

REVIEW ARTICLE

Prokaryotic reverse transcriptases: from retroelements to specialized defense systems

Alejandro González-Delgado^{1,†}, Mario Rodríguez Mestre^{1,2,‡}, Francisco Martínez-Abarca^{1,#} and Nicolás Toro^{1,*,\$}

¹Department of Soil Microbiology and Symbiotic Systems, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas, Structure, Dynamics and Function of Rhizobacterial Genomes, Grupo de Ecología Genética de la Rizosfera, C/ Profesor Albareda 1, 18008 Granada, Spain and ²Department of Biochemistry, Universidad Autónoma de Madrid and Instituto de Investigaciones Biomédicas “Alberto Sols”, CSIC-UAM, Madrid, Spain

*Corresponding author: Department of Soil Microbiology and Symbiotic Systems, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas, Structure, Dynamics and Function of Rhizobacterial Genomes, Grupo de Ecología Genética de la Rizosfera, C/ Profesor Albareda 1, 18008 Granada, Spain. Tel: (+34) 958 181600; Fax: +34 958129600; E-mail: nicolas.toro@eez.csic.es

One sentence summary: Prokaryotic reverse transcriptases (RTs) are attracting considerable attention, both for use in cutting-edge technologies, such as genome editing, and as an emerging research topic. In this review, the authors discuss what is known about prokaryotic RTs, and the exciting evidence for their domestication to create specialized defense systems.

Editor: Suzana Salcedo

†Alejandro González-Delgado, <http://orcid.org/0000-0002-8357-3204>

‡Mario Rodríguez Mestre, <http://orcid.org/0000-0003-3695-5370>

#Francisco Martínez-Abarca, <http://orcid.org/0000-0003-0417-6217>

\$Nicolás Toro, <http://orcid.org/0000-0002-3614-807X>

ABSTRACT

Reverse transcriptases (RTs) catalyze the polymerization of DNA from an RNA template. These enzymes were first discovered in RNA tumor viruses in 1970, but it was not until 1989 that they were found in prokaryotes as a key component of retrons. Apart from RTs encoded by the ‘selfish’ mobile retroelements known as group II introns, prokaryotic RTs are extraordinarily diverse, but their function has remained elusive. However, recent studies have revealed that different lineages of prokaryotic RTs, including retrons, those associated with CRISPR-Cas systems, Abi-like RTs and other yet uncharacterized RTs, are key components of different lines of defense against phages and other mobile genetic elements. Prokaryotic RTs participate in various antiviral strategies, including abortive infection (Abi), in which the infected cell is induced to commit suicide to protect the host population, adaptive immunity, in which a memory of previous infection is used to build an efficient defense, and other as yet unidentified mechanisms. These prokaryotic enzymes are attracting considerable attention, both for use in cutting-edge technologies, such as genome editing, and as an emerging research topic. In this review, we discuss what is known about prokaryotic RTs, and the exciting evidence for their domestication from retroelements to create specialized defense systems.

Received: 15 January 2021; Accepted: 7 May 2021

© The Author(s) 2021. Published by Oxford University Press on behalf of FEMS. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Keywords: reverse transcriptases; group II introns; retrons; DGRs; CRISPR-Cas systems; Abi systems

INTRODUCTION

According to the central dogma of molecular biology proposed by Francis Crick in 1958, once genetic information was converted to proteins the flow could not be reversed. However, it left open the possibility that information could be transferred from RNA to DNA (Crick 1958). The discovery of enzymes capable of synthesizing DNA from an RNA template demonstrated this hypothesis (Crick 1970). RNA-dependent DNA polymerases, also known as reverse transcriptases (RTs), were discovered in 1970 in RNA tumor viruses (Baltimore 1970; Temin and Mizutani 1970), and this finding reshaped existing views on the functioning of all forms of life. In prokaryotes, RTs were discovered almost 20 years later as a key component of retrons (Lampson et al. 1989; Lim and Maas 1989). Currently, prokaryotic RTs display an extraordinary diversity, with most of them (80%) classified into three main groups: those encoded by group II introns, retron/retron-like sequences and diversity-generating retroelements (DGRs). The remaining RT sequences form distinct lineages, including those associated with CRISPR-Cas systems, abortive phage infection systems (Abi-like) and other uncharacterized RTs clustering in the so-called G2L (group-II-like) and unknown groups (UG; Toro et al. 2019a). While the RTs of eukaryotes and viruses have been characterized in detail, it has taken over 30 years for us to begin to understand the function and mechanisms of the highly diverse RTs of prokaryotic organisms (Fig. 1).

Recent investigations have shown that several prokaryotic RT lineages have evolved to provide new strategies for dealing with phages. Thus, together with Abi-RTs (Fortier, Bouchard and Moineau 2005; Odegrip, Nilsson and Haggård-Ljungquist 2006; Durmaz and Klaenhammer 2007) and CRISPR-RTs (Kojima and Kanehisa 2008; Toro and Nisa-Martínez 2014), searches for new antiviral immunity systems in defense islands have led to the discovery of anti-phage properties in a majority of retron types, and several RTs from uncharacterized groups (Gao et al. 2020; Millman et al. 2020). The relevance of these enzymes also lies in the use of different RT lineages as promising tools for diverse applications, including genome editing, recombination-mediated genetic engineering applications and the recording of biological information in bacterial genomes (Schmidt, Cherepkova and Platt 2018; Belfort and Lambowitz 2019; Simon, Ellington and Finkelstein 2019). These potential uses have boosted interest in these prokaryotic enzymes and their associated systems.

Group II intron-RTs are the best-characterized in prokaryotes (Ferat and Michel 1993; Lambowitz and Zimmerly 2011), but very little is known about their ecological role. These mobile retroelements are generally considered to be selfish units, but a few studies have provided evidence to suggest that they can disrupt other mobile genetic elements (MGE), participating in host defense against these potentially harmful elements (Chillón, Martínez-Abarca and Toro 2011; Qu et al. 2018). DGRs provide their hosts with an adaptive advantage due to the great sequence variation that they cause in a target gene through a mutagenic reverse transcription reaction (Liu et al. 2002; Wu et al. 2018). The wide range of functions performed by DGRs have yet to be elucidated, but they are known to participate in tropism switching and signaling pathways potentially involved in virus-host interactions. Thus, our current knowledge of prokaryotic RTs suggests that they tend to be domesticated to perform

immune functions in the host cell through a wide range of biological mechanisms that remain incompletely understood. The overall purpose of this review is to highlight the role of prokaryotic RTs in immunity to phages and MGEs mediated by selfish genetic elements to more specialized systems, and to summarize the use of these enzymes in cutting-edge technologies.

Groups of prokaryotic RTs

Group II introns

Group II introns were first identified in the mitochondrial and chloroplast genomes of lower eukaryotes and plants. They were not found in prokaryotes until 1993, and their mobility has since been characterized in detail (Ferat and Michel 1993; Michel and Ferat 1995; Dai and Zimmerly 2003; Toro 2003; Lambowitz and Zimmerly 2004). Group II introns are self-splicing RNAs that require an ancient form of RT to act as mobile retroelements. This type of RT is the most abundant form in bacteria, accounting for almost 50% of total RT diversity (Toro et al. 2019a). Bacterial group II introns generally insert into sequences outside of essential genes or into non-essential genes, and show an extinction-recolonization dynamics suggesting that they are deleterious to the host (Leclercq and Cordaux 2012). With few exceptions, these introns are typically found as only one or two copies per genome (Dai and Zimmerly 2002). This tendency for group II introns to be deleterious may drive the domestication of RT-containing genes, and their conversion into non-mobile RTs, presumably with novel functions (Fig. 2). Indeed, although group II introns have given rise to more complex retroelements such as eukaryotic non-LTR retrotransposons, they are also considered the evolutionary ancestors of the spliceosome complex and the telomerase, which represent great examples of how group II introns could generate new functions by domestication (Box 1).

Group II introns consist of a catalytic RNA, with characteristic conserved 5'- and 3'-end sequences, GUGYG and AY, respectively, resembling those of spliceosomal eukaryotic introns, and an intron-encoded protein (IEP; Lambowitz and Zimmerly 2011). Catalytic intron RNAs have a conserved secondary structure of 400–800 nts in length and organized into six domains, DIVI, radiating from a central ‘wheel’ (Michel, Costa and Westhof 2009). The IEPs, which are encoded in DIV, have several different domains involved in group II intron retromobility (Lambowitz and Zimmerly 2004): an N-terminal RT domain, an X domain or maturase involved in facilitating RNA splicing, a D domain involved in DNA binding and a metal-dependent DNA endonuclease domain of the HNH family that cleaves a target DNA strand to generate the primer for reverse transcription (San Filippo and Lambowitz 2002). However, a large number of bacterial group II introns encode IEPs without an endonuclease domain. The best studied of introns that lack this domain is the *Sinorhizobium meliloti* RmInt1, which uses a mechanism associated with DNA replication to prime reverse transcription (Martínez-Abarca et al. 2004; García-Rodríguez et al. 2019).

