hypothesized if a pair of proteins have an identity to the sequences of another pair of genes or proteins with described interaction in other species (orthologous proteins), they are supposed to have similar functions that infer the relationship of interactions. This approach uses multiple sequence alignments to define the similarity of full sequences or residues as an index to predict interactions between proteins (Zhang et al. 2013; Kotlyar et al. 2015; Chang et al. 2016)

1.2.5 Deep Learning

Deep-learning algorithms mimic the deep neural connections and learning processes of the human brain. In the last years, they have increased in attention due to their successful applications in different areas such as machine vision, voice, and signal processing, sequence and text prediction, and computational biology topics, altogether shaping the productive AI fields. Compared to older machine learning methods, deep-learning algorithms can process large-scale data and learn useful and more abstract features. Recently, these algorithms have been applied to bioinformatics in order to manage

increasing amounts and dimensions of data generated by next-generation techniques (Min, Lee, and Yoon 2017; Zhang et al. 2017).

Deep learning methods are extensively used in biology to complement experimental data, otherwise hard and costly to acquire in areas (for more information see https://github.com/hussius/deeplearning-biology). The recent progress in deep learning have unraveled new possibilities for reportedly complex predictions in biology, as it was recently demonstrated for protein structure prediction by AlphaFold, a Deep Learning approach developed by Google that is able to predict the folding of proteins based solely on the amino acid sequence of a given protein (AlQuraishi 2019) (https://deepmind.com/blog/article/alphafold) and annotate the protein universe (Bileschi et al. 2019).

Several groups have developed machine learning and deep learning methods to predict Protein-Protein interactions (Wang et al. 2017; Sun et al. 2017; Hashemifar et al. 2018), but they are not being used extensively. As the scientific community gains knowledge on how to use and interpret the results rendered by these techniques, more advances in the area will come.

1.2.6 Text mining

Text-mining based techniques try to automate the extraction of interconnected proteins through their coexistence in sentences, abstracts, or paragraphs within texts present in databases such as Pubmed. This can be done by searching for statistically significant co-occurrences between gene names in public repositories and online resources. Such approaches are very promising as they significantly expand the available PPI knowledge. More complex Text Mining (TM) methodologies use advanced dictionaries and generate networks by Natural Language Processing (NLP) of text, considering gene names as nodes and verbs as edges giving a semantic notion on the graphs. Notably, even newer developments use kernel methods to predict protein interactions from literature (Tsai 2011; Shatkay, Brady, and Wong 2015; Subramani et al. 2015; Papanikolaou et al. 2015).

2. Objective of the study

Inspired by recent works that analyzed large-scale associations of RTs with CRISPR/Cas systems and Argonaute proteins (Shmakov et al. 2018; Ryazansky, Kulbachinskiy, and Aravin 2018; Shah et al. 2019), which are systems that posses shared properties with retroelements, we explore the putative systems in which Reverse Transcriptases are involved. Due to the nature of our data (already used in previous works (Toro et al. 2019)) and taking in mind that we wanted to search large phylogenetically consistent patterns of associations, we combine several described methods to predict these associations.

Our main objective is to expand the knowledge about RTs and to explore the different subgroups by performing phylogenetical and genomic neighborhood analysis and to test whether the incorporation of RTs into novel systems resulted in a functional association that persisted across evolution.

In order to achieve the proposed goals, the specific objectives of this study are:

1. Design a method to retrieve RT's neighbor proteins
2. Execute a clustering method to group proteins by sequences
3. Identify novel putative retroelements based on specific phylogenetic patterns
4. Analyze novel putative retroelements
5. Hypothesize the function of novel RT-containing systems based on the available data

3. Materials and Methods

As the main objective of this study is to search for novel RT-containing putative systems, we designed a computational pipeline in order to achieve it based on previous analysis but with a conceptual shift regarding data analysis. It consisted of 6 different and subsequential steps that are summarized in **Figure 10** and explained in detail in the next sections.

3.1. Prokaryotic Reverse Transcriptases dataset

To explore the diversity of prokaryotic RTs, we used a previously published dataset consisting of 9141 unique representative entries derived from the clustering at 85% sequence identity of 198760 reverse transcriptases (Toro et al. 2019). This heterogeneous dataset was built through several different approaches. RT sequences annotated as RNA-directed DNA polymerases (145,379) or reverse transcriptases (52,684) were downloaded from the PATRIC web server, and 133 new protein entries (23 February 2018) with RT-Cas1 architecture were retrieved from the Conserved Domain Architecture Retrieval Tool (CDART). These two datasets were merged and incorporated into a previously analyzed dataset of 558 RT sequences, including 137 type III CRISPR RT/RT-Cas1 proteins closely related to group-II intron-encoded RTs. Six additional proteins from *Streptomyces* species (annotated as hypothetical proteins) predicted to be RTs linked to type I-E CRISPR-Cas systems were also included in the analysis. The resulting dataset of 198,760 proteins was then filtered by selecting the RT domain (RT0-7) of the proteins with a length $\geq$ 200 amino acids, in multiple-step clustering with a threshold of 85% sequence identity. This procedure resulted in the construction of a set of 9,141 diverse unique RTs for further analysis. The various steps used to compile the final dataset are described in **Figure 9.**

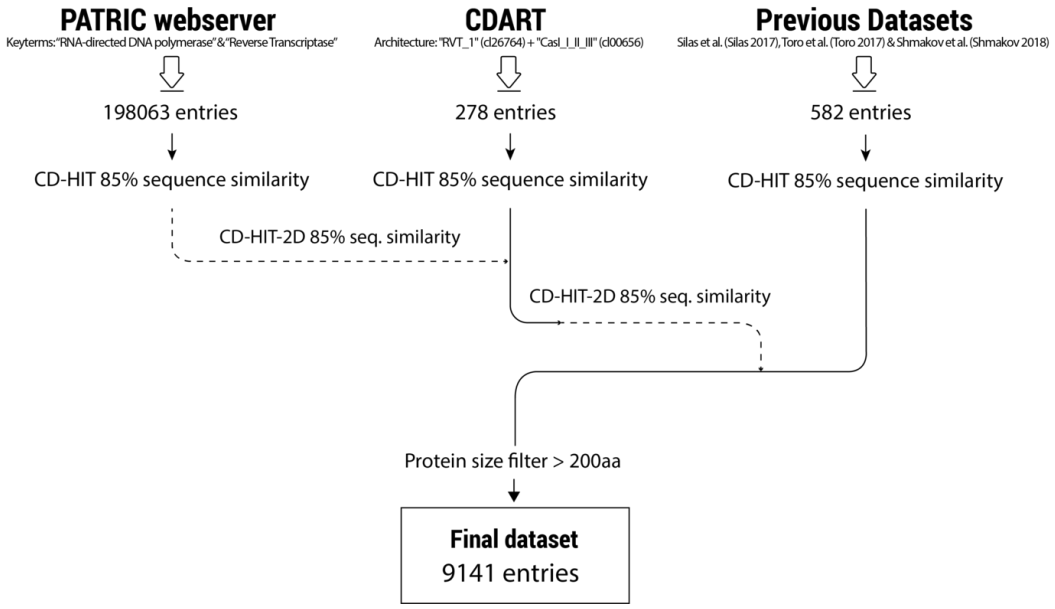


Figure 9. *Schematic representation of the procedure leading to obtaining the final dataset. (Toro et al. 2019)*

After this, MAFFT software was used and progressive methods to perform the MSAs. The MSA corresponding to the RT0-7 domain of the 9,141 entries in the final dataset was filtered to remove sites containing gaps in more than 50% of all sequences. The phylogenetic trees were constructed with the *FastTree* program and the WAG evolutionary model, using pseudo counts (recommended for sequences containing large numbers of gaps) and a discrete gamma model with 20 rate categories. The clades were assigned to the inner nodes showing a high local support value (≥0.92), and subclades were assigned when a large number of sequences were restricted to particular phyla.

3.2. Clustering of neighbor proteins

In order to analyze the genomic context of the above mentioned RTs, we retrieved the annotated proteins located within ± 30 kb of each of them. Proteins shorter than 30aa or larger than 3000aa were discarded from this analysis due to poor downstream alignment results. This procedure resulted in a dataset containing 190277 protein sequences that were further clustered using the suite MMseqs2 (Steinegger and Söding 2017).

As it was needed to cluster these proteins at low sequence identity (below 40%), we used *mmseqs cluster* with parameters *-s 15 –cluster-mode 1 –cluster-steps 9*, which resulted in [92017] different clusters. After this, we performed more in-depth iterative

profile-profile searches until convergence to merge close clusters and cover more remote homologs using *mmseqs search* and *mmseqs result2profile* utils, as described by Martin Steinegger et al. This rendered 62277 different clusters. After clustering, multiple sequence alignments were performed for every cluster using MAFFT with parameters *–globalpair –maxiterate 1000 –reorder*, and Hidden Markov-Models protein profiles were built using the alignment and hmmbuild.

We used –cluster-mode 1 as indicated by Steinegger et al. to increase the sensitivity of the clustering and cover more remote homologs. It consists of a connected component clustering based on transitive clustering starting at the mostly connecter node, and where all the nodes that are reachable in a breadth-first search are members of the cluster. Although it was recommended to not to run iterative searches using –cluster-mode 1, we obtained better preliminary results with this procedure than using typical Greedy Set cover (–cluster-mode 0).

With the aim of diminishing the singleton sequences and increasing the cluster size, we compared all singletons to the previously constructed protein profiles for each cluster using hmmsearch. Singletons were included in the best-reported score cluster with an evalue above 1e-10. This allowed us to avoid the loss of information regarding non-clustered sequences and detect remote homologies while augmenting the intra-cluster diversity. However, this increased the level of noise and the number of false positives. After the clusters were made, we annotated each of them by comparing the representative sequence of every cluster to the PFAM database.

At the end of the procedure, we obtained 62277 different clusters, of which 5413 had more than 5 members. The number of members that define the threshold to be selected is somehow arbitrary. Different authors used 3 members as a threshold, but we decided to go with 5, the same number as the minimum row number for biclustering (see Clustering of presence/absence matrix). Preliminary tests were made (not included) with different values, and the results indicated that the more computing-time affordable threshold was above 5. However, sensitivity decreases as we increase this number and then it is not recommended to choose thresholds above this level.

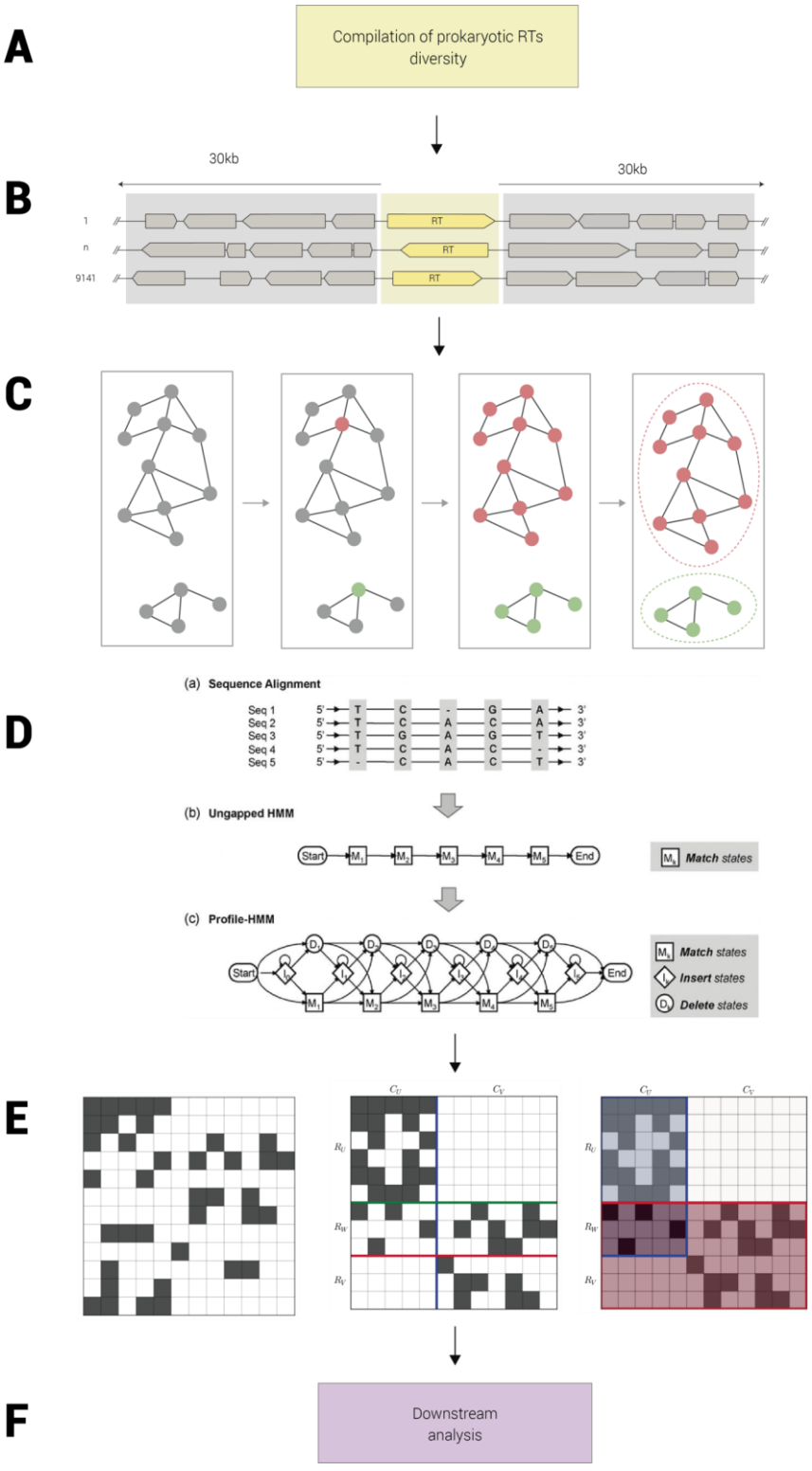


Figure 10. Schematic summary of the computational pipeline used in this study (A) Compilation of prokaryotic RTs diversity. (B) Retrieval of neighbor proteins. (C) Clustering of proteins using mmseqs2. (D) Construction of protein Hidden Markov Models Models profiles. (E) Building and clustering of the matrix. (F) Downstream analysis including expansion of clusters, phylogenetic analysis and domain prediction

3.3. Presence/absence matrix

As described in previous works (Toro et al. 2018; Toro et al. 2019), the different groups of prokaryotic RTs are phylogenetically consistent, meaning that they should be restricted to a given branch. In this way, we used the phylogenetic information derived from the tree to test the association of different proteins with RTs, and we constructed a presence/absence matrix, where the rows were the 9141 RTs, and the columns were the different neighbor protein clusters. In case of RT nº X having a neighbor RT cluster Y, the matrix should have a 1 at the position [X , Y], and 0 on the contrary case. This procedure rendered a binary matrix with 9141 x 5413 cells. Due to the size of the data to analyze, we decided to implement a cluster analysis algorithm with the aim of selecting those clusters that met a criterion of association.

3.4. Clustering of presence/absence matrix

Cluster analysis or clustering is the organization of a collection of elements into groups based on their similarity (according to some criteria that need to be defined), that can be useful in a wide variety of task including machine learning, gene expression analysis or image analysis (Jain, Murty, and Flynn 1999). In this way, clustering the presence/absence matrix would allow us to detect putative associations between RTs (rows) and neighbor proteins (columns). However, simple clustering algorithms would be limited in the detection of multi-protein systems (i.e., different protein clusters associated with the same RT).

Biclustering (also known as block clustering, co-clustering or two-mode clustering) is a non-supervised technique associated with data mining that relies on the simultaneous clustering of the rows and columns of a matrix. In this way, the biclustering algorithm generates biclusters, which are a subset of rows that exhibit similar behavior across a subset of columns.

Due to the richness of algorithms, we found that choosing and comparing reliable clustering methods for a particular problem is an inherently difficult task given that each method uses a different approach that may be only useful in specific scenarios. However, recent systematic comparative evaluations of biclustering techniques helped us to solve this issue. The decision of which algorithm to choose relied on the nature of our data, which is a binary matrix. One of the most used algorithms to bi-cluster binary data is

Bimax, which is a divided-and-conquer method, that is, the input matrix is iteratively partitioned into a set of sub-matrices, until n matrices are obtained, where n is an input parameter for the number of desired or expected biclusters. Bimax also requires minimum number of rows (mnr) and minimum number of columns (mnc) of a bicluster as necessary inputs.

However, there are two main objections that should be considered before Bimax choice. Firstly, we don't know *a priori* the number of biclusters in our data, and secondly, some biclusters might be lost due to premature division of the data matrix, since partitions cannot be reconsidered once they have been divided

Another biclustering algorithm for binary data is Bibit, introduced by Rodríguez-Baena et al. Bibit algorithm searches for maximal biclusters in binary data sets by applying the logical AND operator over all possible gene pairs. Bibit algorithm belongs tho exhaustive enumeration algorithms, which assume the idea that the best biclusters can only be identified using an enumeration of all possible biclusters existent in the matrix. Moreover, exhaustive enumeration algorithms usually try to avoid an exponential running time by restricting the size of the searched biclusters. For that reason, the minimum number of rows (mnr) and minimum number of columns (mnc) of a bicluster should be necessary inputs.

Retroelements are usually located within the named "prokaryotic dark matter" where the genes don't seem to follow a specific pattern due to the evolution from transposable elements and the high recombination frequency. In addition, mobile retroelement can have defined retrohoming targets, and it is possible that they will be inserted in a designated genomic region, and it is also frequent that retroelement has undergone horizontal genetic transfer as a single unit or being part of a broader genomic region.

In this way, taking into account that retroelements can have determined integration targets, that it is common to find ubiquitous genes surrounding RT-coding genes and that they can undergo horizontal genetic transfer, it is logical to suppose that phylogenetically distant RTs not necessarily related show similar genomic neighborhood.

To avoid that, we clustered the matrix using Bimax to detect putative biclusters with at least 4 rows of dimension (mnr =4, i.e., it should be present at least in 4 contiguous RTs). However, we only allowed reordering and expansion of the columns and not the rows.

This allowed us to maintain the phylogenetic information of RTs in search of specific patterns of co-evolution that would add strength to the detection of putative associations while avoiding spurious or false associations due to the artifacts aforecited. In this way, after selecting the biclusters, we removed the columns non-related to any cluster.

To reduce the time of execution of the algorithm, and taking into account that we could be clustering neighbor proteins of phylogenetically different RTs, we divided the binary matrix into 5 different submatrices based on previously described groups.

3.5. Downstream analysis

To explore the putative associations discovered by the pipeline, we performed a downstream manual analysis of the most relevant protein clusters and systems obtained (summarized in Table X). This consisted of expanding the number of cluster sequences using hmmsearch, aligning the resulting expanded cluster, constructing phylogenetic trees and measuring coevolution between proteins in a given system.

3.5.1. Cluster expansion by hmmsearch searches

To enlarge the size of the protein clusters, we searched the UniprotKB database with hmmsearch or jackhmmer webserver using the previously constructed hmm protein profiles, given that they allow recovering more remote homologs in comparison to typical BLASTp searches. Iterative searches using jackhmmer were only used when hmmsearch rendered few matches, as it often produced false positives.

3.5.2. Alignment, Phylogenetic Trees and HHpred searches

Alignments were performed using MAFFT with parameters *–globalpair –maxiterate 1000 –reorder*. After this, phylogenetic trees were built using IQ-TREE with automatic detection of substitution model and 1000 iterations of ultrafast bootstrap. When PFAM searches performed in step C did not yield any results, or when the sequence analysis required a more profound understanding, we used HHpred to search PDB, CD, PFAM, COG and PRK databases using the previously built alignments as input.

3.5.3. Coevolution measure

Several methods exist to study and predict the interaction between proteins, such as co-purification, gene neighborhood analysis (the main focus of this study), Rosetta Stone or

Domain Fusion, similarity of phylogenetic trees, structure analysis, or predictive machine learning software based on experimentally validated protein interactions as training set (see Introduction).

One of the most used is the comparison of the evolutionary histories of the proteins based on their phylogenetic trees. If we assume that interacting proteins experience similar evolutionary pressures, their trees will have similar topology and measuring these similitudes can reveal the degree of coevolution. Using this approach, different groups have tried to develop tools and algorithms to determine and predict the coevolution of proteins (Goh et al. 2000; Pazos and Valencia 2001; Pazos et al. 2005; Ochoa et al. 2015).

Although in this study, we were already analyzing proteins predicted to interact based on gene-neighborhood, further and deeper coevolution analysis using the mirror-tree approach can provide strength to the assumption that those proteins are physically interacting. However, special care needs to be taken with this procedure when analyzing probable horizontal genetic transfer events often involving reverse transcriptases or retroelement, or when analyzing modular cassettes such as CRISPR/Cas systems. As Pazos et al. pointed out, "non-standard evolutionary events leave landmarks in the trees of the proteins resulting in species being far from where they should be in the canonical tree of life and close to species not related with them by the canonical phylogeny. Both factors affect the similarity between trees and their application for prediction of interactions ".

When analyzing neighbor proteins that could be acquired as a modular gene system, it is crucial taking into account the 16S rRNA distances between the species compared to the given protein-coding genes that could be acquired via horizontal gene transfer. One approach that overcame this was used by Shmakov et al. to find coevolution between core and accessory CRISPR-linked genes (Shmakov et al. 2018) and consisted of calculating the Pearson correlation coefficient between the alignment distances of proteins and the 16S RNA. However, RNA and proteins have different evolutionary rates, and the distances between them need to be corrected. Shkmakov et al. do not specify if they corrected the distances of rRNA genes taking into account their differential evolutionary rate compared to proteins, but Pazos et al. calculated the average ratio of distance of proteins/distance of RNA to be, on average, 0,42/1. To improve the quality of the results predicted by mirror tree, we also plotted the alignment distances between the proteins of

a given system and their respective 16S RNA. 16S RNA were downloaded for each of the species harboring proteins from a given system/cluster and aligned using MAFFT with parameters –globalpair –maxiterate 1000 –reorder. However, as performed by Shmakov et al. and Pazos et al., using the tree of life distances or a proxy of 16S RNA phylogeny with the species harboring the 9141 RTs could render different and better results. Nevertheless, as pointed out by Talavera et al. (Talavera, Lovell, and Whelan 2015), covariation is not a reliable method to measure coevolution, and thus the calculated score coevolution should be treated with caution.

4. Results & Discussion

4.1 RT association matrix

The designed computational pipeline has led to the construction of a 9141x62277 matrix that represents every possible association of these RTs with neighbor proteins. As there were more than 60.000 clusters, it was needed to set a cutoff from which to consider a significant association. This cutoff was set to 5 (see Figure 11 and Methods), although other works usually set it to 3 ((Ryazansky, Kulbachinskiy, and Aravin 2018; Shmakov et al. 2018; Shah et al. 2019) and are somehow arbitrary. The resulting matrix applying this criterion ended up having 5413 columns, which reduced drastically the computational times required to analyze it. Each column represents a cluster of neighbor proteins, ordered from larger (named RT1) to smaller clusters (named RT5413), and the rows represent each of the phylogenetically ordered RTs (see Methods).

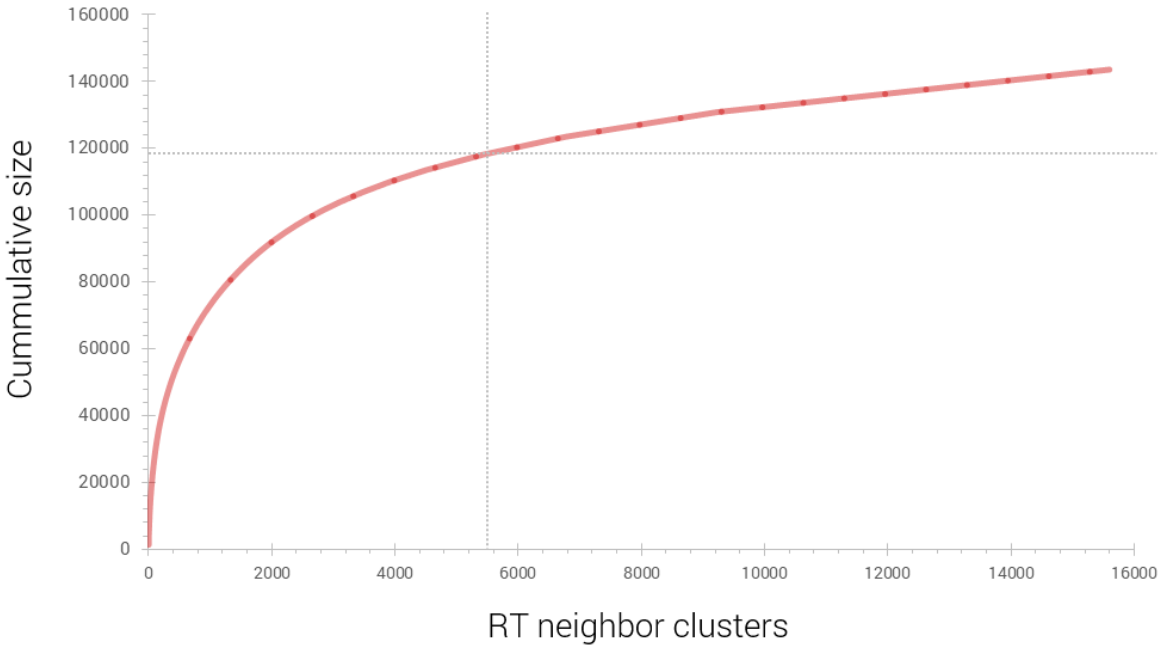


Figure 11. *Cumulative count of protein cluster sizes. Dotted lines represent the selected cutoff*

At first glance, the clustering of the matrix and the search for association patterns revealed at least 30 different cluster groups (i.e., groups of columns associated with a given group of RTs) to be widely associated with RTs (see Table X). Although this number is expected

to increase, and further analyses are already planned in the future, we will only consider these results from now on.

The most exhaustive previous analysis of prokaryotic RTs present in the literature analyzed 3044 different prokaryotic RTs (Zimmerly and Wu 2015), and we almost tripled this number in the present study. In addition, it is needed to take into account that the 9141 RTs analyzed represent a diversity of more than 198000 annotated RTs, thus increasing the magnitude of the described associations. An example of the complexity of the handled data is shown in Figure 12.

Our results showed in (Toro et al. 2019), which uses the same dataset to this work, are in line with previous studies (Kojima and Kanehisa 2008; Zimmerly and Wu 2015). Most of the RTs in this set are classified as group-II introns (47%), Retron/retron-like sequences (25%) and DGRs (12%), whereas the remaining 16% clustered into distinct groups including RTs previously reported to be linked to type III CRISPR-Cas systems, Abi/Abi-like systems or uncharacterized groups (UG groups) (Toro et al. 2019).

As we pointed out in the Introduction, previous analyses have already described associations of RTs with proteins that perform useful cellular functions (Simon and Zimmerly 2008; Zimmerly and Wu 2015). This is the case of Reverse Transcriptases associated with CRISPR/Cas systems, which is the most studied case of domestication of a prokaryotic RT (Silas et al. 2016; Silas et al. 2017; Toro, Martínez-Abarca, and González-Delgado 2017; Mohr et al. 2018; Schmidt, Cherepkova, and Platt 2018; Toro et al. 2018; Toro et al. 2019). As happened with CRISPR/Cas systems, it is possible that RTs can be part of some other functional systems.

To confirm the validity of the pipeline, it should be able to identify the complex association of RT to CRISPR/Cas systems previously described that were found in phylogenetically different groups of RTs such as Abi-like, Group-II like and Retrons in the previous analysis. In addition, it should be also capable of detecting previously described but non-explored associations such as UG3-UG8 Reverse Transcriptases with the aim of demonstrating that it is also adequate to analyze the group of Unknowns RTs.