The mobility of group II introns requires RNA splicing via two transesterification reactions, the first of which starts with a nucleophilic attack of 2'-OH on a bulged adenosine in DIVI, with the second reaction resulting in the generation of an excised intron lariat and exon ligation (Fig. 3A). The IEP facilitates this stage, by acting as a maturase, with the RT and X

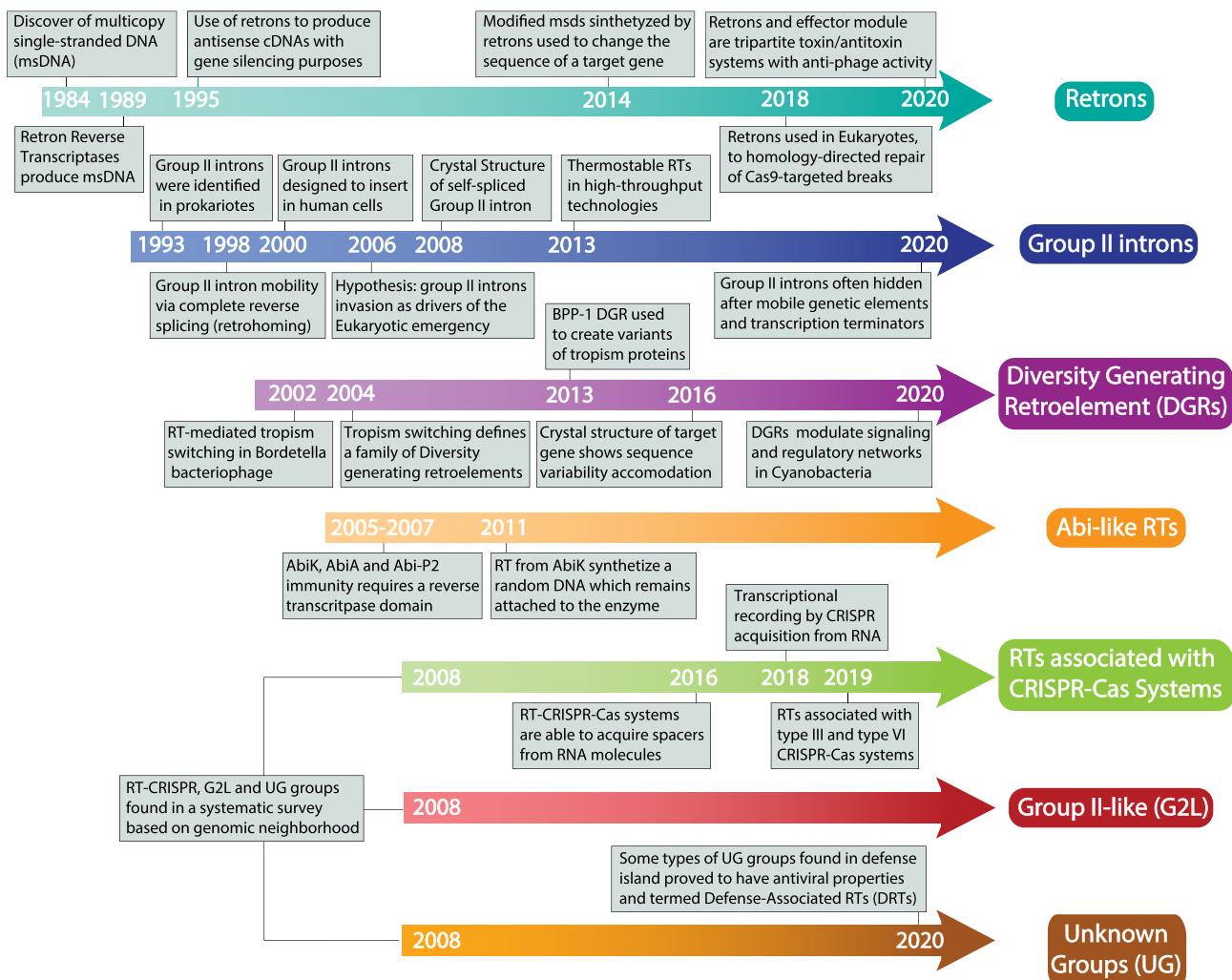


Figure 1. Timeline of research on prokaryotic RTs. Over the last almost 40 years since the discovery of a small DNA satellite in 1984, which turned out to be synthesized by an RT, research in this field has expanded spectacularly. Here, we show the main breakthroughs concerning the various functional groups of prokaryotic RTs.

domains functioning together to bind the intron RNA specifically, thereby promoting the formation of a stable ribonucleoprotein (RNP) complex (Belfort and Lambowitz 2019). This active ribozyme can migrate to new DNA targets in the host genome by a process involving a target-primed reverse transcription (TPRT) mechanism, in which the RT domain synthesizes the intron cDNA of the reverse-spliced intron RNA in one strand of a double-stranded-DNA target site (Lambowitz and Zimmerly 2011). Intron mobility can occur via retrotransposition, wherein the intron is introduced into ectopic sites, but the principal mobility pathway of group II introns is retrohoming, in which the intron is inserted into a target region of the host genome (Belfort and Lambowitz 2019).

Despite detailed characterization of the mobility pathways of group II introns, the ecological implications of these retroelements in host cells are poorly understood, with only a few studies published to date. Group II introns, which are considered to be selfish elements, tend to localize at higher densities on plasmids than on chromosomes and frequently hide in other MGEs, such as other group II introns, a broad range of transposases and some phage-related proteins, disrupting these elements and their functions (Waldern et al. 2020). In one study, following the acquisition of the RmInt1 group II intron by

conjugative transfer, the colonization of homing sites, typically insertion sequences (ISRm2011-2 and close homologs), was found to occur at high frequency via the preferred retrohoming pathway, with sites located on the template for lagging-strand synthesis invaded first, followed by those on the leading strand template (Nisa-Martínez et al. 2007). The splicing of RmInt1 naturally inserted into an IS interrupting a transposase gene is almost completely abolished, but this intron retains its invasion capacity, suggesting that group II introns may, to some extent, control the spread of other MGEs (Chillón, Martínez-Abarca and Toro 2011). These findings suggest that group II introns may have an evolutionary role in circumventing splicing, preventing the mobility of harmful elements in the bacterial cell and forming a particular defense system.

Consistent with this hypothesis, another group II intron integrated into a relaxase gene on a conjugative plasmid has been shown to inhibit host gene expression and to restrain the naturally cohabiting mobile element, preventing its conjugative horizontal transfer by decreasing the levels of spliced mRNA (Qu et al. 2018). This process seems to function as a defensive barrier, limiting the spread of other mobile elements acting as general inhibitors of gene expression. However, the relaxase stimulates intron dispersal by nicking the conjugative plasmid and