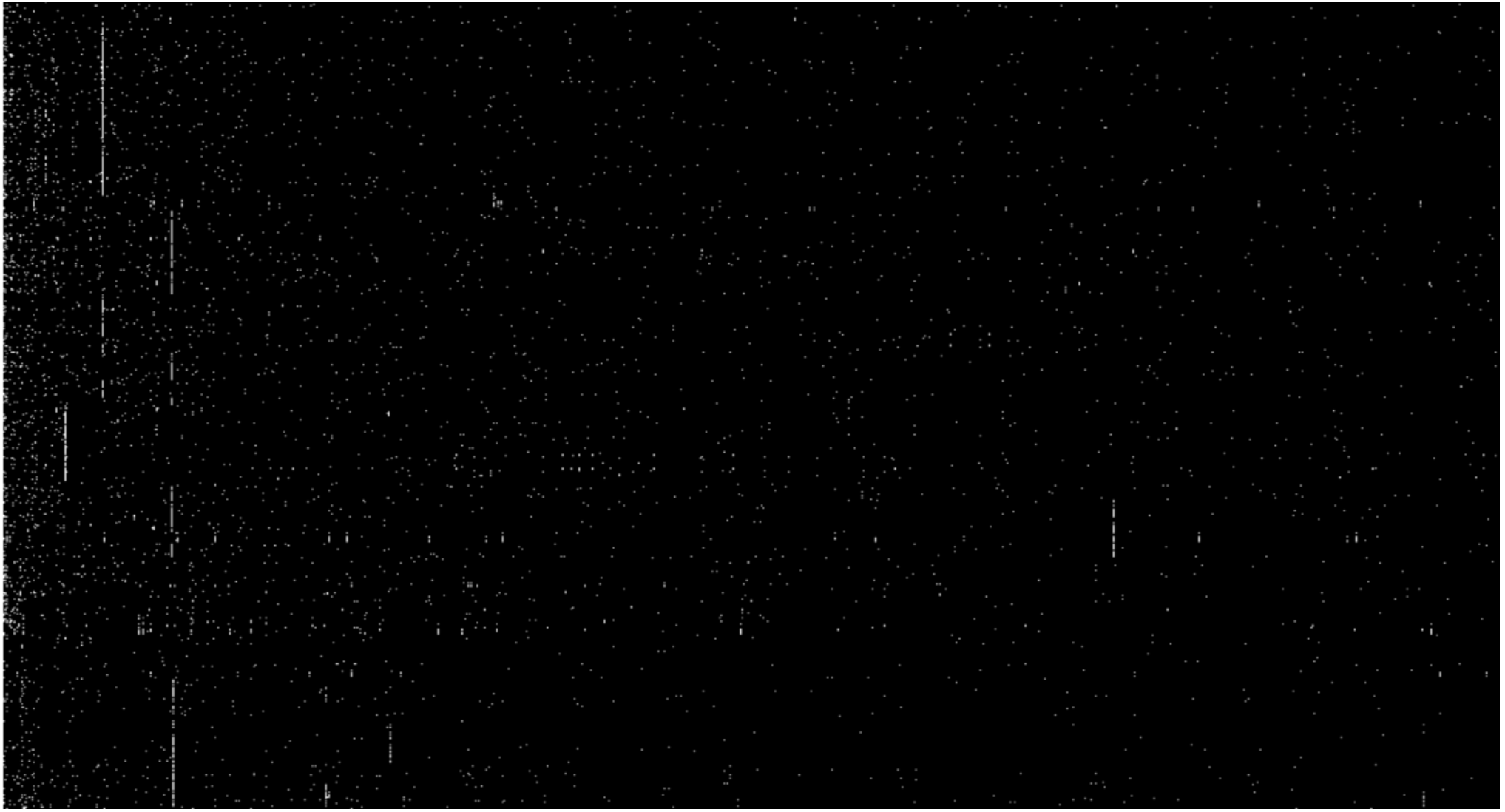


Figure 12. Submatrix of the presence/absence matrix corresponding to RTs associated with CRISPR/Cas and G2L RTs representing 500 rows (RTs) and 2000 columns (RTclusters).

4.2. Confirmatory results

4.2.1 CRISPR/Cas

RTs are commonly found to be associated with Type III CRISPR/Cas systems. Type III systems are complex multiprotein systems that are not yet fully understood. For this reason, some studies have tried to expand the knowledge about these systems by searching novel associations (Shmakov et al. 2018; Shah et al. 2019), and our results are in line with them (see Table). These studies revealed, for example, the association of CorA-like proteins with CRISPR/Cas systems, which corresponds with cluster RT1276 (see table). In addition to previously described associations, we noticed that Clade 3 CRISPR/Cas systems are associated with clusters RT1797, RT643, RT2989, RT1527 and RT1699 too. Sequence analysis revealed that RT1797 is a group of GAF proteins, which are known to be cGMP-specific phosphodiesterases or adenylyl cyclases. On another hand, RT643 contained a CHAT protease domain that was fused in some cases with TPR domains.

Clade 3 RT-Cas1 systems are a particular type of system in which there is a Cas1 protein in addition to the RT-Cas1. While RT1797 and RT643 were co-located to RT-Cas1, RT2989, RT1527 and RT1699 were located near to the other Cas1 (Figure 13), suggesting a possible crosslink.

Apart from the association with CRISPR/Cas systems in G2L RTs, nascent associations of Group-II intron RTs, Retron-like RTs and Abi-like RTs with these systems have already been described (Toro et al. 2019). In line with our assumptions, we were able to identify Retron-like RTs associated with CRISPR/Cas by using the computational approach subject of this work. However, in the case of Abi-like and Group-II intron RTs (not showed), we weren't able to detect it, as the number of representants harboring these associations was smaller than the selected cutoff.

Nonetheless, the purpose of this study is to explore strong relevant relationships between RTs and other putative systems, and not to determine associations that only take place in a small number of species. In this way, the inability to detect these associations does not affect the performance of the pipeline.

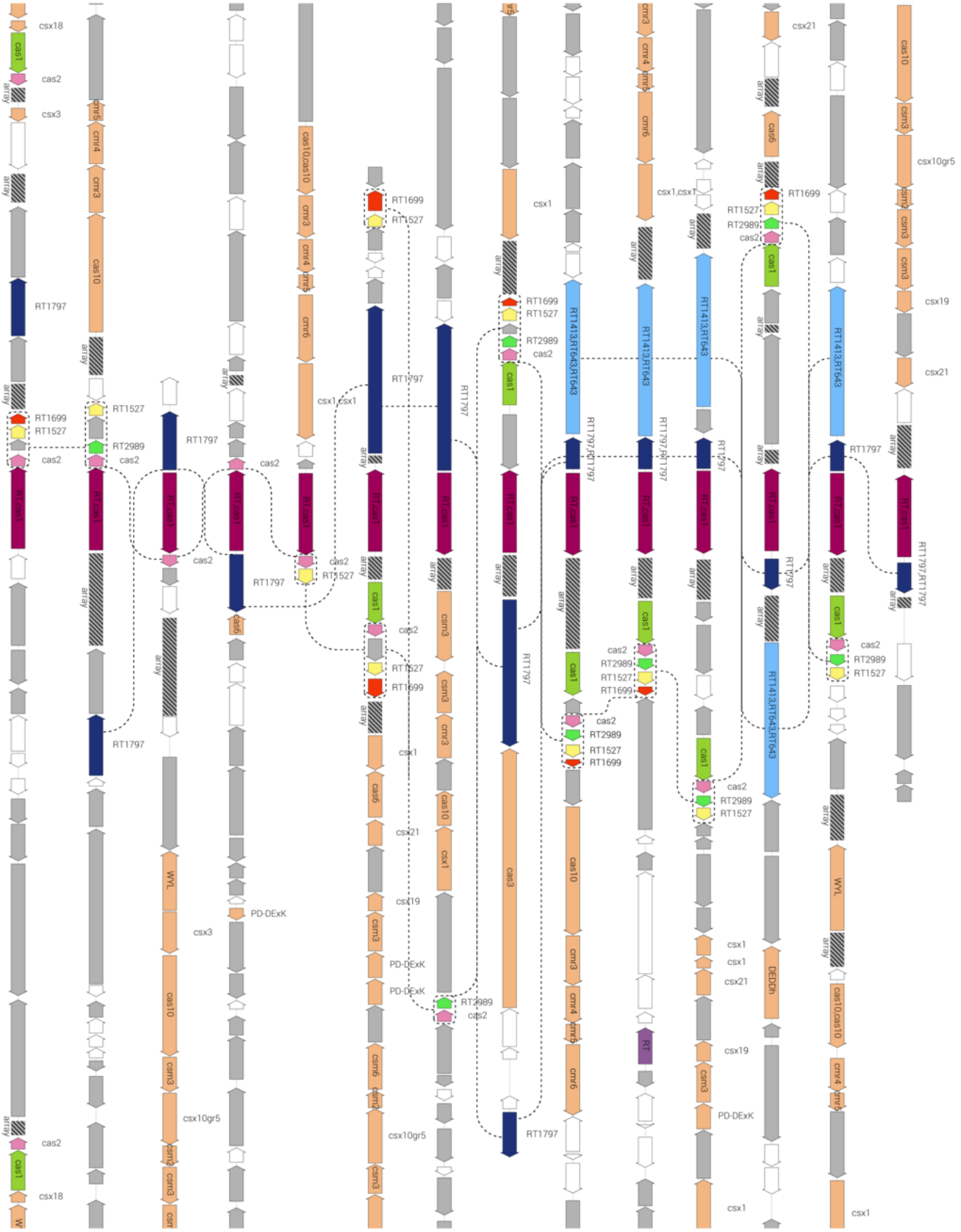


Figure 13. Microsynteny of genomes harboring Clade 3 RT-CRISPR/Cas systems.

The genomes harboring thes systems are described in Appendix II.

4.2.2 UG3/UG8

Reverse Transcriptases from phyletic groups 3 and 8 were found together to each other, as described previously (Kojima and Kanehisa 2008; Zimmerly and Wu 2015). UG8 RTs are slightly larger (~800aa) than UG3 RTs (~400aa), and the difference in size relies on a putative domain located at the C-terminal region of UG8 RTs. Sequence analysis using HHpred did not yield any significant match. Prediction of protein structure using a Deep Learning approach in RaptorX web server revealed that UG8 proteins contain a C-terminal domain that consists of a four bundle folding, composed of four interacting alpha-helices (Figure 14). Although this fold is highly ubiquitous and it is not possible to infer any function, we can find it on thumb domains of RNA/DNA polymerases (including RTs) and sensory proteins.

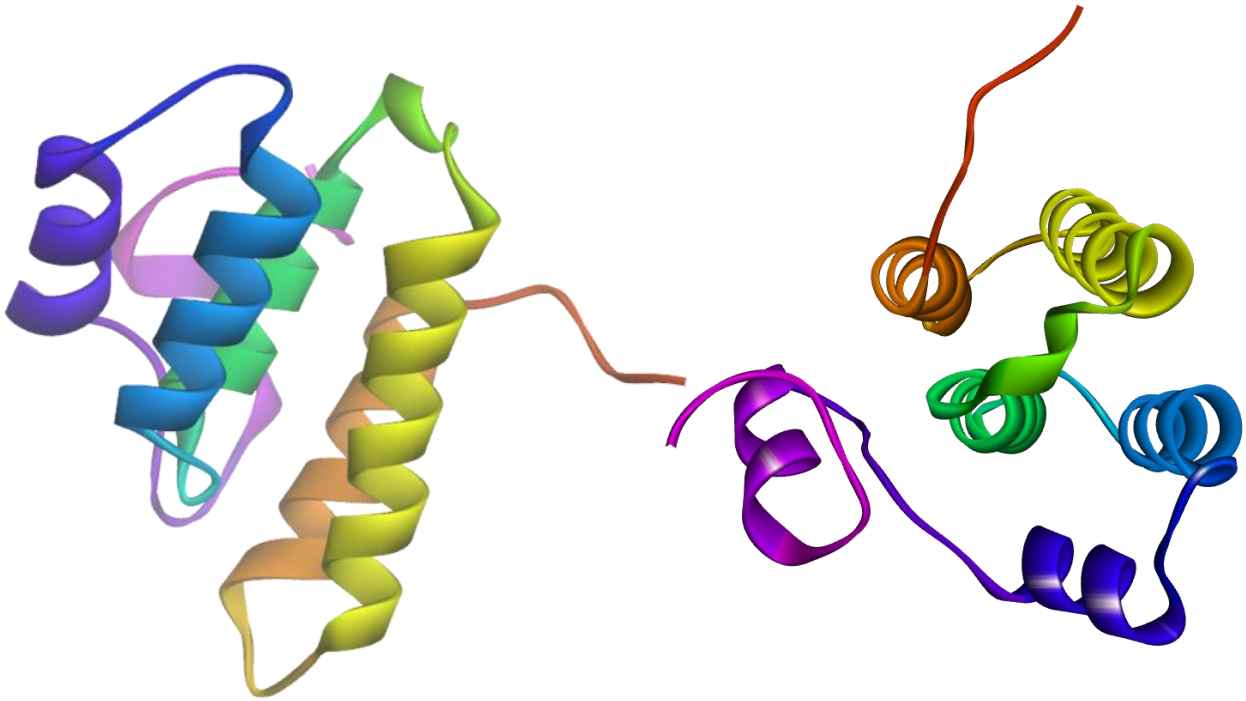


Figure 14. *Predicted structure of the C-terminal domain of UG8 Reverse Transcriptase modelled by RaptorX webserver. The four alpha-helices are colored in orange, yellow, green and pale blue.*

Apart from the co-localization, UG3 and UG8 showed a high degree of coevolution, as MirrorTree analysis suggested (0,848). Although this association has been previously described (Kojima and Kanehisa 2008; Zimmerly and Wu 2015), its possible function or origin has not been hypothesized yet. The genomic neighborhood of the genes encoding this putative protein complex is highly diverse and does not seem to follow a specific

pattern, suggesting that the system forms as a two-component unit and it could be a mobile retroelement. Based on the results obtained, we hypothesize that this system could be a putative sensor complex that detects RNA and amplifies the signal via Reverse Transcription, allowing for a more sensitive response.

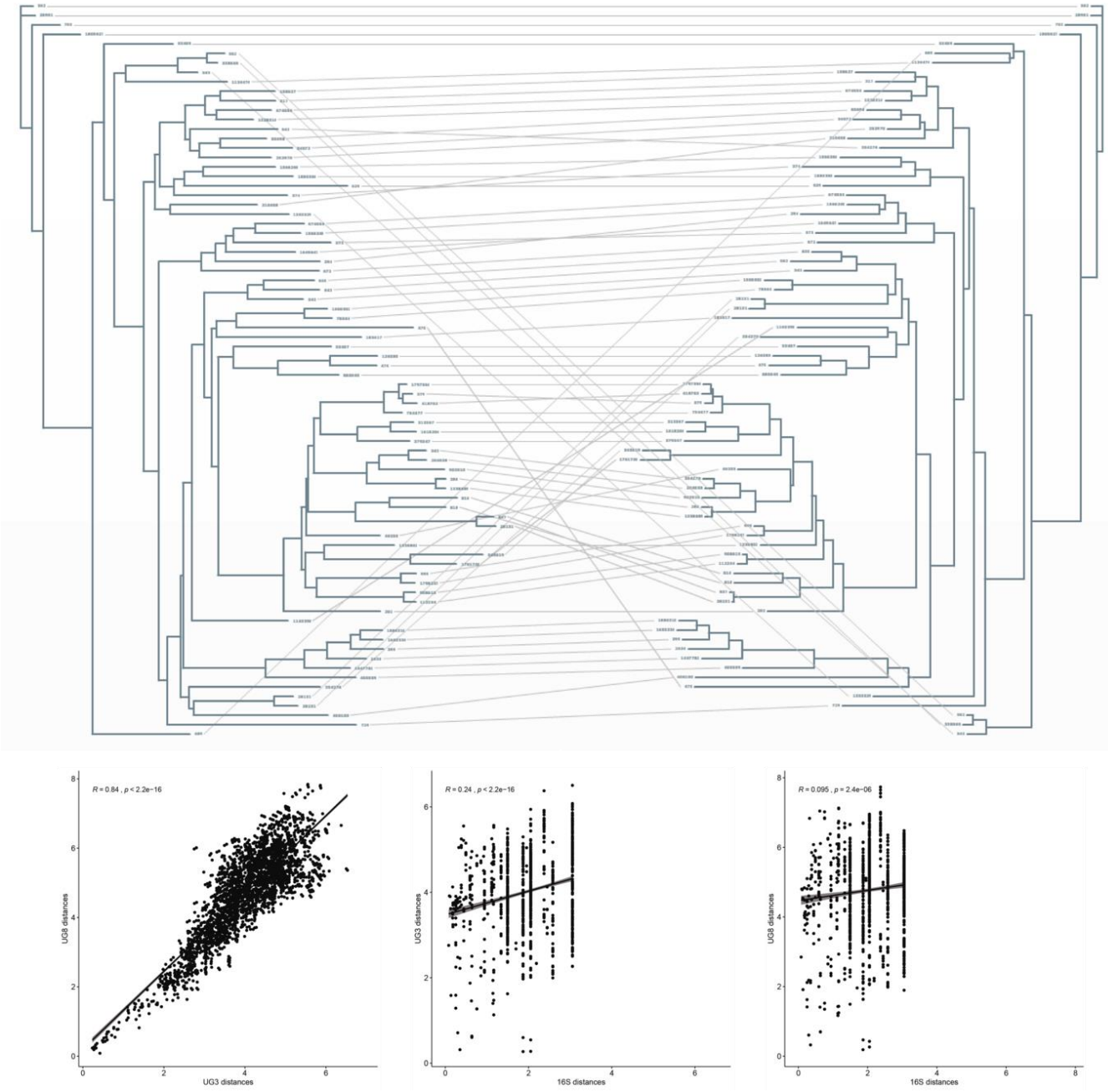


Figure 15. *Coevolution prediction of UG3 (left) and UG8 (right) RTs based on mirrortree approach and distances plot of UG3/UG8 RTs and 16S rRNA of species harboring these RTs.*

As we have shown with the example of RT-CRISPR/Cas and UG8/UG3 RTs, this pipeline not only has been able to identify previously described associations, but it also showed novel associations for Clade 3 RT-CRISPR/Cas. However, these associations need to be explored using other methodological pipelines, as the modularity of CRISPR/Cas systems is hard to study by only performing phylogenetic analysis.

In the next sections, we will describe some of the associations found for the rest of the phylogenetical groups.

4.3 UG groups

4.3.1 Uncharacterized groups (Unknown Groups 26, 27, 7, 2, 25)

UG26

Within UG26 RTs, we can find an unclassified protein cluster (RT3503) associated with the RT. The protein size varies between 200-300~ aa, and sequence analysis using DeepCoil revealed that it contains coiled-coil segments, and it is a probable nuclease. Further analysis using HHpred revealed that it shares similitudes at the C-terminus with YcjD (COG2852), a very-short-patch-repair endonuclease.

Due to the small size of this cluster, we tried to expand it by using hmmsearch and previously built profiles. However, this procedure did not yield a significative augment of the cluster sizes. In this way, we cannot go further in the study of this association, and we will need to explore novel methodologies in order to retrieve more information.

UG27

Group 27 Reverse Transcriptases are associated with clusters RT1931, RT2783, RT896, RT2166, RT2244, RT2893, RT3246, RT4709, RT4776, and RT4846. Species harboring these genetic cassettes belong to Bacteroidales and Clostridiales taxonomic orders. Of the abovementioned clusters, RT1931 and RT2783 coding genes were present in almost all of the cases, and its location relative to the RT coding gene was uniform, whereas the rest of the clusters showed an irregular pattern and were absent in some cases. To increase the number of candidates to analyze, we searched the UniprotKB database with hmmsearch using previously constructed profiles for all of the clusters above mentioned. The results were clustered at 95% sequence identity in order to remove redundant entries. We found

that nearly all of the proteins similar to RT1931 and RT2783 were located near to an RT, whereas this didn't happen for the rest of the clusters. RT1931 protein size varies between 300-400aa, and RT2783 proteins size is around 150aa. Both of them showed high pairwise similitude when aligned (~50%), and its evolutionary rates were pretty similar, suggesting a possible functional link.

Genomic analysis showed that the distribution of these protein-coding genes surrounding the RT is not uniform, reminiscent of what happens with the distribution of CRISPR genes associated with RT. This modularity suggests that protein from this group could be part of a functional putative system with accessory components. However, It is not clear whether the RT could play a central role in these putative systems or just contributes by providing extra functionalities.

Sequence analysis of RT1931 and RT2832 did not render any useful information, but annotations of RT1931 links it to Ubiquitin carboxyl-hydrolase. Previous studies (Chatzidaki-Livanis et al. 2017) have shown that some strains of *Bacteroides fragilis* secrete an ubiquitin-like protein with inhibitory activity against other *B. fragilis* strains. Thus, this could be a putative toxin system or a functional attack system against co-resident cells. In any of the cases, the role of an RT associated with clusters RT1931 and RT2832 remain enigmatic and needs to be further explored.

UG7

Reverse Transcriptases from UG7 are associated with RT629. Sequence analysis using HHpred revealed that it contains a PD-(D/E)XK GxxExxY nuclease domain (TIGR04256), which is characterized by having a Restriction endonuclease-like fold. In this way, this represents another group of Reverse Transcriptase associated with a putative DNA nuclease, similar to what happens with UG26 RTs.

4.3.2 System A: YodC (UG2)

The phyletic group of Unknown proteins 2 is widely associated with clusters RT2002 and RT470. Sequence analysis revealed that both of them share similitudes with a characteristic N-terminal YodC (COG5475) domain with an extension of about ~90aa. In contrast, RT2002 members have a size of about 200-250aa, and RT470 members sizes range from 360 to 480aa. In addition, RT470 proteins display a WYL-like domain at the C-terminus that explains its larger extension.

Very little is known about YodC, but protein models and profiles in databases are restricted to proteins present in *E. coli*, *Salmonella* and *Shigella*, while RT2002 and RT470 clusters from UG2 are encoded by species classified under the *Bacteroidetes* phylum. In addition, this domain seemed to be repeated in both proteins at least two times, having the first of them a more reliable identification.

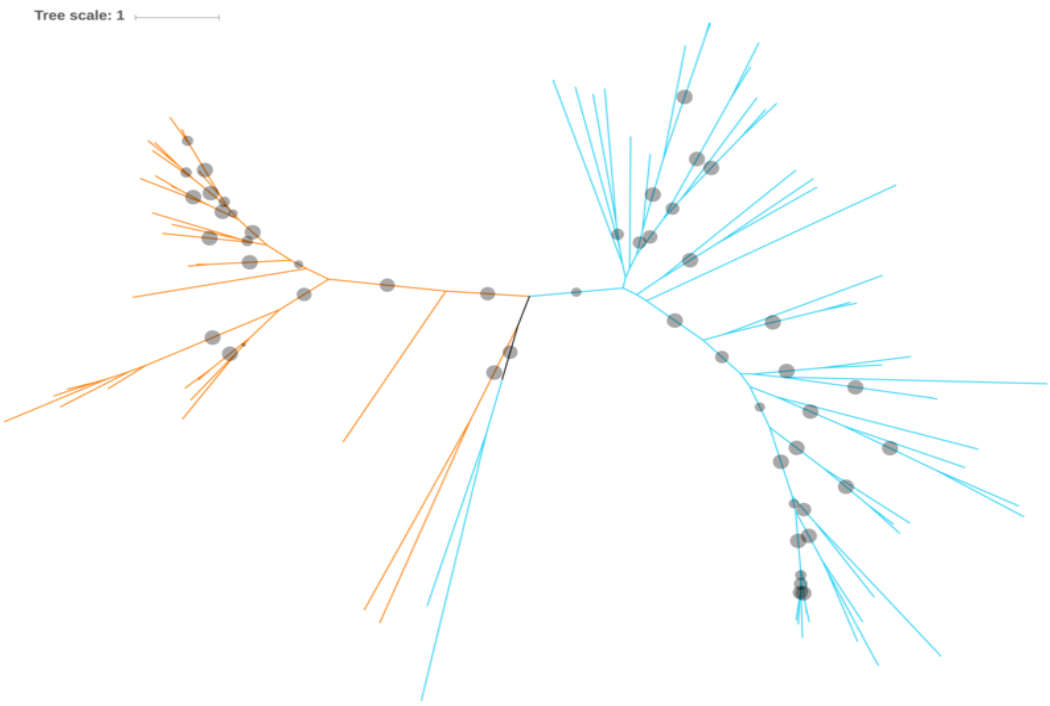


Figure 16. *Phylogenetic reconstruction of YodC-containing proteins located next to UG2 RTs. Black circles represent Bootstrap support values greater than 90% in a size range from 90% to 100%. Nodes highlighted in orange correspond to RT470 proteins whereas nodes highlighted in blue correspond to RT2002.*

Cluster expansion by hmm searches on the UniproKB database and subsequent gene neighborhood analysis showed that all of these proteins are co-localized with RTs. We performed a phylogenetic analysis of these proteins clustered at 95% sequence identity, and it showed that the two domain architectures branched into different phylogenetic groups (*see Figure 16*). In addition, this putative system is frequently surrounded by known defense systems such as RM or Abi systems. As described by Koonin et al., defense systems are nonrandomly co-located, thus suggesting that it could be part of a defense/attack system or provide this function on its own.

On the other hand, the WYL domain is known to be a transcription factor that regulates a response when binding to RNA molecules. This domain is present in CRISPR/Cas systems (4461 a 4470, Clade 9, identified in previous analysis as WYL protein and it has been described to be a transcription factor) and previous observations suggest that WYL is a ligand-sensing domain that could bind negatively charged ligands, such as nucleotides or nucleic acid fragments, to regulate CRISPR-Cas and other defense systems such as the abortive infection AbiG system (Makarova et al. 2014), thus providing strength to our assumptions.

4.3.3 System B: DUF1848 (UG25)

The proteins belonging to the UG25 group were found to be associated with cluster RT2632. Pfam searches suggest that it is a DUF1848-containing protein. This domain is currently functionally uncharacterized and little is know about it, but its C-terminus is linked to the iron-sulfur cluster found at the N-terminus of pfam04055, as they share a cluster of cysteines. Moreover, further analysis of the RT belonging to this group revealed that they are abundant proteins (>900aa) with an N-terminal RNA methylase domain (PF09243), a central RT domain and an unknown C-terminal domain. Bacteria harboring these RTs belong to *Firmicutes* phylum.

To test whether this peculiar architecture was present among other organisms, we searched the UniprotKB database using an alignment of UG25 RTs and hmmsearch. The results were filtered in order to retain proteins larger than 600aa, and further alignments were performed to elucidate if these homologs retained the N-terminal RNA methylase domain.

After manually remove those not showing this domain, we obtained 16 non-redundant proteins with the given architecture. Surprisingly, we found that while 6 of these proteins were co-located to RT2632 cluster and belong to *Firmicutes* phylum, the remaining 10 proteins were co-located to UvrD helicase (pfam13361-pfam00580) and belong to *Cyanobacteria* phylum. We then constructed a phylogenetic tree (see methods) and observed that they grouped into different branches.

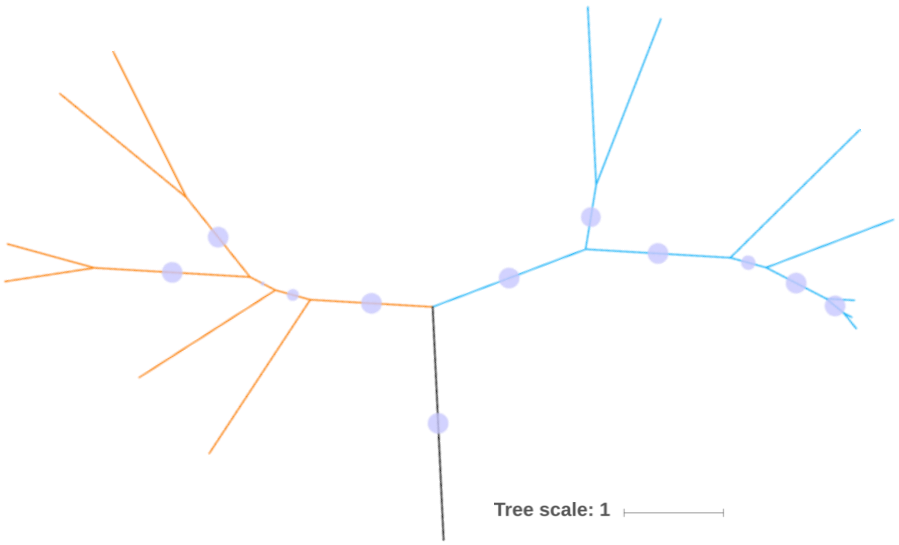


Figure 17. *Phylogenetic tree RNA methylase/RT proteins. The branch of this proteins associated with UvrD is highlighted in blue. Nodes highlighted in orange corresponds to those associated with RT2632. Bootstrap support values > 90% are shown in purple circles.*

4.3.4 System C: VirE/Pri-CT2 (UG28)

A subgroup of RTs belonging to UG28 (Unknowns group 28) phyletic group appeared to be associated with the protein cluster RT1301. According to protein sequence analysis, the RT1301 cluster corresponds to a group of proteins containing an N-terminal virE domain (Virulence protein E; PF08800) and a C-terminal PriCT-2 domain (Primase C terminal 2; PF08707).

The species carrying protein-coding genes from this particular subgroup belong to the phylum *Bacteroidetes*. To test whether this peculiar architecture was typical among other organisms, we used the CDART retrieval system in order to obtain proteins with similar architectures. We found 157 different proteins with virE and PriCT-2 domains, of which 155 derived from *Bacteriodetes* species. The two remaining were carried by *Cellulophaga* phages PhiSM and phi3:1 which are known to infect hosts from the genus *Cellulophaga* (phylum *Bacteroidetes*).