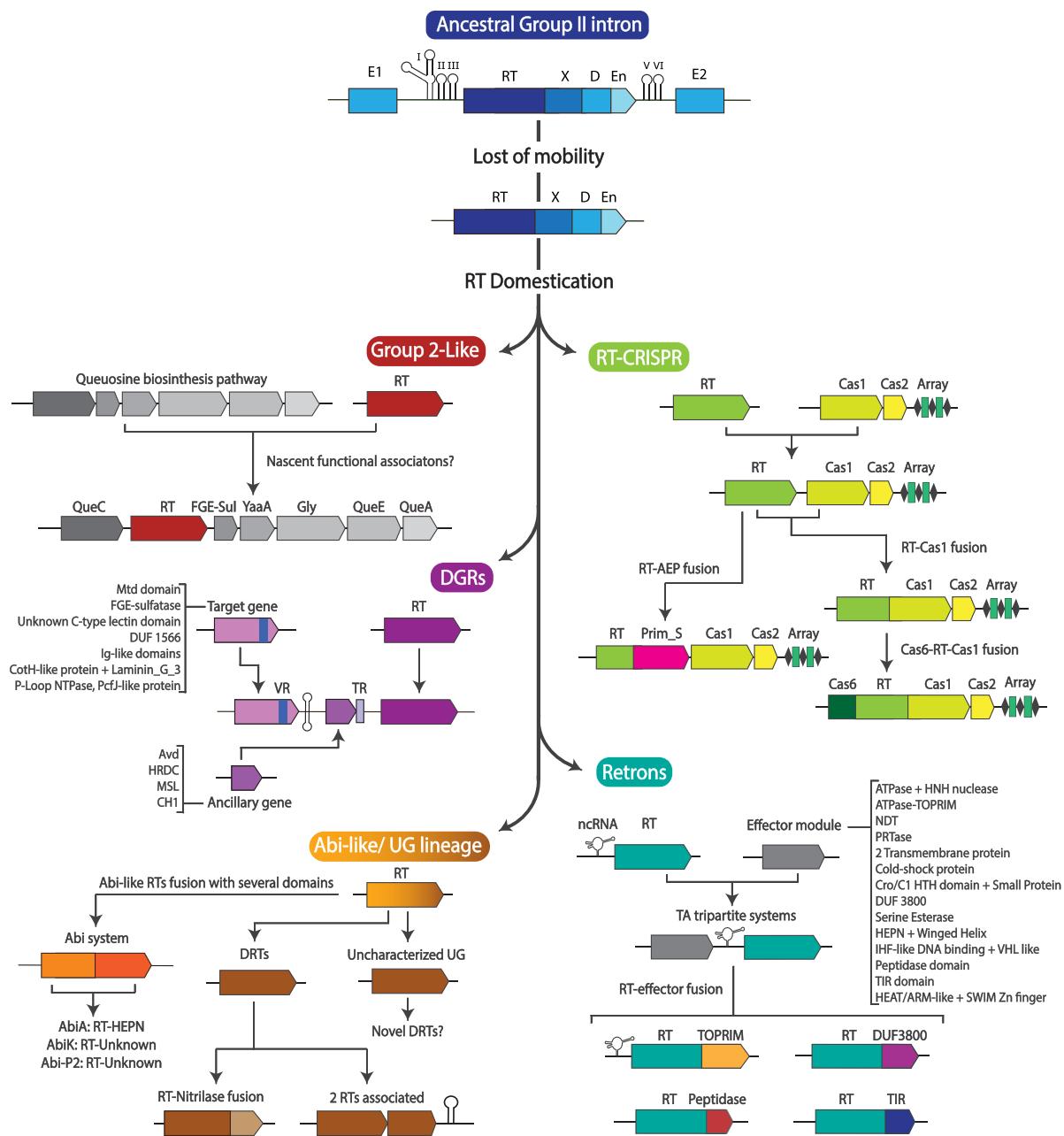


Figure 2. RT domestication from an ancestral retroelement. The scheme depicts a hypothetical scenario of the domestication of the different RT lineages from an autonomous mobile retroelement, probably an ancestral group II intron. The RNA domains (I–VI) as well as the exon regions (E1 and E2) and the different domains of the intron encoded protein (RT: Reverse Transcriptase; X: Maturase; D: DNA-binding and En: Endonuclease) are indicated. At some point, the intron RNA component was lost, and the remaining RTs coevolved with their genomic context, resulting in the recruitment of RTs to various specialized systems. The G2L RTs may represent an intermediate state between a mobile group II intron and a domestication of these RTs leading to nascent functional associations. Here, an example of a putative association between G2L-RTs and a multicomponent system comprised by the queuosine biosynthesis pathway QueA, QueC, QueE (queuosine biosynthesis pathway) and FGE-s (FGE sulfatase), YaaA (DNA-binding protein) and Gly (Glycosyltransferase) proteins is shown. The RTs associated with CRISPR-Cas systems present different stages of association: first a group II intron was inserted into the genomic context of a cas1 gene. Following a loss of mobility, the remaining RT coevolved with cas1, resulting in a functional association. Subsequently, RT and Cas1 were fused and, later, a Cas6 domain was acquired independently. Alternatively, the RT and AEP primase domain (Prim.S) may have fused to form a particular group of RT-CRISPR systems. DGR RTs have evolved to hypermutate the variable region (VR) of several target genes with a specific fold domain and through a mutagenic reverse transcription of a target region (TR). DGRs are typically assisted by various ancillary proteins including Avd (Accessory variability determinant), HRDC (Helicase and RNaseD C-terminal domain), MSL (MutS-like) and CH1 (Conserver Hypothetical Gene 1). In the case of retrons, RTs have become associated with small ncRNAs and an effector module, some of them forming tripartite toxin/antitoxin systems with antiviral properties. The wide variety of effectors suggests that the retron unit is highly modular and, in some cases, the RT and other domains have fused (TOPRIM: topoisomerase-primase; DUF3800; peptidase and TIR: toll-interleukin receptor). The RTs from the Abi-like/UG lineage are highly divergent and phylogenetically distant from those of group II introns, suggesting that this lineage may represent an old domestication event, creating a new mechanism of defense against phages. The Abi RTs are fused to unknown domains, except for AbiA, in which the RT is fused to a HEPN domain. Most RTs from the UG lineage remain uncharacterized, but some have been shown to confer resistance to phages and are now known as defense-associated RTs (DRTs). DRTs may consist of the RT itself, but the RT is also often fused to a nitrilase domain or associated with other RT proteins. The arrows denote events that have been inferred during the domestication of the different RT lineages.

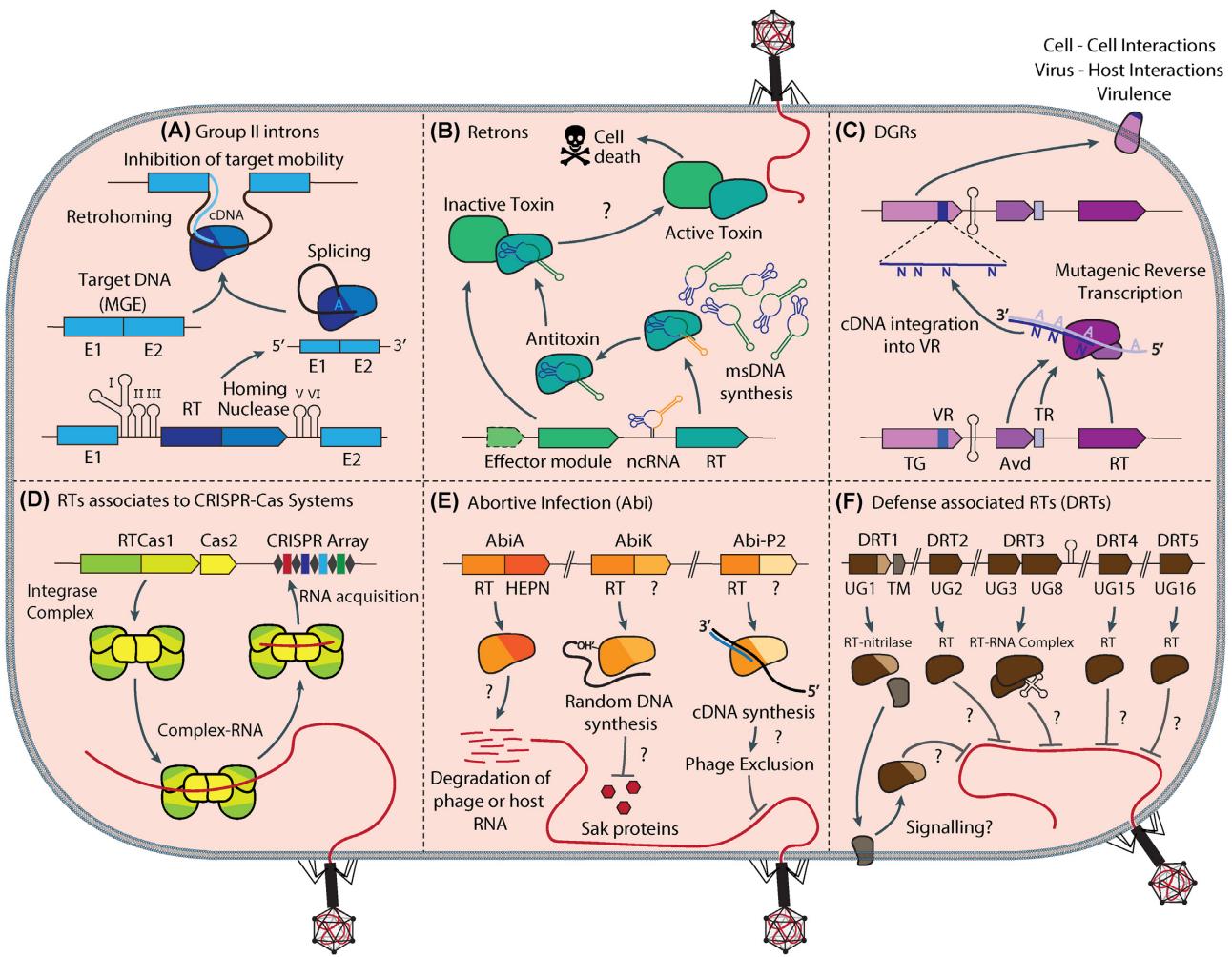


Figure 3. Role of RTs in prokaryotic immunity. (A), Group II intron life cycle. The RNA domains (I–VI) as well as the exon regions (E1 and E2) and the RT and homing nuclease domains of intron encoded protein are indicated. Intron splicing steps form a lariat RNA which constitute a complex with the IEP. As soon as the target DNA is found, intron reverse splices into the target DNA, frequently a mobile genetic element, abolishing its mobility via a process termed retrohoming. (B), Retrons are tripartite modular systems, some of which have been shown to act as toxin/antitoxin systems. The retron unit is comprised by a non-coding RNA (ncRNA) and the RT gene. The RT produces an msDNA at high copy number using the ncRNA as template. This msDNA remaining bound to the RT, forming the antitoxin unit. The effector module, formed by one or more genes with different enzymatic functions, constituting the toxin effector, is inhibited by direct contact with the antitoxin. By an as yet uncharacterized mechanism, when a phage infects the cell, triggers the msDNA processing/degradation, allowing the toxin to cause cell death, to protect cell population. (C), Diversity-generating retroelements (DGRs) introduce random mutations into the template repeat (TR), in a reaction known as mutagenic reverse transcription, which is performed by the RT with the help of an ancillary protein (typically Avd: Accessory variability determinant). The mutated cDNA is then integrated into the variable region (VR) of a target gene (TG), generating considerable sequence variability. The TG generally encodes a protein displaying protein–protein interactions or with surface activities enabling the host cell to adapt to different conditions, such as host–phage interactions. (D), RTs associated with RNA-targeting CRISPR-Cas systems (type III and VI) form an integrase complex together with Cas1 and Cas2, facilitating the acquisition of RNA molecules, which are integrated into the CRISPR array as a new spacer. (E), RTs involved in abortive infection (Abi) systems are part of AbiA, AbiK and Abi-P2, but their mechanisms of action remain unknown. In AbiA, the RT is fused to a HEPN domain and is thought to degrade host or phage RNA to confer resistance. In AbiK, a random-sequence DNA molecule remains covalently attached to an OH-group on a tyrosine residue in the RT domain, which is fused to an unknown domain. AbiK uses an as yet incompletely undeciphered mechanism to block phage Sak proteins, conferring immunity. Abi-P2, which has demonstrated reverse transcriptase activity, is formed by a RT and a domain of unknown function capable of excluding phages by an as yet undetermined mechanism. (F), Defense-associated RTs (DRTs) are novel antiviral systems in which RT activity is required for resistance, through an unknown mechanism. These systems consist of RTs from different unknown groups (UGs) that act alone (DRT2, DRT4 and DRT5), together with small membrane proteins (TM; DRT1), or, in case of DRT3, two RTs from two different UGs and a ncRNA are required for anti-phage properties.

the chromosome (Novikova et al. 2014). Thus, the relaxase facilitates plasmid dispersal and retrotransposition events, whereas the group II intron regulates relaxase expression, maintaining a balance that may be positive for the host.

Contrasting with this inhibitory function, one recent study showed that group II introns could increase genetic diversity by creating chimeric relaxase variants through the shuffling of coding sequences in RNA and DNA, thereby showing that these retroelements can be beneficial to the conjugative

elements that harbor them and to their bacterial host (LaRoche-Johnston, Bosan and Cousineau 2020). Moreover, the existence of host factors acting as global regulators of intron mobility has been demonstrated. Some of these factors, such as RNase E, act as depressors, whereas others, such as the alarmones ppGpp and cAMP, act as stimulators, demonstrating the role of nutritional stress in the activation of these retroelements (Coros et al. 2008, 2009; Nisa-Martínez, Molina-Sánchez and Toro 2016). Thus, group II introns may act by preventing the damage caused by

other MGEs activated by stress conditions. Although these findings support that group II introns occasionally facilitate host adaptation, they are also in agreement with the reported selfish behavior of these retroelements that allows them to spread and survive within their bacterial host.

Retrons

In 1984, an unusual single-stranded DNA, known as multicopy single-stranded DNA (msDNA) was found to accumulate to high levels in the bacterium *Myxococcus xanthus* (Yee et al. 1984). It was subsequently shown that msDNA was produced by an RT using a two-region (*msr* and *msd*) non-coding RNA (ncRNA) as a template, forming a unit called retron (Inouye et al. 1989; Lampson et al. 1989; Lim and Maas 1989). This was the first demonstration of RT activity in prokaryotic organisms (Inouye 2017). Retrons have been studied in detail biochemically (Lampson, Inouye and Inouye 2005), but their biological role has remained unknown more than 30 years after their discovery (Simon, Ellington and Finkelstein 2019). However, several independent studies have shed light on the function of retrons, proposing that they act as a novel prokaryotic defense system against phages (Gao et al. 2020; Millman et al. 2020). Interestingly, these studies show that the msDNA molecule is crucial for the antiviral activity of retron systems (Gao et al. 2020; Millman et al. 2020; Bobonis et al. 2020a).

During msDNA synthesis, the RT protein is bound to the transcribed ncRNA just downstream from the *msd* region, where it initiates a reverse transcription reaction using a 2'-OH group present in a conserved branching G residue in the *msr* region as a primer. The resulting msDNA remains covalently attached to the *msr* RNA as a single branched molecule through a 2'-5' phosphodiester bond (Shimamoto, Inouye and Inouye 1995). Despite the considerable divergence of *msr/msd* sequences in the small number of experimentally validated retrons, all these sequences have a number of structural properties in common, including complementary 5' and 3' ends of the ncRNA, facilitating formation of the secondary structure of the RNA. The *msr* region presents a variable number of short stem-loops and the *msd* region folds into a single hairpin with a long stem, all of these features being indispensable for msDNA production (Lampson, Inouye and Inouye 2005; Simon, Ellington and Finkelstein 2019).

The recent expansion of the range and diversity of known retrons based on genome survey analyses has made it possible to increase the number of putative retrons from tens to thousands, with most of these elements containing in RT motif 7 the characteristic 'VTG' signature (Toro et al. 2019a). About a third of annotated retrons were thought to encode an ancillary gene (Simon, Ellington and Finkelstein 2019), but a computer pipeline for the systematic prediction of genes specifically associated with retrons has shown that approximately 75% of retrons include an additional component, as an independent gene or a RT-fused domain (Mestre et al. 2020); retrons should, therefore, be considered tripartite systems (Fig. 2). Interestingly, this report showed that the use of covariance models and consensus RNA structure detection makes it possible to identify putative retron ncRNA consensus structures, even in groups with no experimentally validated representatives. A comparison of the phylogenies of the three retron components indicates that retrons are not only highly modular, with the same type of RT associated with different domains and vice versa, but that they have also co-evolved, suggesting that they may act as a functional unit. Moreover, based on the high diversity of putative enzymatic activities encoded by the genes or associated domains, retrons have been classified into 13 types and 25 subtypes, revealing

a tremendous diversity of possible mechanisms and biological functions not only related to defense (Mestre et al. 2020).

Interest in prokaryotic defense mechanisms against phages has led to several strategies for searching for novel immunity systems. The most successful has involved searching for clusters of antiviral system genes in defense islands (Doron et al. 2018). This approach has led to some retron types being identified as abundant in these islands, and two independent reports (Gao et al. 2020; Millman et al. 2020) have recently shown that some retron systems confer resistance to a wide range of phages. In addition, mutations of the three components of the system have been shown to abolish immunity, indicating that all three are required for correct activity. The isolation of phages able to overcome the resistance conferred by Retron-Eco6 (Ec48), a type IV retron system according to the recent classification (Mestre et al. 2020), led to the detection of genetic mutations inactivating a phage RecBCD inhibitor. This suggests that Retron-Eco6 acts as a RecBCD guardian, sensing the presence of phage-encoded RecBCD inhibitors and somehow activating the associated protein, in this case, a two-transmembrane domain protein, resulting in cell death (Millman et al. 2020). However, Retron-Eco8, a type I-B2 retron system (Mestre et al. 2020), has been shown to act independently of RecBCD (Millman et al. 2020), highlighting the diverse possible modes of action underlying the antiviral activity of retrons, potentially due to highly diverse enzyme activities.

In parallel, Retron-Sen2 (St85), a type I-B1 retron system (Mestre et al. 2020), has been shown to act as a novel type of tripartite toxin/antitoxin (TA) system, in which the RT and msDNA form the antitoxin directly blocking the toxin unit called RcaT. This protein has been predicted to present a N-terminal nucleoside deoxyribosyltransferase-like (NDT) domain and a C-terminal DNA binding domain (Mestre et al. 2020). Furthermore, the toxicity of RcaT increases in cold temperatures or anaerobic conditions (Elfenbein et al. 2015; Bobonis et al. 2020a) and its binding to the RT protein does not interfere with its toxicity (Bobonis et al. 2020a). However, the presence of the RT-msDNA complex binding the effector protein determines anti-toxin specificity (Fig. 3B). Phage-origin triggers and prophages-encoded blockers of this novel TA system have also been identified, suggesting an extensive arms-race between retron systems and phages (Bobonis et al. 2020b). Some of the triggers identified (Dam and RecE) have been shown to directly interact with the RT-msDNA complex, methylating, cleaving and/or binding the msDNA component and, therefore, triggering the activation of the toxin. Moreover, Dam and RecC have anti-restriction properties and could lead to Abi mediated by RcaT through the inactivation of RT-msDNA antitoxins, suggesting crosstalk between the innate/adaptive immunity systems and this tripartite TA system. One of the identified blockers is RacC, a small protein encoded by Rac-prophage which acts directly against the toxin activity, raising the possibility of the existence of phage-encoded anti-retron systems.