An exhaustive analysis of these phage genomes revealed that their virE/PriCT-2 protein was not associated with an RT, as we weren't able to find any coding region with similitudes to Reverse Transcriptases. In contrast, we observed that virE/PriCT-2 was co-located in both cases to a DNA polymerase III and proteins involved in Queuosine biosynthesis, suggesting a possible link with RTs from G2L/Que group

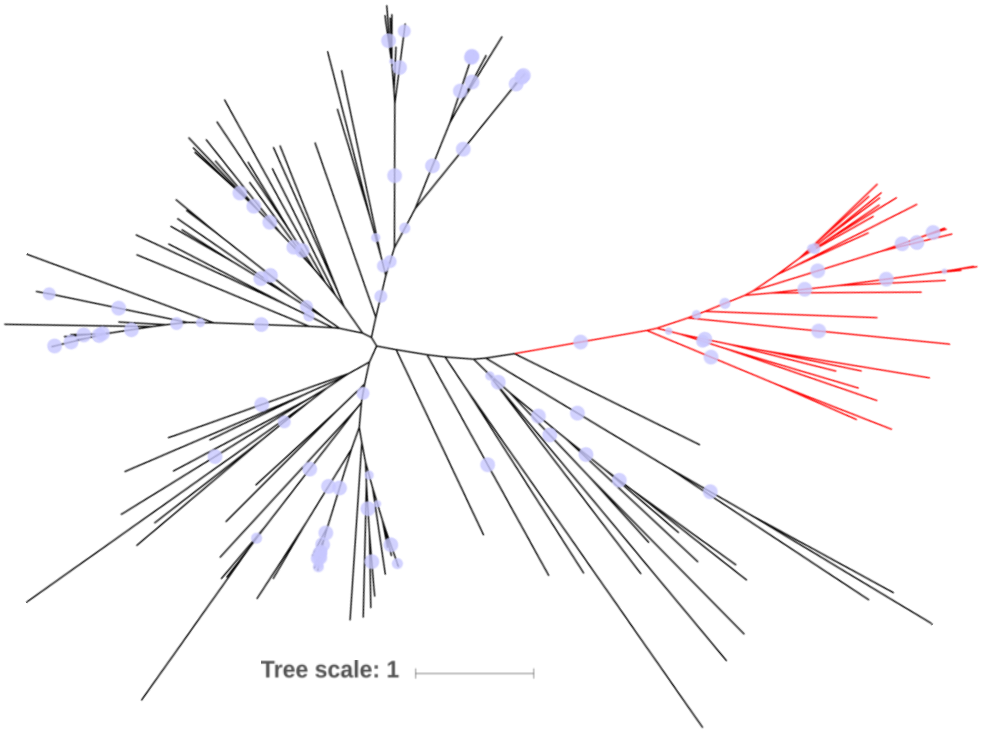


Figure 18. *Phylogenetic tree of 157 N-virE/priCT proteins. The branch of N-virE/priCT associated with RTs is highlighted in red. Bootstrap support values > 90% are shown in purpe circles.*

We then analyzed the neighborhood of virE/PriCT-2 proteins in the Bacteroidetes phylum and constructed a phylogenetic tree based on alignment distances of the 157 proteins. Multiple alignment was performed using MAFFT with parameters *–globalpair –maxiterate 1000 –reorder*. After aligning, misaligned N-terminal and C-terminus were trimmed, identical sequences were filtered, and alignment was repeated. After this, we constructed a phylogenetic tree using FastTree with default parameters. Consistent with the data, we noticed that only virE/PriCT-2 sequences from a particular branch were located near to an RT (see ???).

After this, we analyzed the coevolution of virE/PriCT-2 and neighbor RTs using MirrorTree, and MatrixMatchMaker approaches (see ???) to test whether these two proteins were interacting or probably acting together as part of a system. We obtained very high levels of coevolution (VALORES) between these proteins, consistent with the hypothesis that they could be part of a putative functional system.

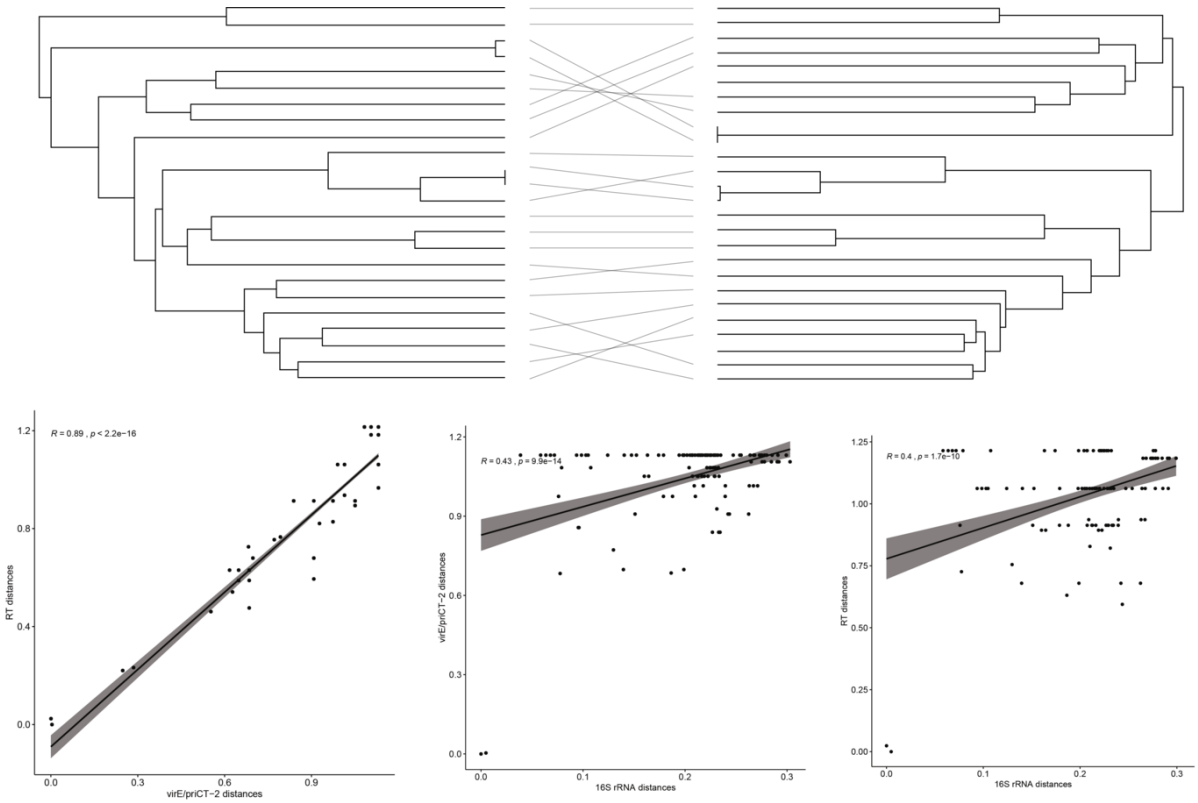


Figure 19. *Coevolution prediction of VirE/Pri-CT2 (left) and UG28 (right) RTs based on mirrortree approach and Distance plot of VirE/Pri-CT2/UG28 RTs and 16S rRNA of species harboring this system.*

In addition, we found this domain architecture to be fused with domains VirE, COG5545, DUF3987, COG3378, and VirE in other proteins present in Bacteriodetes and phages too, but we did not analyze them due to the lack of a significant number.

The association of an RT to a primase has been previously described at least three different times by Kojima & Kanehisa (Kojima and Kanehisa 2008), Zimmerly & Wu (Zimmerly and Wu 2015) and Toro et al. (Toro et al. 2019). Proteins reported by Kojima & Kanehisa are known to belong to the dnaG primase family, similar (or identical) to those reported in Class ML Group II introns (entries 999-1004). Those reported by Zimmerly & Wu belong to the unknown group 10 and consist of an RT-like

domain with an N-terminal primase domain (PriS) and an unknown C-terminal domain. The RT-PriS fusions reported by Toro et al. are predicted to function as an accessory protein of Type III CRISPR/Cas systems. In this way, the virE/priCT-2 domain-containing proteins form a new type of association with RTs. The N-terminal domain of the virE protein has been also found in CRISPR/Cas systems, suggesting that the role of RT within these putative systems can be linked somehow to defense systems. (Zhang, Doak, and Ye 2014). Prototypical Group II intron possessing En domain can use cleaved target DNA as a primer (Zimmerly et al. 1995), and some Group II intron lacking En domain depend on random cleaving of target DNA sequences (Martínez-Abarca et al. 2004). Retrons use a 2'OH of branching Guanosine as a primer (Inouye et al. 1999), and some other RT such as HIV RT uses a tRNA as a primer to polymerize cDNA (Marquet et al. 1995). Finally, other RT-like containing elements such as AbiK (with untemplated DNA polymerase activity) could use an amino-acid a self primer for this reaction (Wang et al. 2011). This shows that RTs can adapt significantly to different forms of priming, suggesting that this protein could be part of another type of priming mechanism.

4.3.5 System D: SLATT proteins (UG17)

Reverse Transcriptases belonging to group 17 are associated with RT269 and RT1640 protein clusters, which are annotated as SLATT (SMODS and SLOG-associating 2TM effector domain family 5) containing proteins. This link between RTs and SLATT proteins has been previously described and discussed in (Burroughs et al. 2015).

SLATT is a superfamily of domains with two transmembrane helices. Multiple alignments revealed a conserved core for the domain consisting of a pair of N-terminal TM helices and a mainly helical C-terminal cytoplasmic region. The TM helices often contain family-specific polar residues that are likely to form an intramembrane aqueous channel that might facilitate the transport of molecules across the membrane.

SLATT proteins associated with RTs proteins are distinguished by the presence of a third TM segment after the C-terminal helical cytoplasmic tail (See Figures 20 and 21) and display rapid sequence divergence relative to all other prokaryotic SLATT domains.

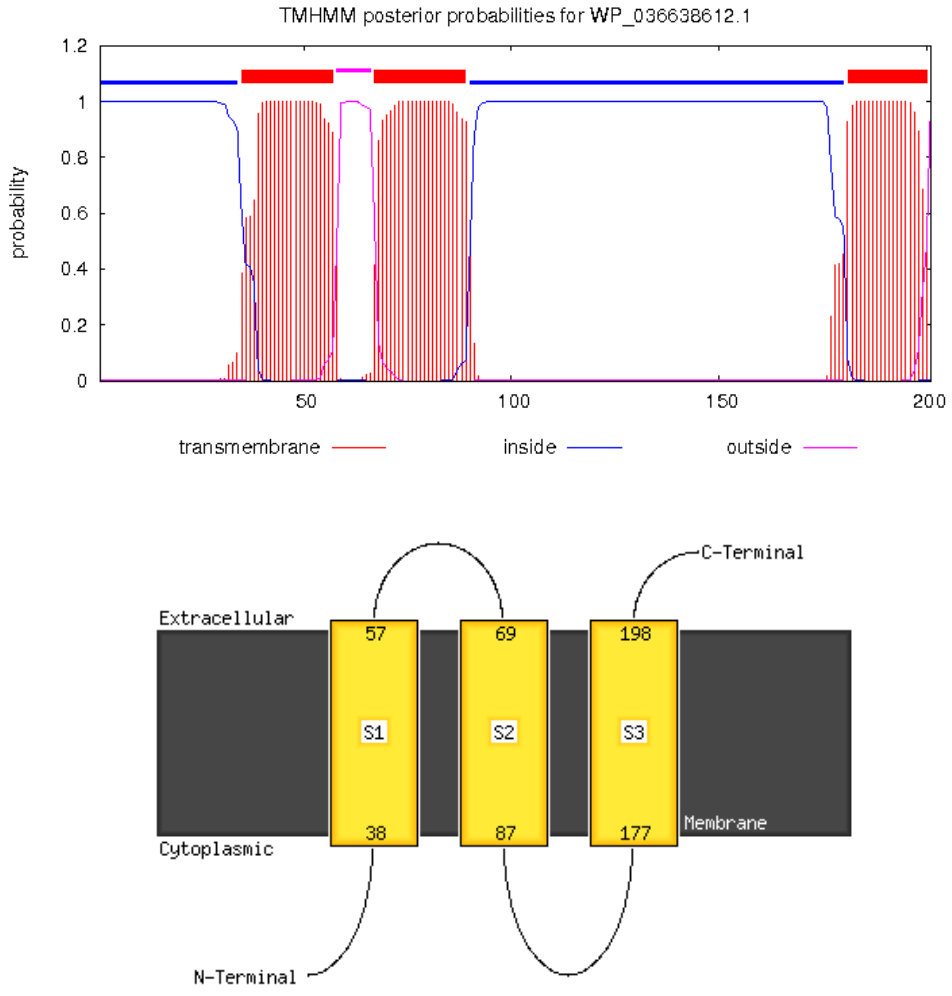


Figure 20. *Transmembrane and intra/extracellular domains prediction of SLATT protein from Pseudomonas Aeruginosa using THMM server v2.0 (up) (http://www.cbs.dtu.dk/services/TMHMM/) and MEMSAT_SVM from the PSI-PRED suite (down) (http://bioinf.cs.ucl.ac.uk/psipred/)*

RTs from this group have the so-called ‘domain X,’ which is thought to bind RNA during reverse transcription. However, like the DGRs, this novel system does not contain the HNH endonuclease found fused in many group-II intron RTs. It has been proposed that this putative system is highly mobile and is seen in diverse proteobacteria, firmicutes, bacteroidetes, and fusobacteria.

Burroughs et al. (2015) hypothesized a novel mechanism of dispersal for this putative retroelement. In their model, RNA copies of the element would be transcribed and translated by host machinery. The resulting RNA transcript would then been reverse-transcribed into DNA by the RT domain and transported out of host bacteria via pores formed by the SLATT protein. The specificity in DNA transport could be mediated by the distinctive C-terminal region of these SLATT domains. Interestingly, given the rapid

sequence divergence of both components of this system, it is likely that the RT is error-prone and has the potential to generate diversity, and it is possible that diversification of the SLATT protein by this mechanism might confer some advantage to the host cells.

4.3.6 System E: HEPN domain-containing proteins (UG22)

Reverse Transcriptases belonging to Unknown group 22 are associated with cluster RT3633, which is identified by HHpred as a HEPN domain-containing protein (COG2445) with an extension of about 200aa. Due to the small number of members of this cluster, we expanded it by searching on the UniprotKB database, and we obtained 22 protein hits with an E-value smaller than 1e-10, being all of them associated with RT. These proteins shared a particular domain located at the C-terminus predicted to possess the catalytic residues **(Figure 21).**

HEPN domain (Higher Eukaryotes and Prokaryotes Nucleotide-binding domain) is known for having RNA binding and cleaving properties, and it has been found in CRISPR-associated proteins such as Cas13, an RNA-guided ribonuclease, or Csm6, an endoribonuclease involved in nucleotide signaling mediated by CRISPR/Cas systems (Anantharaman et al. 2013).

In prokaryotes, HEPN domains are also essential components of numerous toxin-antitoxin (TA), abortive infection (Abi) systems, many restriction-modification (R-M) and occasionally with other defense systems such as Pgl and Ter (Anantharaman et al. 2013). In this way, we hypothesize that this system could be a defense system in which the RT plays an important role.