Despite the great progress in our understanding of retrons based on these recent discoveries, many biological questions remain unanswered, concerning the ways in which different types of retron systems sense phages, how the antitoxin is inactivated, or the role of the toxin in the final step in this form of immunity. Moreover, new types of retrons with new ncRNA structures have recently been identified but have not yet been experimentally characterized (Mestre et al. 2020). Further studies are therefore required to improve our understanding of the biological role of retrons as an anti-phage system and of the other

putative functions of these elements derived from their potential for sensing other cell signals.

Diversity-generating retroelements

DGRs are a unique type of domesticated RT-containing system that have evolved to provide benefits to the host through a reverse transcription reaction generating broad sequence variability in a specific target gene (Zimmerly and Wu 2015). It has been suggested that they originated from a loss of movement capacity in another type of retroelement followed by diversification, and they are widespread in phages, plasmids, bacterial and archaeal genomes (Paul et al. 2017; Yan et al. 2019; Roux et al. 2020). The functional unit of DGRs is highly diverse and formed by a variable gene cassette, but all DGRs have at least three essential components: a reverse transcriptase (RT), a template repeat (TR) and a target gene (TG) with a variable region (VR) displaying $\approx 90\%$ sequence identity to the TR (Fig. 3C). DGRs increase the ability of the host to adapt to changing environmental conditions through a reaction called mutagenic retrohoming, during which the TR RNA is randomly modified by a mutagenic reverse transcription process, and the resulting cDNA, typically with random A-to-N mutations, is inserted into the VR region, replacing the native sequence and creating multiple novel versions of the TG (Medhekar and Miller 2007; Guo et al. 2014).

Several DGRs have been characterized in bacterial genomes (Le Coq and Ghosh 2011; Arambula et al. 2013), but the best-known and understood example of mutagenic retrohoming is that of the DGR of the *Bordetella* phage BPP-1 (Liu et al. 2002). DGR activity controls phage tropism switching, by generating new variants of the major tropism determinant (Mtd) protein of the tail. This protein is responsible for binding to pertactin, an adhesin on the cell surface of *Bordetella* species that is expressed only during the virulent Bvg⁺ phase. DGR hypermutation in Mtd therefore facilitates the adaptation of phage tropism to surface modifications in the host bacterium (Liu et al. 2002; Doulatov et al. 2004). The VR is found at the 3' end of the *mtd* gene, corresponding to a Clec (C-type lectin) fold consisting of a structural scaffold and a final region in which massive mutations can occur, resulting in functional protein variants (McMahon et al. 2005). Furthermore, the mutagenic retrohoming performed by the BPP-1 phage DGR requires several ancillary elements for efficacy. An accessory variability determinant (*avd*) gene is involved in the mutagenic reverse transcription reaction, binding both the RT protein and the RNA of the TR (Alayyoubi et al. 2013). Following cDNA synthesis, recognition between TR and VR requires a GC-rich sequence called the initiation of mutagenic homing (IMH) sequence at the 3' end of the VR, together with a slightly different IMH sequence (IMH*) in TR. A DNA stem-loop structure just downstream from the IMH sequence facilitates IMH–IMH* recognition, ensuring directional retrohoming and, therefore, resulting in the correct insertion of a novel VR in the target gene (Guo et al. 2011; Naorem et al. 2017).

Over the years, research has greatly expanded the number of putative DGRs identified, with the prediction of these retroelements in genomic (Park et al. 2012; Schillinger and Zingler 2012; Nimkulrat et al. 2016; Wu et al. 2018) and metagenomic data (Yan et al. 2019; Roux et al. 2020). Furthermore, bioinformatics tools have been developed to identify and characterize DGRs. These tools include DiGReF (Schillinger et al. 2012), DGRscan (Ye 2014) and MyDGR (Sharifi and Ye 2019). All these reports have progressively revealed the widespread presence of DGRs in prokaryotes and phages and have demonstrated the great variability of their genetic components and their functional diversity. Indeed, many bacterial DGRs have been shown

to be encoded by temperate phages inserted into bacterial chromosomes as prophages (Benler et al. 2018). Based on RT sequences, the largest DGR dataset available compiles 32 321 sequences, grouped into 1318 clusters ($\geq 50\%$ identity), including DGRs from phages and prokaryotic organisms, in both genomes and metagenomes (Roux et al. 2020). This survey revealed that DGRs predominate in continually changing environments, in which hypermutation is highly beneficial to the host. In these ecological conditions, continual attempts at the horizontal gene transfer of DGR cassettes are made between phylogenetically distant organisms, enhancing the adaptation of a broad range of biological entities. Furthermore, using non-synonymous single nucleotide variants (SNVs) has been shown that most DGRs (50–75%) analyzed present signs of recent activity, with higher activity levels in phage-associated than in cellular DGRs, in which hypermutation may be induced under stress conditions (Roux et al. 2020).

DGRs present a broad range of cassette architectures, based on the order, number and orientation of their components (Fig. 2). For example, there are four classes of accessory genes, with *avd* the most common (over 70%), but some DGR loci lack this ancillary ORF. Furthermore, DGRs can present multiple target genes (from 2 to 8) and can act in *trans* (Wu et al. 2018). Despite this modular organization, the target genes typically encode multidomain proteins with the VR located at the C-terminus. These regions are associated only with the C-type lectin fold and with an uncharacterized domain next to Ig-like fold protein sequences (Roux et al. 2020). In addition, VRs share a highly conserved sequence on either side of the fold sequence, whereas the internal sequence is subject to mutagenesis (Wu et al. 2018). The crystal structure of the C-type lectin fold has revealed an unusually large capacity to accommodate massive sequence variation (Handa et al. 2016). This and the conserved bias towards adenine mutation indicate that DGRs are mechanistically limited in terms of how and where they can produce diversity (Wu et al. 2018; Roux et al. 2020). However, the target proteins have also been shown to be highly modular, suggesting that genetic recombination occurs between independent folding domains and a C-terminal C-type lectin fold to generate chimeric targets (Roux et al. 2020). This process may be the evolutionary source of the involvement of DGRs in various functions beneficial for the host. Target genes are currently classified on the basis of the putative functions of the domains outside the VR sequence, mostly involved in protein–protein binding, ligand binding or surface displays activities, suggesting a broad range of potential biological functions, including virulence, virus-host or cell-cell interactions (Fig. 3C). Most phage target genes encode structural proteins, the variability of which enables phages to overcome bacterial defenses based on cell wall modifications, whereas cellular target genes encode proteins involved in binding extracellular substrates (Roux et al. 2020). A new target function has recently been described, with a group of specific cyanobacterial DGRs able to hypermutate a small pocket in binding domains of multidomain proteins broadly involved in regulatory pathways (Vallota-Eastman et al. 2020). Given the broad array of possible scenarios, further characterization of novel types of DGRs is required to develop a full understanding of the role of these elements in their biological niches.

RTs associated with CRISPR–Cas systems

CRISPR–Cas (Clustered regularly interspaced short palindromic repeats and CRISPR-associated proteins) systems are an adaptive immune system harbored by archaea and bacteria for

defense against foreign nucleic acids, mostly MGes, such as phages, transposons and plasmids (Barrangou *et al.* 2007). CRISPR-mediated immunity occurs in three stages: adaptation, expression and interference (Van der Oost *et al.* 2014). However, the mechanism of the immune response is highly variable, thanks to the considerable diversity of CRISPRs, comprising two classes, six types and 33 subtypes (Makarova *et al.* 2020). The relationship between RTs and these very diverse defense systems was described for the first time in 2008 (Kojima and Kanehisa 2008; Simon and Zimmerly 2008).

Both these studies showed that the reverse transcriptase-coding region was adjacent or fused to the *cas1* gene, which is present in most CRISPR-Cas systems. This observation was endorsed by broader studies of RTs linked to these systems (Toro and Nisa-Martínez 2014; Toro, Martínez-Abarca and González-Delgado 2017). Simultaneously, it was shown that Cas1 and Cas2 form an integrase complex responsible for performing the adaptive step of CRISPR-Cas immunity, which involves the insertion of a genetic fragment of the invading agent, known as the protospacer, into the CRISPR array as a new spacer between two repeat sequences, the direct repeats (DRs; Yosef, Goren and Qimron 2012; Nuñez *et al.* 2014; Jackson *et al.* 2017). This particular group of RTs may, therefore, be involved in acquiring new spacers from RNA sources.