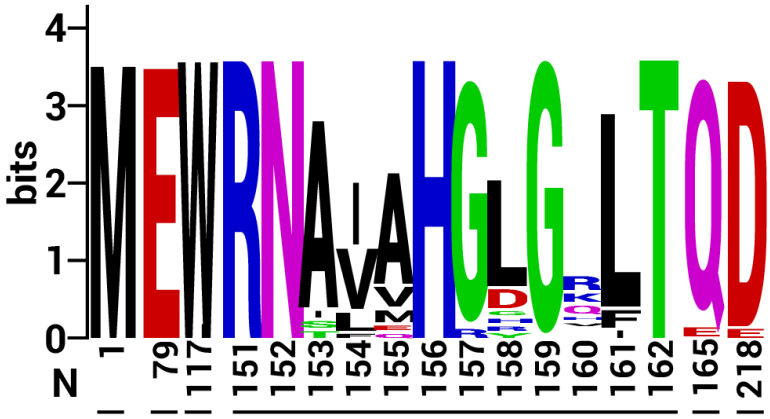


Figure 21. Domain conservation across HEPN proteins located next to UG22 proteins. There is a highly conserved domain from aminoacid 151 to 162, with a characteristic “RNA/HGG” domain.

4.3.7 System F: DNA polymerase I (UG9)

UG9 Reverse Transcriptases are associated with the RT161 group, which is identified as a DNA polymerase I family (PF00476.20). Its members have about ~500aa, and more in-depth sequence analysis revealed that it contains the canonical 5'->3' polymerase domain, but the 5'-3' exonuclease domain slightly varies from those present in modeled DNA polymerase I, and it contains an unknown N-terminal domain with an extension of about 120 aa.

We expanded the cluster by searching the UniprotKB database with hmmsearch using an alignment of RT161 from UG9 as a query. We retrieved 67 different protein hits with an E-value smaller than 1e-100. Genomic neighborhood analysis revealed that all of them are co-located with an RT.

Similarly, proteins from the RT161 group has been found to be associated with RTs from Cluster 4. However, they failed to align together, were slightly smaller (~290aa) and seemed to correspond only with the 5'-3' exonuclease domain.

4.4 Group II-like (G2L)

Reverse Transcriptases belonging to Cluster 4 (G2L) have been found to be associated with clusters RT2, RT13, RT26, RT32, RT130, RT161, RT580, RT632, RT793, RT813, RT981, RT1596 and RT2386.

RT2	HTH (Helix-turn-helix)
RT1596	FGE-sulfatase/C-Lec fold
RT32	Glycosyltransferase
RT580	QueC
RT161	DNApolA
RT793	QueA
RT981	QueE
RT813	QueD
RT632	N-acetylglucosamine deacetylase
RT2386	YaaA
RT26	Glycosyltransferase
RT130	QueE

Table 1. Table showing the annotation of the different clusters associated with Cluster 4 G2L RTs

Some of these proteins belong to the biosynthetic pathway of queuosines, a modified nucleoside that is commonly present in tRNAs and that has been described to be part of putative defense systems. In addition, the RT161 cluster is identified as a DNA polymerase I, and RT32 and RT26 are classified as Glycosyltransferases **(see Table 1).**

As It has been shown recently, Queuosine modification not only acts on RNA but in DNA too. The discovery of ~20-kb guanosine transglycosylase (tgt) genes, alongside 7-cyano-7-deazaguanine ($preQ_0$) synthesis and DNA metabolism genes, led to the hypothesis that 7-deazaguanine derivatives are inserted in DNA. The functions of most DNA hypermodifications are still not known, but some have roles in protection against restriction enzymes, whereas others affect thermal stability temperature, DNA packaging, or transcription regulation (Thiaville et al. 2016).

R–M (Restriction-Modification) systems and genomic islands are often transferred through phage transduction. Several examples of phage-encoded *tgt*-like genes and $preQ_0$ genes have already been reported in the literature, In the characterization of phage 9g, Kulikov et al. (2014) speculated that the restriction endonuclease-resistant nature of the phage DNA suggested it was heavily modified, and they proposed that *tgt* and $preQ_0$ synthesis genes were involved in inserting Q into the DNA (Thiaville et al. 2016)..

We hypothesize that the Queuosine modification island was transferred from phage and remained in the host due to evolutionary advantages such as R-M avoiding. However, the function of this island is not clear, and the role of RT within this system remains enigmatic.

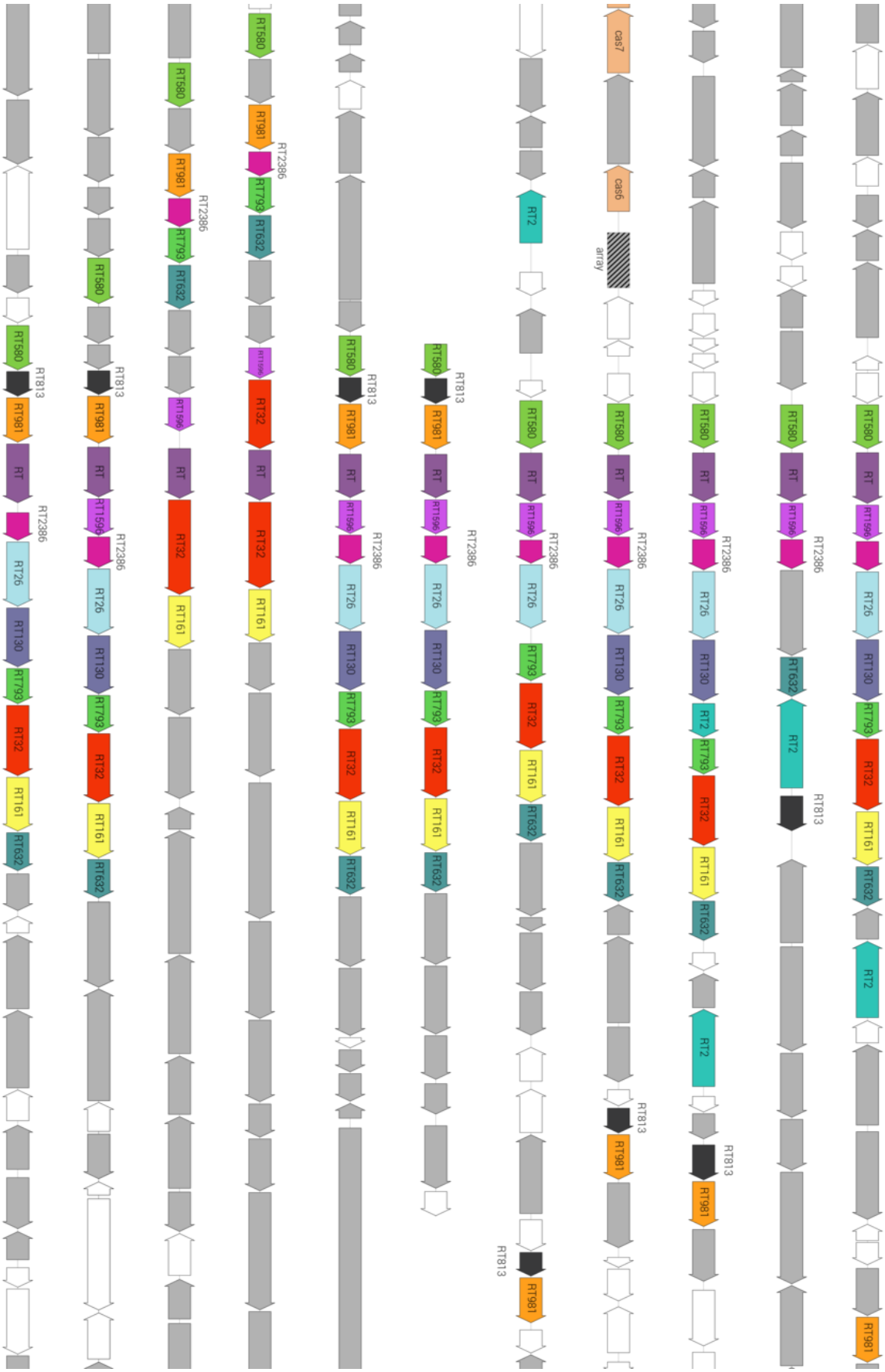


Figure 13. Microsynteny of genomes harboring G2L/Que systems. The genomes harboring thes systems are described in Appendix II.

5. Concluding remarks

In summary, we can conclude that:

1. The large volume of data analyzed coerced us to design a simple yet useful high-performance pipeline that rendered results greater than expected. It has been demonstrated to be valid in different scenarios in which the RTs are present, and It could be applied to the study of other systems.

2. The pipeline revealed novel associations of RTs to different proteins such as HEPN or VirE/PriCT2, suggesting roles beyond mobility. In these systems, RTs can act as an accessory module, but they can also be a key factor in some others.

3. The role of RTs in prokaryotic genomes remains enigmatic, and more studies in this area are needed in order to understand the crosstalk between RNA and DNA. The study of the association of RTs to novel putative systems open many doors in the understanding of their biological role, and can illuminate on how RT operate through RNA intermediaries, and what are the possible roles of cDNAs in bacterial cells.

4. Some of the described systems are susceptible to be biotechnologically exploited, such as DNA polymerase/RT tandem or HEPN systems, but mechanistic studies need to be done previous to that point.

7. References

AlQuraishi, M. 2019. “AlphaFold at CASP13.”. *Bioinformatics*.

Anantharaman, V, KS Makarova, AM Burroughs, EV Koonin, and L Aravind. 2013. “Comprehensive Analysis of the HEPN Superfamily: Identification of Novel Roles in Intra-Genomic Conflicts, Defense, Pathogenesis and RNA Processing.”. *Biol Direct* 8: 15.

Arambula, D, W Wong, BA Medhekar, H Guo, M Gingery, E Czornyj, M Liu, S Dey, P Ghosh, and JF Miller. 2013. “Surface Display of a Massively Variable Lipoprotein by a Legionella Diversity-Generating Retroelement.”. *Proc Natl Acad Sci U S A* 110: 8212–17.

Baltimore, D. 1970. “RNA-Dependent DNA Polymerase in Virions of RNA Tumour Viruses.”. *Nature* 226: 1209–11.

Bannert, N, and R Kurth. 2004. “Retroelements and the Human Genome: New Perspectives on an Old Relation.”. *Proc Natl Acad Sci U S A* 101 Suppl 2: 14572–79.

Barrangou, R, C Fremaux, H Deveau, M Richards, P Boyaval, S Moineau, DA Romero, and P Horvath. 2007. “CRISPR Provides Acquired Resistance against Viruses in Prokaryotes.”. *Science* 315: 1709–12.

Bernardes, de Jesus B, and MA Blasco. 2013. “Telomerase at the Intersection of Cancer and Aging.”. *Trends Genet* 29: 513–20.

Bezginov, A, GW Clark, RL Charlebois, VU Dar, and ER Tillier. 2013. “Coevolution Reveals a Network of Human Proteins Originating with Multicellularity.”. *Mol Biol Evol* 30: 332–46.

Bileschi, M. L., Belanger, D., Bryant, D. H., Sanderson, T., Carter, B., Sculley, D., ... & Colwell, L. J. (2019). Using Deep Learning to Annotate the Protein Universe. bioRxiv, 626507.

Burroughs, AM, D Zhang, DE Schäffer, LM Iyer, and L Aravind. 2015. “Comparative Genomic Analyses Reveal a Vast, Novel Network of Nucleotide-Centric Systems in Biological Conflicts, Immunity and Signaling.”. *Nucleic Acids Res* 43: 10633–54.

Casari, G, C Sander, and A Valencia. 1995. “A Method to Predict Functional Residues in Proteins.”. *Nat Struct Biol* 2: 171–78.

Chang, JW, YQ Zhou, Qamar MT Ul, LL Chen, and YD Ding. 2016. “Prediction of Protein-Protein Interactions by Evidence Combining Methods.”. *Int J Mol Sci* 17.

Chatzidaki-Livanis, M, MJ Coyne, KG Roelofs, RR Gentyala, JM Caldwell, and LE Comstock. 2017. “Gut Symbiont *Bacteroides Fragilis* Secretes a Eukaryotic-Like Ubiquitin Protein That Mediates Intraspecies Antagonism.”. *MBio* 8.

Coffin, JM, and H Fan. 2016. “The Discovery of Reverse Transcriptase.”. *Annu Rev Virol* 3: 29–51.

Dai, L, and S Zimmerly. 2003. “ORF-Less and Reverse-Transcriptase-Encoding Group II Introns in Archaebacteria, with a Pattern of Homing into Related Group II Intron ORFs.”. *RNA* 9: 14–19.

Date, SV. 2008. “The Rosetta Stone Method.”. *Methods Mol Biol* 453: 169–80.

Devos, DP, C Jogler, and JA Fuerst. 2013. “The 1st EMBO Workshop on PVC Bacteria-Planctomycetes-Verrucomicrobia-Chlamydiae Superphylum: Exceptions to the Bacterial Definition?”. *Antonie Van Leeuwenhoek* 104: 443–49.

Dong, S, and NJ Provart. 2018. “Analyses of Protein Interaction Networks Using Computational Tools.”. *Methods Mol Biol* 1794: 97–117.

Doulatov, S, A Hodes, L Dai, N Mandhana, M Liu, R Deora, RW Simons, S Zimmerly, and JF Miller. 2004. “Tropism Switching in *Bordetella Bacteriophage* Defines a Family of Diversity-Generating Retroelements.”. *Nature* 431: 476–81.

Durmaz, E, and TR Klaenhammer. 2007. “Abortive Phage Resistance Mechanism AbiZ Speeds the Lysis Clock to Cause Premature Lysis of Phage-Infected *Lactococcus Lactis*.”. *J Bacteriol* 189: 1417–25.

Emond, E, BJ Holler, I Boucher, PA Vandenbergh, ER Vedamuthu, JK Kondo, and S Moineau. 1997. “Phenotypic and Genetic Characterization of the Bacteriophage Abortive Infection Mechanism AbiK from *Lactococcus Lactis*.”. *Appl Environ Microbiol* 63: 1274–83.

Enright, AJ, I Iliopoulos, NC Kyrpides, and CA Ouzounis. 1999. “Protein Interaction Maps for Complete Genomes Based on Gene Fusion Events.”. *Nature* 402: 86–90.

Fernández-Recio, J, M Totrov, and R Abagyan. 2004. “Identification of Protein-Protein Interaction Sites from Docking Energy Landscapes.”. *J Mol Biol* 335: 843–65.

Finnegan, DJ. 2012. “Retrotransposons.”. *Curr Biol* 22: R432–7.

Fortier, LC, JD Bouchard, and S Moineau. 2005. “Expression and Site-Directed Mutagenesis of the Lactococcal Abortive Phage Infection Protein AbiK.”. *J Bacteriol* 187: 3721–30.

Gogvadze, E, and A Buzdin. 2009. “Retroelements and Their Impact on Genome Evolution and Functioning.”. *Cell Mol Life Sci* 66: 3727–42.