This hypothesis was recently validated by three independent studies in different systems: the Cas6RTCas1 adaptive operon from *Marinomas mediterranea* MMB-1 (Silas *et al.* 2016), the RTCas1 system from *Fuscanibacter saccharivorans* (Schmidt, Cherepkova and Platt 2018) and the RTCas1 system from *Vibrio vulnificus* YJ016 (González-Delgado *et al.* 2019). All three studies showed that RT-CRISPR systems can acquire new spacers directly from RNA *in vivo*, in an RT-dependent manner, via a mechanism with features common to group II intron retrohoming (Fig. 3D). It has also been demonstrated that the additional Cas6 domain in the *M. mediterranea* system participates in CRISPR-RNA (crRNA) biogenesis and is required for RT activity and its regulation (Mohr *et al.* 2018). This crosstalk between the different components of RT-CRISPR systems has been validated by the cryo-EM structure of a Cas6RTCas1-Cas2 complex from *Thiomicrospira* (Wang *et al.* 2021), in which the linking RT helix regulates both RT and Cas1 activities.

In the conditions analyzed, the spacers acquired by the RT-containing adaptive operon in these systems were found to be biased towards highly transcribed regions (Silas *et al.* 2016; Schmidt, Cherepkova and Platt 2018) or highly abundant rRNAs (González-Delgado *et al.* 2019), reflecting a correlation between spacer acquisition events and RNA abundance. The systems studied displayed a bias towards the antisense strand of coding sequences in the newly acquired spacers of the CRISPR Array when the adaptive unit was present alone. Most RTs are associated with type III CRISPR-Cas systems (Silas *et al.* 2017; Toro, Martínez-Abarca and González-Delgado 2017; Toro *et al.* 2019a), which recognize both DNA and RNA targets for target degradation. This slight preference for the antisense strand during acquisition therefore improves the performance of the interference step of type III systems, thanks to the complementarity between the crRNA and the target transcript (Pyenson and Marraffini 2017). However, the natural sources of the spacers found in CRISPR Arrays associated with RT-containing CRISPR-Cas loci remain to be determined. Due to the limited abundance and distribution of RNA phages, only one natural example of CRISPR immunity to an RNA phage has been described (Wolf *et al.* 2020), and only a few matches to DNA phage-like sequences have been found (Silas *et al.* 2017). Different approaches, such as exploring

the larger number of metagenomes available from databases, as well and the recent expansion of known ssRNA phage genomes from tens to 15 611 near-complete genomes (Callanan *et al.* 2020), might help to shed light on this topic and improve our understanding of the biological role of RT-CRISPR-Cas systems in their environmental niche.

The origin and evolutionary relationships of RTs functionally associated with CRISPR-Cas systems have been a matter of debate since the proposal of two models, one suggesting a ‘single-point origin’, in which RTs are domesticated by the CRISPR-Cas locus via the retrotransposition of a group II intron (Silas *et al.* 2017), and the other postulating ‘multiple origins’, in which RTs related to group II introns have been recruited by CRISPR-Cas adaptation modules several times during evolution (Toro *et al.* 2018). However, new insight into the ‘multiple origins’ model was provided by a recent analysis of the putative RT sequences in bacterial genomes (Toro *et al.* 2019a). This study identified a total of 15 clades of CRISPR-RTs: 13 clades closely related to group II introns already reported in the previous analysis (Toro *et al.* 2018), and two new clades, one evolved from retron/retron-like sequences (clade 14) and the other from RTs related to the RT domain present in Abi-P2 systems (clade 15). Furthermore, novel RT sequences adjacent or fused to an Archaeo-eukaryotic primase (AEP) domain and belonging to clade 12 have been described (Fig. 2), suggesting the possibility of cDNA production without the use of a primer to facilitate the acquisition of spacers in these particular systems. All the RTs described in these studies were found to be functionally linked to type III CRISPR-Cas systems, with the exception of those AbiP2-like RTs linked to a type I-C system, but a recent study reported the recruitment of RTCas1 fusion proteins by type VI-A CRISPR-Cas systems (Toro *et al.* 2019b). Type VI systems are class 2 CRISPR-Cas systems in which a unique multidomain protein is responsible for performing the interference step specifically targeting RNA (East-Seletsky *et al.* 2016; O'Connell 2019). The presence of novel RTs in some type III and VI CRISPR-Cas systems suggests an adaptive advantage of these systems in the RNA world that it would be worth characterizing further in the search for unusual mechanisms of spacer acquisition that would help us to understand CRISPR biology.

RTs involved in Abortive bacteriophage infection (Abi) systems

Abi systems are a prokaryotic defense mechanism against bacteriophages in which the infection cycle of the virus is blocked, by halting host metabolism or driving the cell to its death. In this way, Abi systems prevent phage multiplication, thereby protecting the rest of the population (Bernheim and Sorek 2020). A tremendous variety of Abi genes has been described, with about 20 different Abi systems present in *Lactococcus* spp. alone. Another indicator of the heterogeneity of Abi systems is that three such systems contain an RT domain: AbiA, AbiK and Abi-P2 (Fortier, Bouchard and Moineau 2005; Odegrip, Nilsson and Haggard-Ljungquist 2006; Durmaz and Klaenhammer 2007). However, the role of the putative reverse transcriptase activity in these systems remains unclear. Only in the case of AbiK has been shown that the RT domain *per se* is sufficient to confer phage resistance (Fortier, Bouchard and Moineau 2005). For AbiA and Abi-P2, it remains unclear whether other genetic elements present at the same locus are required for defense activity.

Each of these systems has a different mode of operation, but the N-terminal RT domains of Abi proteins have several features in common, including the conservation of a potentially active site with a Y(R/V)DD sequence and the absence of the RT motif 7 (Simon and Zimmerly 2008; Toro and Nisa-Martínez

2014). A C-terminal domain of variable length and uncharacterized function fused to this domain is also present in all types of RT-based Abi systems (Fig. 2). It has been suggested that the AbiA C-terminal domain is a novel version of the higher eukaryote and prokaryote nucleotide-binding (HEPN) domain (Anantharaman et al. 2013). In many defense systems, including restriction-modification (R-M), toxin-antitoxin (TA), CRISPR-Cas and other Abi systems, HEPN domains play a key role due to their RNase activity, which may directly attack viral RNAs or induce host suicide or dormancy by attacking self-RNAs. The presence of RT and HEPN domains in AbiA suggests that phage multiplication may be inhibited through interaction between phage-encoded proteins and a DNA molecule covalently linked to the RT, with the HEPN domain degrading host RNA, driving the cell to suicide and protecting the surrounding community (Fig. 3E). Indeed, the best-characterized AbiA protein, which is found in a lactococcal plasmid (Hill, Pierce and Klaenhammer 1989), is able to stop phage replication. Furthermore, a mutant in a phage recombinase have been shown to be insensitive to AbiA-mediated immunity (Dinsmore and Klaenhammer 1997), suggesting that this recombinase could trigger AbiA. Loci containing AbiA sequences have been shown to confer immunity to a wide set of lactococcal phages (Hill, Miller and Klaenhammer 1990; Tangney and Fitzgerald 2002a). The lactococcal AbiA also displays versatile heterologous activity, conferring resistance to several phages in a *Streptococcus thermophilus* strain (Tangney and Fitzgerald 2002b).

The best-known RT-based Abi system is AbiK, which has also been found in a plasmid native to a lactococcal strain (Emond et al. 1997). In addition to AbiA, this system provides resistance to a broad range of lactococcal phages (936, c2 and P335), reducing infectivity by six orders of magnitude (Fortier, Bouchard and Moineau 2005). The AbiK protein has a DNA polymerase activity divergent from that of canonical RTs and acts like a terminal transferase, polymerizing DNAs of random sequence (Wang et al. 2011). Moreover, the synthesized product remains covalently attached to the enzyme, possibly via a hydroxide group on a tyrosine located in the C-terminal domain, which could act as the primer for a reaction that would be similar to that observed in hepadnavirus self-priming (Wang and Seeger 1992). Conversely, studies of phage mutants able to escape AbiK have identified different viral proteins sensitive to AbiK-mediated immunity. All these proteins, denoted Sak (sensitivity to AbiK), participate in the phage replication process (Ploquin et al. 2008; Lopes et al. 2010; Scaltriti et al. 2010; Scaltriti et al. 2011). In this way, the AbiK system works like AbiA, preventing phage maturation by a functional interaction with Sak proteins and provoking cell death by an unknown mechanism (Fig. 3E).

Unlike the previously described systems, Abi-P2 is encoded in a highly variable region of several P2 prophages present in different *E. coli* strains, with a higher AT content than the host genome, suggesting that this region may have been acquired by HGT (Odegrip, Nilsson and Haggård-Ljungquist 2006). Furthermore, Abi-P2 displays reverse transcriptase activity and confers resistance to the T5 phage by decreasing the plating efficiency of this phage by more than seven orders of magnitude (Fig. 3E). Moreover, deletion of the region of the gene containing the sequence encoding the putative active site of Abi-P2 (YRDD) abolishes resistance to the phage. A recent genome survey analysis enlarged the number of known RT-based Abi systems by an order of magnitude, with Abi-P2 accounting for 75% of these systems (Toro et al. 2019a). Non-redundant homologs of AbiA, AbiK and Abi-P2 have been found in many bacterial phyla and are particularly diverse in the phylum Firmicutes. By

contrast, no RT-containing Abi systems has ever been found in Archaea. With the increasing number of examples identified, research in this field merits a new impetus, to shed light on the role of these Abi-like RTs in phage resistance.

RTs from unknown groups (UG)

Over the years, novel uncharacterized RTs more distantly related to group II intron RTs and clustering in the so-called ‘unknown groups’ have emerged in analyses of increasing numbers of RT sequences (Kojima and Kanehisa 2008; Simon and Zimmerly 2008; Toro and Nisa-Martínez 2014). There are now 28 different UGs covering 10% of prokaryotic RT diversity (Toro et al. 2019a) showing considerable sequence diversity. For example, UG1 and UG5 RTs have a C-terminal nitrilase domain (Simon and Zimmerly 2008), whereas, in UG6, this domain is located further downstream in an adjacent ORF (Gao et al. 2020). UG3 and UG8 represent a unique case in which RT sequences are always located next to each other, suggesting that they act as a functional unit (Fig. 2; Kojima and Kanehisa 2008).

RTs from UG groups have recently been found in defense islands and have been experimentally validated as new anti-phage systems known as Defense-associated RTs (DRTs), in which immunity is dependent on the RT domain (Gao et al. 2020). In DRT type 1 (UG1) both the C-terminal nitrilase domain and a small membrane protein are required for defense (Fig. 3F). In DRT type 3, corresponding to UG3 and UG8, a structured ncRNA downstream from the UG8 gene is required for immunity. By contrast, in DRT type 2 (UG2), type 4 (UG15) and type 5 (UG16), the RT alone is sufficient for resistance against phages. However, the mechanism of action of these novel systems remains undeciphered, and additional experimental data are required to determine how UG-RTs provide protection against different types of phages. Moreover, not all UG groups have been tested, and only some of those investigated display antiviral properties (Gao et al. 2020). The characterization of additional types of UG-RTs is therefore required, for the discovery of new DRT systems.

Group II-like (G2L) uncharacterized RTs

In addition to the RT lineages described above, novel phylogenetic groups may arise through different types of domestication of ancient retroelements, the function of which remains to be elucidated. The RTs of one of these groups are most closely related phylogenetically to group II introns and were therefore named ‘group II-like’ or ‘G2L’ RTs. The main difference between the RTs of this group and those classified as group II introns is the absence of the characteristic intronic RNA structure in the G2L group (Simon and Zimmerly 2008). A previous analysis of the prokaryotic reverse transcriptase landscape (Toro et al. 2019a) led to an increase in the number of identified G2L-type sequences, with these sequences being found to branch off from a node common to CRISPR-associated RTs. There is a close relationship between G2L and CRISPR-RTs, and it should be pointed out that two groups previously classified as G2L (G2L1 and G2L2) have since been reclassified as new clades of RT-CRISPR systems (clade4/6 and clade 9), suggesting that both RT lineages may have evolved from a common ancestral group II intron RT. Moreover, the previously named G2L3 (Simon and Zimmerly 2008) was noticed to be phylogenetically separated from the G2L lineage (Toro and Nisa-Martínez 2014) and more recently, this monophyletic group has been reclassified as a novel UG group named as UG17 (Toro et al. 2019a). The G2L lineage has been greatly expanded and comprises RT sequences clustered within the G2L4 and G2L5 groups (Simon and Zimmerly 2008) and four additional clades (Toro et al. 2019a).

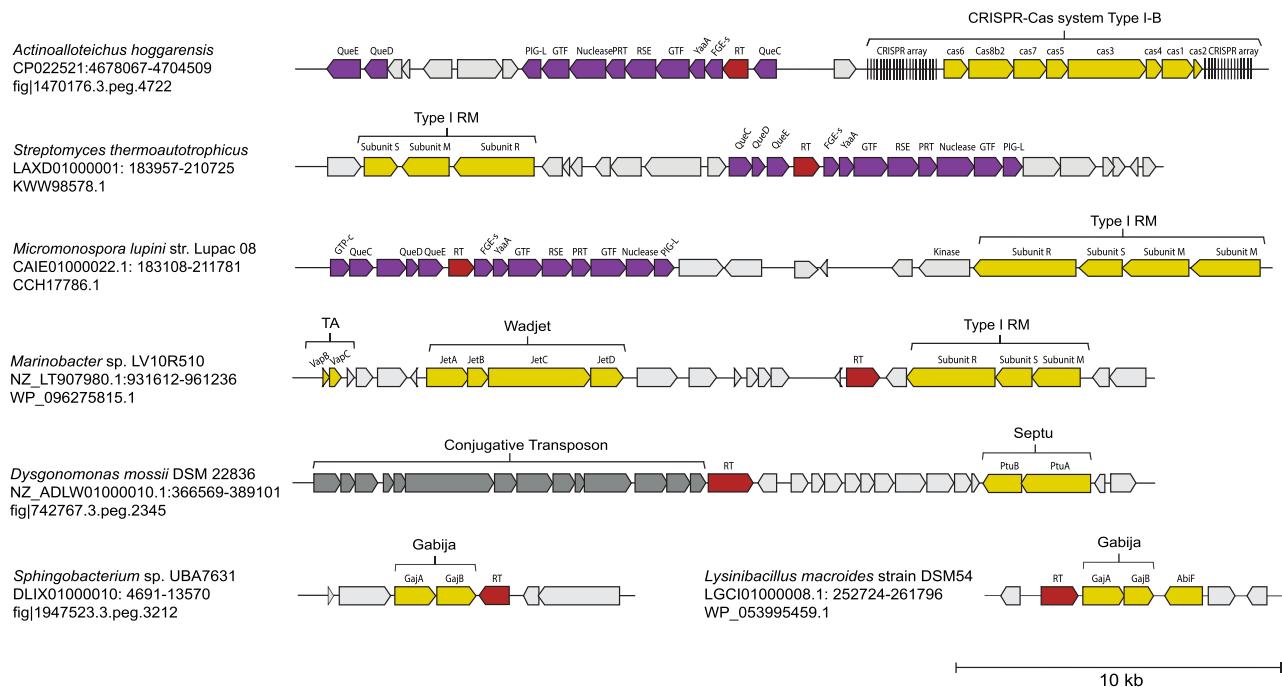


Figure 4. Discovery of putative functional associations of G2L-RTs. Representative instances of RTs from G2L group found in defense islands. Some of these RTs are embedded in a multicomponent system (colored in purple), comprised by QueC, QueD, QueE (queuosine biosynthesis pathway), FGE-s (FGE sulfatase), YaaA (DNA-binding protein), GTF (Glycosyltransferase), RSE (Radical Sam Enzyme), PRT (phosphoribosyltransferase), Nuclease (Flap Endonuclease), PIG-L (GlcNAc-PI de-N-acetylase) and GTP-c (GTP cyclohydrolase). Genes known to be involved in defense are shown in yellow: CRISPR-Cas system type I-B, Type I RM (Restriction-Modification), TA (Toxin-Antitoxin), Wadjet, Septu and Gabija are recently described defense systems (Doron et al. 2018). The genes encoded by a conjugative transposon are shown in dark grey. The encoding strain, the accession number and the genomic coordinates and the protein ID on NCBI or PATRIC databases are indicated to the left.

In this way, G2L may constitute an evolutionary record of an intermediate state between the autonomous mobilization of group II intron RTs and the domestication of these RTs for the performance of useful cellular functions. This hypothesis was tested by analyzing the genomic neighborhood (Mestre et al. 2020) of G2L sequences, to search for possible evidence of emerging functional associations (see Supplementary Methods). Interestingly, a particular subgroup of G2L RTs (cluster 4; Toro et al. 2019a) appears to be embedded in a multicomponent system composed of the products of the queACDE genes involved in the queuosine biosynthesis pathway (Thiaville et al. 2016), two different glycosyltransferases, the 5'-3' exonuclease domain of a family A DNA polymerase, a GlcNAc-PI de-N-acetylase (LmbE), a phosphoribosyltransferase, YaaA (DUF328), which is a DNA-binding protein with a novel fold (Prahlad et al. 2020), and a FGE sulfatase with a C-lec fold (Fig. 4 and Figure S1, Supporting Information). QueC was recently reported to be a component of novel defense systems (Gao et al. 2020), which suggest that this genomic cassette may also be involved in immunity functions. Moreover, 25% of known G2L-RT sequences were located close to known defense system sequences (Fig. 4). Thus, functional associations between G2L and neighboring genes could generate specialized systems with antiviral activities worthy of further investigation.