Goh, CS, AA Bogan, M Joachimiak, D Walther, and FE Cohen. 2000. “Co-Evolution of Proteins with Their Interaction Partners.”. *J Mol Biol* 299: 283–93.

González-Delgado, A., **Mestre, M. R**., Martínez-Abarca, F., & Toro, N. 2019. Spacer acquisition from RNA mediated by a natural reverse transcriptase-Cas1 fusion protein associated with a type III-D CRISPR–Cas system in *Vibrio vulnificus*. Nucleic Acids Research.

Handa, S, Y Jiang, S Tao, R Foreman, RF Schinazi, JF Miller, and P Ghosh. 2018. “Template-Assisted Synthesis of Adenine-Mutagenized CDNA by a Retroelement Protein Complex.”. *Nucleic Acids Res* 46: 9711–25.

Hashemifar, S, B Neyshabur, AA Khan, and J Xu. 2018. “Predicting Protein-Protein Interactions through Sequence-Based Deep Learning.”. *Bioinformatics* 34: i802–i810.

Hayashi, T, Y Matsuzaki, K Yanagisawa, M Ohue, and Y Akiyama. 2018. “MEGADOCK-Web: an Integrated Database of High-Throughput Structure-Based Protein-Protein Interaction Predictions.”. *BMC Bioinformatics* 19: 62.

Hew, BE, R Sato, D Mauro, I Stoytchev, and JB Owens. 2019. “RNA-Guided *PiggyBac* Transposition in Human Cells.”. *Synth Biol (Oxf)* 4: ysz018.

Hill, C, LA Miller, and TR Klaenhammer. 1990. “Nucleotide Sequence and Distribution of the pTR2030 Resistance Determinant (Hsp) Which Aborts Bacteriophage Infection in Lactococci.”. *Appl Environ Microbiol* 56: 2255–58.

Hille, F, H Richter, SP Wong, M Bratovič, S Ressel, and E Charpentier. 2018. “The Biology of CRISPR-Cas: Backward and Forward.”. *Cell* 172: 1239–59.

Hsu, MY, M Inouye, and S Inouye. 1990. “Retron for the 67-Base Multicopy Single-Stranded DNA from Escherichia Coli: a Potential Transposable Element Encoding Both Reverse Transcriptase and Dam Methylase Functions.”. *Proc Natl Acad Sci U S A* 87: 9454–58.

Inouye, M. 2017. “The First Demonstration of the Existence of Reverse Transcriptases in Bacteria.”. *Gene* 597: 76–77.

Inouye, S, MY Hsu, A Xu, and M Inouye. 1999. “Highly Specific Recognition of Primer RNA Structures for 2’-OH Priming Reaction by Bacterial Reverse Transcriptases.”. *J Biol Chem* 274: 31236–44.

Iyer, LM, EV Koonin, and L Aravind. 2002. “Extensive Domain Shuffling in Transcription Regulators of DNA Viruses and Implications for the Origin of Fungal APSES Transcription Factors.”. *Genome Biol* 3: RESEARCH0012.

Jackson, SA, RE McKenzie, RD Fagerlund, SN Kieper, PC Fineran, and SJ Brouns. 2017. “CRISPR-Cas: Adapting to Change.”. *Science* 356.

Jain, A. K., M. N. Murty, and P. J. Flynn. 1999. “Data Clustering: a Review”. *ACM Computing Surveys* 31 (3): 264–323.

Jung, H, J Liang, Y Jung, and D Lim. 2015. “Characterization of Cell Death in Escherichia Coli Mediated by XseA, a Large Subunit of Exonuclease VII.”. *J Microbiol* 53: 820–28.

Kanehisa, M, S Goto, Y Sato, M Furumichi, and M Tanabe. 2012. “KEGG for Integration and Interpretation of Large-Scale Molecular Data Sets.”. *Nucleic Acids Res* 40: D109–14.

Kojima, KK, and M Kanehisa. 2008. “Systematic Survey for Novel Types of Prokaryotic Retroelements Based on Gene Neighborhood and Protein Architecture.”. *Mol Biol Evol* 25: 1395–1404.

Kotlyar, M, C Pastrello, F Pivetta, Sardo A Lo, C Cumbaa, H Li, T Naranian, et al. 2015. “In Silico Prediction of Physical Protein Interactions and Characterization of Interactome Orphans.”. *Nat Methods* 12: 79–84.

Kulikov, E., Golomidova, A., Letarova, M., Kostryukova, E., Zelenin, A., Prokhorov, N., & Letarov, A. (2014). Genomic sequencing and biological characteristics of a novel Escherichia coli bacteriophage 9g, a putative representative of a new Siphoviridae genus. Viruses, 6(12), 5077-5092.

Lagkouvardos, I, MA Jehl, T Rattei, and M Horn. 2014. “Signature Protein of the PVC Superphylum.”. *Appl Environ Microbiol* 80: 440–45.

Lambowitz, AM, and S Zimmerly. 2004. “Mobile Group II Introns.”. *Annu Rev Genet* 38: 1–35.

Lampson, BC, M Inouye, and S Inouye. 1989. “Reverse Transcriptase with Concomitant Ribonuclease H Activity in the Cell-Free Synthesis of Branched RNA-Linked MsDNA of Myxococcus Xanthus.”. *Cell* 56: 701–7.

Lampson, BC, M Inouye, and S Inouye. 2005. “Retrons, MsDNA, and the Bacterial Genome.”. *Cytogenet Genome Res* 110: 491–99.

Lemay, DG, WF Martin, AS Hinrichs, M Rijnkels, JB German, I Korf, and KS Pollard. 2012. “G-NEST: a Gene Neighborhood Scoring Tool to Identify Co-Conserved, Co-Expressed Genes.”. *BMC Bioinformatics* 13: 253.

Li, Y, S Wang, R Umarov, B Xie, M Fan, L Li, and X Gao. 2018. “DEEPre: Sequence-Based Enzyme EC Number Prediction by Deep Learning.”. *Bioinformatics* 34: 760–69.

Lim, D, and WK Maas. 1989. “Reverse Transcriptase-Dependent Synthesis of a Covalently Linked, Branched DNA-RNA Compound in E. Coli B.”. *Cell* 56: 891–904.

Maas, WK, C Wang, T Lima, A Hach, and D Lim. 1996. “Multicopy Single-Stranded DNA of Escherichia Coli Enhances Mutation and Recombination Frequencies by Titrating MutS Protein.”. *Mol Microbiol* 19: 505–9.

Maas, WK, C Wang, T Lima, G Zubay, and D Lim. 1994. “Multicopy Single-Stranded DNAs with Mismatched Base Pairs Are Mutagenic in Escherichia Coli.”. *Mol Microbiol* 14: 437–41.

Makarova, KS, V Anantharaman, NV Grishin, EV Koonin, and L Aravind. 2014. “CARF and WYL Domains: Ligand-Binding Regulators of Prokaryotic Defense Systems.”. *Front Genet* 5: 102.

Makarova, KS, YI Wolf, S Snir, and EV Koonin. 2011. “Defense Islands in Bacterial and Archaeal Genomes and Prediction of Novel Defense Systems.”. *J Bacteriol* 193: 6039–56.

Makarova, KS, and EV Koonin. 2015. “Annotation and Classification of CRISPR-Cas Systems.”. *Methods Mol Biol* 1311: 47–75.

Makarova, KS, YI Wolf, and EV Koonin. 2018. “Classification and Nomenclature of CRISPR-Cas Systems: Where from Here?”. *CRISPR J* 1: 325–36.

Malik, HS, WD Burke, and TH Eickbush. 1999. “The Age and Evolution of Non-LTR Retrotransposable Elements.”. *Mol Biol Evol* 16: 793–805.

Marquet, R, C Isel, C Ehresmann, and B Ehresmann. 1995. “TRNAs as Primer of Reverse Transcriptases.”. *Biochimie* 77: 113–24.

Martínez-Abarca, F, S Zekri, and N Toro. 1998. “Characterization and Splicing in Vivo of a Sinorhizobium Meliloti Group II Intron Associated with Particular Insertion Sequences of the IS630-Tc1/IS3 Retroposon Superfamily.”. *Mol Microbiol* 28: 1295–1306.

Martínez-Abarca, F, A Barrientos-Durán, M Fernández-López, and N Toro. 2004. “The RmInt1 Group II Intron Has Two Different Retrohoming Pathways for Mobility Using Predominantly the Nascent Lagging Strand at DNA Replication Forks for Priming.”. *Nucleic Acids Res* 32: 2880–88.

Mattoo, S, AK Foreman-Wykert, PA Cotter, and JF Miller. 2001. “Mechanisms of Bordetella Pathogenesis.”. *Front Biosci* 6: E168–86.

Melamed, D, DL Young, CR Miller, and S Fields. 2015. “Combining Natural Sequence Variation with High Throughput Mutational Data to Reveal Protein Interaction Sites.”. *PLoS Genet* 11: e1004918.

Menéndez-Arias, L, A Sebastián-Martín, and M Álvarez. 2017. “Viral Reverse Transcriptases.”. *Virus Res* 234: 153–76.

Michel, F, and JL Ferat. 1995. “Structure and Activities of Group II Introns.”. *Annu Rev Biochem* 64: 435–61.

Min, S, B Lee, and S Yoon. 2017. “Deep Learning in Bioinformatics.”. *Brief Bioinform* 18: 851–69.

Mohr, G, S Silas, JL Stamos, KS Makarova, LM Markham, J Yao, P Lucas-Elío, et al. 2018. “A Reverse Transcriptase-Cas1 Fusion Protein Contains a Cas6 Domain Required for Both CRISPR RNA Biogenesis and RNA Spacer Acquisition.”. *Mol Cell* 72: 700–714.e8.

Murakami, Y, and S Jones. 2006. “SHARP2: Protein-Protein Interaction Predictions Using Patch Analysis.”. *Bioinformatics* 22: 1794–95.

Nicholson, Joshua M, Joana C Macedo, Aaron J Mattingly, Darawalee Wangsa, Jordi Camps, Vera Lima, Ana M Gomes, et al. 2015. “Chromosome Mis-Segregation and Cytokinesis Failure in Trisomic Human Cells”. *ELife* 4 (May).

Nimkulrat, S, H Lee, TG Doak, and Y Ye. 2016. “Genomic and Metagenomic Analysis of Diversity-Generating Retroelements Associated with Treponema Denticola.”. *Front Microbiol* 7: 852.

Niu, Y, C Liu, S Moghimyfiroozabad, Y Yang, and KN Alavian. 2017. “PrePhyloPro: Phylogenetic Profile-Based Prediction of Whole Proteome Linkages.”. *PeerJ* 5: e3712.

Ochoa, D, and F Pazos. 2010. “Studying the Co-Evolution of Protein Families with the Mirrortree Web Server.”. *Bioinformatics* 26: 1370–71.

Ochoa, D, and F Pazos. 2014. “Practical Aspects of Protein Co-Evolution.”. *Front Cell Dev Biol* 2: 14.

Ochoa, D, D Juan, A Valencia, and F Pazos. 2015. “Detection of Significant Protein Coevolution.”. *Bioinformatics* 31: 2166–73.

Odegrip, R, AS Nilsson, and E Haggård-Ljungquist. 2006. “Identification of a Gene Encoding a Functional Reverse Transcriptase within a Highly Variable Locus in the P2-like Coliphages.”. *J Bacteriol* 188: 1643–47.

Papanikolaou, N, GA Pavlopoulos, T Theodosiou, and I Iliopoulos. 2015. “Protein-Protein Interaction Predictions Using Text Mining Methods.”. *Methods* 74: 47–53.

Paul, BG, D Burstein, CJ Castelle, S Handa, D Arambula, E Czornyj, BC Thomas, et al. 2017. “Retroelement-Guided Protein Diversification Abounds in Vast Lineages of Bacteria and Archaea.”. *Nat Microbiol* 2: 17045.

Pazos, F, M Helmer-Citterich, G Ausiello, and A Valencia. 1997. “Correlated Mutations Contain Information about Protein-Protein Interaction.”. *J Mol Biol* 271: 511–23.

Pazos, F, and A Valencia. 2001. “Similarity of Phylogenetic Trees as Indicator of Protein-Protein Interaction.”. *Protein Eng* 14: 609–14.

Pazos, F, and A Valencia. 2002. “In Silico Two-Hybrid System for the Selection of Physically Interacting Protein Pairs.”. *Proteins* 47: 219–27.

Pazos, F, JA Ranea, D Juan, and MJ Sternberg. 2005. “Assessing Protein Co-Evolution in the Context of the Tree of Life Assists in the Prediction of the Interactome.”. *J Mol Biol* 352: 1002–15.

Pazos, F, and A Valencia. 2008. “Protein Co-Evolution, Co-Adaptation and Interactions.”. *EMBO J* 27: 2648–55.

Pellegrini, M. 2012. “Using Phylogenetic Profiles to Predict Functional Relationships.”. *Methods Mol Biol* 804: 167–77.

Rivas-Marín, E, I Canosa, and DP Devos. 2016. “Evolutionary Cell Biology of Division Mode in the Bacterial *Planctomycetes-Verrucomicrobia-Chlamydiae* Superphylum.”. *Front Microbiol* 7: 1964.

Rivas-Marín, E, and DP Devos. 2018. “The Paradigms They Are a-Changin’: Past, Present and Future of PVC Bacteria Research.”. *Antonie Van Leeuwenhoek* 111: 785–99.

Rodionov, A, A Bezginov, J Rose, and ER Tillier. 2011. “A New, Fast Algorithm for Detecting Protein Coevolution Using Maximum Compatible Cliques.”. *Algorithms Mol Biol* 6: 17.

Ryazansky, S, A Kulbachinskiy, and AA Aravin. 2018. “The Expanded Universe of Prokaryotic Argonaute Proteins.”. *MBio* 9.

San, Filippo J, and AM Lambowitz. 2002. “Characterization of the C-Terminal DNA-Binding/DNA Endonuclease Region of a Group II Intron-Encoded Protein.”. *J Mol Biol* 324: 933–51.

Schmidt, F, MY Cherepkova, and RJ Platt. 2018. “Transcriptional Recording by CRISPR Spacer Acquisition from RNA.”. *Nature* 562: 380–85.

Shah, SA, OS Alkhnbashi, J Behler, W Han, Q She, WR Hess, RA Garrett, and R Backofen. 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: 530–42.

Sharifi, F, and Y Ye. 2019. “MyDGR: a Server for Identification and Characterization of Diversity-Generating Retroelements.”. *Nucleic Acids Res* 47: W289–W294.

Shatkay, H, S Brady, and A Wong. 2015. “Text as Data: Using Text-Based Features for Proteins Representation and for Computational Prediction of Their Characteristics.”. *Methods* 74: 54–64.

Shmakov, SA, KS Makarova, YI Wolf, KV Severinov, and EV Koonin. 2018. “Systematic Prediction of Genes Functionally Linked to CRISPR-Cas Systems by Gene Neighborhood Analysis.”. *Proc Natl Acad Sci U S A* 115: E5307–E5316.

Shoemaker, BA, and AR Panchenko. 2007. “Deciphering Protein-Protein Interactions. Part I. Experimental Techniques and Databases.”. *PLoS Comput Biol* 3: e42.

Silas, S, G Mohr, DJ Sidote, LM Markham, A Sanchez-Amat, D Bhaya, AM Lambowitz, and AZ Fire. 2016. “Direct CRISPR Spacer Acquisition from RNA by a Natural Reverse Transcriptase-Cas1 Fusion Protein.”. *Science* 351: aad4234.

Silas, S, KS Makarova, S Shmakov, D Páez-Espino, G Mohr, Y Liu, M Davison, et al. 2017. “On the Origin of Reverse Transcriptase-Using CRISPR-Cas Systems and Their Hyperdiverse, Enigmatic Spacer Repertoires.”. *MBio* 8.

Simon, DM, and S Zimmerly. 2008. “A Diversity of Uncharacterized Reverse Transcriptases in Bacteria.”. *Nucleic Acids Res* 36: 7219–29.

Skrabanek, L, HK Saini, GD Bader, and AJ Enright. 2008. “Computational Prediction of Protein-Protein Interactions.”. *Mol Biotechnol* 38: 1–17.

Škunca, N, and C Dessimoz. 2015. “Phylogenetic Profiling: How Much Input Data Is Enough?”. *PLoS One* 10: e0114701.

Staroń, A, HJ Sofia, S Dietrich, LE Ulrich, H Liesegang, and T Mascher. 2009. “The Third Pillar of Bacterial Signal Transduction: Classification of the Extracytoplasmic Function (ECF) Sigma Factor Protein Family.”. *Mol Microbiol* 74: 557–81.

Steinegger, M, and J Söding. 2017. “MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets.”. *Nat Biotechnol* 35: 1026–28.

Subramani, S, R Kalpana, PM Monickaraj, and J Natarajan. 2015. “HPIminer: A Text Mining System for Building and Visualizing Human Protein Interaction Networks and Pathways.”. *J Biomed Inform* 54: 121–31.

Sun, T, B Zhou, L Lai, and J Pei. 2017. “Sequence-Based Prediction of Protein Protein Interaction Using a Deep-Learning Algorithm.”. *BMC Bioinformatics* 18: 277.

Szklarczyk, D, AL Gable, D Lyon, A Junge, S Wyder, J Huerta-Cepas, M Simonovic, et al. 2019. “STRING v11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets.”. *Nucleic Acids Res* 47: D607–D613.

Talavera, D, SC Lovell, and S Whelan. 2015. “Covariation Is a Poor Measure of Molecular Coevolution.”. *Mol Biol Evol* 32: 2456–68.

Tangney, M, and GF Fitzgerald. 2002. “Effectiveness of the Lactococcal Abortive Infection Systems AbiA, AbiE, AbiF and AbiG against P335 Type Phages.”. *FEMS Microbiol Lett* 210: 67–72.

Temin, HM, and S Mizutani. 1970. “RNA-Dependent DNA Polymerase in Virions of Rous Sarcoma Virus.”. *Nature* 226: 1211–13.

Thiaville, J. J., Kellner, S. M., Yuan, Y., Hutinet, G., Thiaville, P. C., Jumpathong, W., & Malik, C. K. (2016). Novel genomic island modifies DNA with 7-deazaguanine derivatives. Proceedings of the National Academy of Sciences, 113(11), E1452-E1459.

Tillier, ER, and RL Charlebois. 2009. “The Human Protein Coevolution Network.”. *Genome Res* 19: 1861–71.

Toro, N. 2003. “Bacteria and Archaea Group II Introns: Additional Mobile Genetic Elements in the Environment.”. *Environ Microbiol* 5: 143–51.

Toro, N, and R Nisa-Martínez. 2014. “Comprehensive Phylogenetic Analysis of Bacterial Reverse Transcriptases.”. *PLoS One* 9: e114083.

Toro, N, JI Jiménez-Zurdo, and FM García-Rodríguez. 2007. “Bacterial Group II Introns: Not Just Splicing.”. *FEMS Microbiol Rev* 31: 342–58.

Toro, N., & Nisa-Martínez, R. (2014). Comprehensive phylogenetic analysis of bacterial reverse transcriptases. PLoS One, 9(11), e114083.

Toro, N, F Martínez-Abarca, and A González-Delgado. 2017. “The Reverse Transcriptases Associated with CRISPR-Cas Systems.”. *Sci Rep* 7: 7089.

Toro, N, F Martínez-Abarca, A González-Delgado, and **MR Mestre**. 2018. “On the Origin and Evolutionary Relationships of the Reverse Transcriptases Associated With Type III CRISPR-Cas Systems.”. *Front Microbiol* 9: 1317.

Toro, N, F Martínez-Abarca, **MR Mestre**, and A González-Delgado. 2019. “Multiple Origins of Reverse Transcriptases Linked to CRISPR-Cas Systems.”. *RNA Biol*, 1–8.

Tsai, FS. 2011. “Text Mining and Visualisation of Protein-Protein Interactions.”. *Int J Comput Biol Drug Des* 4: 239–44.

Vella, D, I Zoppis, G Mauri, P Mauri, and Silvestre D Di. 2017. “From Protein-Protein Interactions to Protein Co-Expression Networks: a New Perspective to Evaluate Large-Scale Proteomic Data.”. *EURASIP J Bioinform Syst Biol* 2017: 6.

Wang, C, M Villion, C Semper, C Coros, S Moineau, and S Zimmerly. 2011. “A Reverse Transcriptase-Related Protein Mediates Phage Resistance and Polymerizes Untemplated DNA in Vitro.”. *Nucleic Acids Res* 39: 7620–29.

Wang, J, L Zhang, L Jia, Y Ren, and G Yu. 2017. “Protein-Protein Interactions Prediction Using a Novel Local Conjoint Triad Descriptor of Amino Acid Sequences.”. *Int J Mol Sci* 18.

Wank, H, J SanFilippo, RN Singh, M Matsuura, and AM Lambowitz. 1999. “A Reverse Transcriptase/Maturase Promotes Splicing by Binding at Its Own Coding Segment in a Group II Intron RNA.”. *Mol Cell* 4: 239–50.

Weinberg, Z, CE Lünse, KA Corbino, TD Ames, JW Nelson, A Roth, KR Perkins, ME Sherlock, and RR Breaker. 2017. “Detection of 224 Candidate Structured RNAs by

Comparative Analysis of Specific Subsets of Intergenic Regions.". *Nucleic Acids Res* 45: 10811–23.

Wu, L, M Gingery, M Abebe, D Arambula, E Czornyj, S Handa, H Khan, et al. 2018. "Diversity-Generating Retroelements: Natural Variation, Classification and Evolution Inferred from a Large-Scale Genomic Survey.". *Nucleic Acids Res* 46: 11–24.

Xie, X, and R Yang. 2017. "Multi-Copy Single-Stranded DNA in Escherichia Coli.". *Microbiology* 163: 1735–39.

Xiong, Y, and TH Eickbush. 1990. "Origin and Evolution of Retroelements Based upon Their Reverse Transcriptase Sequences.". *EMBO J* 9: 3353–62.

Ye, Y. 2014. "Identification of Diversity-Generating Retroelements in Human Microbiomes.". *Int J Mol Sci* 15: 14234–46.

Yin, C, and SS Yau. 2017. "A Coevolution Analysis for Identifying Protein-Protein Interactions by Fourier Transform.". *PLoS One* 12: e0174862.

Zhang, M, Q Su, Y Lu, M Zhao, and B Niu. 2017. "Application of Machine Learning Approaches for Protein-Protein Interactions Prediction.". *Med Chem* 13: 506–14.

Zhang, Q, TG Doak, and Y Ye. 2014. "Expanding the Catalog of Cas Genes with Metagenomes.". *Nucleic Acids Res* 42: 2448–59.

Zhang, QC, D Petrey, JI Garzón, L Deng, and B Honig. 2013. "PrePPI: a Structure-Informed Database of Protein-Protein Interactions.". *Nucleic Acids Res* 41: D828–33.

Zhu, J, S Wang, D Bu, and J Xu. 2018. "Protein Threading Using Residue Co-Variation and Deep Learning.". *Bioinformatics* 34: i263–i273.

Zimmerly, S, H Guo, PS Perlman, and AM Lambowitz. 1995. "Group II Intron Mobility Occurs by Target DNA-Primed Reverse Transcription.". *Cell* 82: 545–54.

Zimmerly, S, and C Semper. 2015. "Evolution of Group II Introns.". *Mob DNA* 6: 7.

Zimmerly, S, and L Wu. 2015. "An Unexplored Diversity of Reverse Transcriptases in Bacteria.". *Microbiol Spectr* 3: MDNA3–0058-2014.

Appendix I: Putative systems found in UG groups

Group/Cluster	Repr. Sequence	Start	End	Size
UG28				
RT1301	WP_079700534.1	9117	9129	12
UG27				
RT1931	fig|445970.5.peg.914	9071	9083	12
RT2783	fig|1339346.3.peg.3671			
RT896	WP_097786714.1			
RT2166	fig|1339346.3.peg.3667			
RT2244	fig|1121096.3.peg.2728			
RT2893	fig|1121097.3.peg.1529			
RT3246	fig|564423.8.peg.1659			
RT4709	fig|1950729.3.peg.1566			
RT4776	fig|445970.5.peg.927			
RT4846	fig|1121096.3.peg.2684			
UG26				
RT3503	WP_093121914.1	9045	9055	10
UG2				
RT2202	WP_091904160.1	8939	8951	12
RT725	fig|1952220.3.peg.892			
RT3	fig|2024833.5.peg.3515			
RT21	fig|853.15.peg.1691			
RT62	fig|1763509.5.peg.892			
RT92	fig|1947945.3.peg.533			
RT115	WP_004150801.1			
RT162	fig|1217627.3.peg.1008			

RT181	fig|1797504.3.peg.917			
RT192	WP_000798650.1			
RT218	fig|1332071.4.peg.4242			
RT310	fig|698758.3.peg.1230			
RT551	fig|1214190.3.peg.2028			
RT933	CBH38826.1			
RT1011	CUN32826.1			
RT1060	WP_092985463.1			
RT3673	fig|1382798.3.peg.2453			
RT3842	fig|732242.4.peg.252			
RT5007	WP_068359943.1			
RT5097	fig|1382798.3.peg.2461			
RT470	WP_132275010.1	8921	8938	17

UG3

RT146	ADT87872.1	8769	8857	88
RT66	fig|2004648.3.peg.1751			
RT5	WP_092169221.1			
RT6	fig|1947874.3.peg.3490			

UG17

RT269	WP_109615051.1	8675	8768	93
RT1640	WP_041928575.1			

UG25

RT2632	WP_077844617.1	8670	8674	4

UG26

RT2534	WP_029811522.1	8646	8651	5

UG5

RT3966	WP_075384214.1	8614	8619	5

UG7

RT629	ANM28510.1	8499	8537	38

UG9

RT161	WP_029507943.1	8485	8498	13
RT6	fig|1947874.3.peg.3490			
RT245	fig|1550240.3.peg.1907			

UG22

RT3633	WP_092555623.1	8441	8448	7

UG8

RT164	WP_071843470.1	8304	8385	81

Appendix II: Genomes harboring Clade 3 RT-CRISPR/Cas systems

Genome_partition	Genome_name	Domain	Phylum	Genus
NZ_RSCN01000022.1	Nostocaceae	Bacteria	Cyanobacteria	
NC_019751.1	Calothrix parietina	Bacteria	Cyanobacteria	Calothrix
NZ_ALWD01000002.1	[Scytonema hofmanni] UTEX B 1581	Bacteria	Cyanobacteria	Tolypothrix
NZ_KQ976354.1	Scytonema hofmannii	Bacteria	Cyanobacteria	Scytonema
NZ_MNPM02000006.1	Mastigocladus laminosus	Bacteria	Cyanobacteria	Mastigocladus
CP003660.1	Anabaena cylindrica PCC 7122	Bacteria	Cyanobacteria	Anabaena
NZ_MRCE01000026.1	Phormidium ambiguum	Bacteria	Cyanobacteria	Phormidium
NZ_KB235908.1	Kamptonema	Bacteria	Cyanobacteria	
NC_019740.1	Microcoleus sp. PCC 7113	Bacteria	Cyanobacteria	Microcoleus
NZ_FSSI02000066.1	Phormidium sp. HE10JO	Bacteria	Cyanobacteria	Phormidium
NZ_CP021983.2	Halomicronema hongdechloris	Bacteria	Cyanobacteria	Halomicronema
NZ_KI913950.1	Leptolyngbya sp. PCC 6406	Bacteria	Cyanobacteria	Leptolyngbya
LJZR01000039.1	Phormidesmis priestleyi Ana	Bacteria	Cyanobacteria	Phormidesmis
NZ_AQPY01001269.1	Microcystis aeruginosa	Bacteria	Cyanobacteria	Microcystis
NC_014533.1	Gloeothece verrucosa	Bacteria	Cyanobacteria	Gloeothece
NZ_KV878782.1	Spirulina major	Bacteria	Cyanobacteria	Spirulina

Appendix III: Genomes harboring G2L/Que systems

Genome_partition	Genome_name	Domain	Phylum	Genus
LN850107.1	Alloactinosynnema sp. L-07	Bacteria	Actinobacteria	Alloactinosynnema
NZ_FQVN01000004.1	Streptoalloteichus hindustanus	Bacteria	Actinobacteria	Streptoalloteichus
NZ_MQUP01000004.1	Amycolatopsis keratiniphila	Bacteria	Actinobacteria	Amycolatopsis
CP022521.1	Actinoalloteichus hoggarensis	Bacteria	Actinobacteria	Actinoalloteichus
NZ_MKKE01000044.1	Saccharomonospora sp. CUA-673	Bacteria	Actinobacteria	Saccharomonospora
CPWZ01000154.1	Mycobacterium tuberculosis	Bacteria	Actinobacteria	Mycobacterium
LAXD01000001.1	Streptomyces thermoautotrophicus	Bacteria	Actinobacteria	Streptomyces
FR845719.1	Streptomyces venezuelae ATCC 10712	Bacteria	Actinobacteria	Streptomyces
NZ_CP007699.2	Streptomyces lydicus	Bacteria	Actinobacteria	Streptomyces
HF570108.1	Micromonospora lupini str. Lupac 08	Bacteria	Actinobacteria	Micromonospora
PJMT01000001.1	Streptomyces sp. GP55	Bacteria	Actinobacteria	Streptomyces