Differential distribution of RTs in bacteria and archaea

The most exhaustive phylogenetic analysis of bacterial RTs used 198 760 predicted RT proteins and addressed the distribution of RTs, mostly in bacteria (Toro et al. 2019a). The classification of the final dataset of 9141 non-redundant RTs supported the classification of most RTs into the three main groups: group-II introns,

the largest group, accounting for 47% of RTs, retrons/retron-like sequences (25%) and DGRs (12%; Fig. 5A and B). The remaining 16% of RTs clustered into distinct groups including RTs previously reported to be linked to CRISPR-Cas systems, group 2-like (G2L), Abi-like or UG (unknown) groups. These data are consistent with previously reported findings (Kojima and Kanehisa 2008; Simon and Zimmerly 2008; Toro and Nisa-Martínez 2014).

Far fewer RT sequences have been identified in archaea than in bacteria, because of the smaller number of complete genome sequences available in databases (i.e. more than 3×10^5 bacterial genome sequences versus just under 5×10^3 archaeal genome sequences in the PATRIC database; Supplementary Methods). A more detailed analysis of archaeal RTs revealed that DGRs accounted for the largest proportion (30% of total archaeal RTs), contrasting with the situation in bacteria (with only 12%). The opposite pattern was observed for retrons, which account for only 4% of total archaeal RTs (Fig. 5B and C; Supplementary File 1). No RTs from the G2L or Abi-like groups have been found in archaea. The remaining RT groups account for similar proportions of the RTs in archaea and bacteria, with group II introns the most abundant (52%), which are probably proliferating in archaea even though they were originally acquired from bacteria (Fig. 5B; Toro 2003).

In bacteria, the Flavobacteria-Chlorobi-Bacteroidetes group (FCB), candidate phyla radiation (CPR) and Cyanobacteria have high rates of non-redundant RTs relative to the number of sequenced genomes for these phyla (Figure S2A, Supporting Information). Group II introns are, predictably, prevalent in most bacterial phyla, but retrons predominate in Proteobacteria, whereas DGRs account for about 80% of total RTs in CPR, the strongest bias towards a specific RT group in bacteria (Fig. 5D).

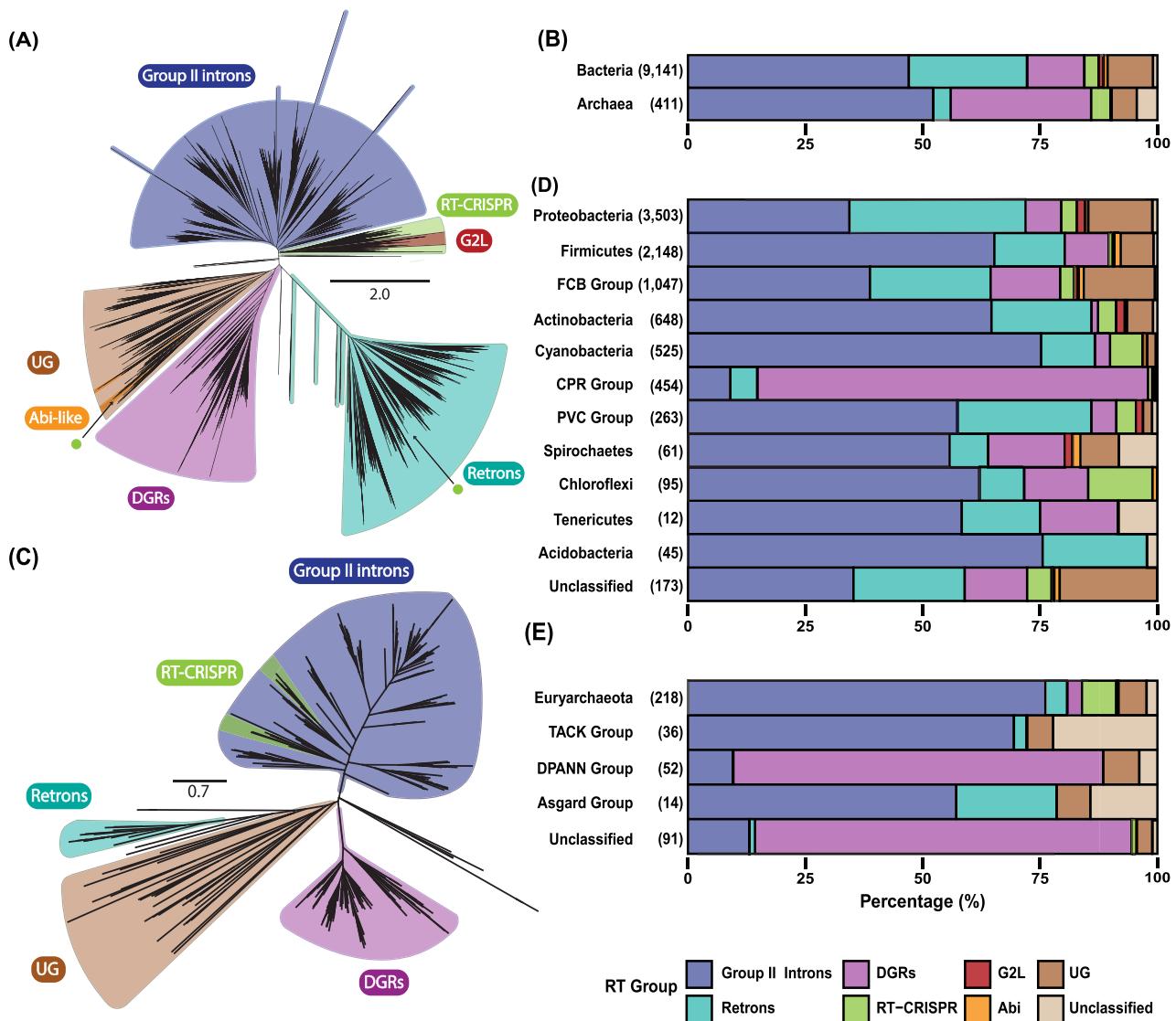


Figure 5. Distribution of prokaryotic RTs. (A) Phylogeny of bacterial RTs. The unrooted trees were constructed from an alignment of 9141 sequences. The tree newick files are provided in Toro et al. (2019) and the redrawn tree is reprinted by permission of the publisher (Taylor & Francis Ltd, <http://www.tandfonline.com>). The branches of each RT type are indicated and highlighted with distinct colors. RTs associated with CRISPR–Cas systems with a retron (clade 14) or Abi-P2 (clade 15) origin are indicated by green points and black arrows. (B) Relative abundance of the different RT groups in bacterial and archaeal genomes (C). Phylogeny of archaeal RTs. The unrooted trees were constructed from an alignment of 411 sequences. The tree newick files is provided in Supplementary File 1. (D) RT distribution in the main bacterial phyla. (E) RT distribution in the main archaeal phyla. Charts showing the proportions of RTs corresponding to group II introns, retrons, DGRs, RTs associated with CRISPR–Cas systems (RT-CRISPR), group II-like RTs (G2L), RT-based Abi systems (Abi), unknown groups (UG) and unclassified RTs. In panels (B, D and E), the number of unique RT sequences is indicated after the name of each domain/phylum.

Looked at the other way (distribution of bacterial phyla for particular RT groups), Proteobacteria is the predominant phylum in most RT lineages, whereas, Firmicutes account for a large proportion of group II introns and Abi-like RTs. Similarly, the CPR group contains about 40% of all singular DGRs (Figure S2B, Supporting Information). These data support the notion that some bacterial phyla have recruited specific groups of RTs to improve their responses to the ecological conditions in their natural environments.

As in bacteria, in most phyla and other taxonomic groups of archaea there is a correlation between the total number of genomes sequenced and the total number of RTs found in the group concerned, except for the TACK group (the Thaumarchaeota, Aigarchaeota, Crenarchaeota and Korarchaeota

phyla), which accounts for 25% of total archaeal genomes but only 7% of non-redundant RT sequences (Figure S2C, Supporting Information). Group II introns predominate in Euryarchaeota, the TACK group and the Asgard group (55–75% of RTs). However, a different scenario emerges for the DPANN group (Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanoarchaeota) and uncultured/unclassified archaea, in which DGRs account for about 80% of RTs (Fig. 5E). Regarding the dominance of phyla for every RT group, it is worth to highlight that 90% of all DGRs were found in DPANN group and uncultured archaea (Figure S2D, Supporting Information).

The abundance of DGRs in the DPANN group and in the bacterial CPR group has been described (Paul et al. 2017; Roux et al. 2020), but we show here that DGRs are, by far, the most abundant

type of RT in these phylogenetic groups. Curiously, the DPANN group is the most divergent of all archaeal groups, resembling the situation of the CPR group in bacteria (Castelle et al. 2018). Indeed, these groups have several characteristics in common and differ from other prokaryotic organisms in having small cell and genome sizes, in their episymbiotic relationships with other organisms and their limited metabolic capacities. In this context, the high abundance of DGRs in the DPANN and CPR groups may reflect a role of RTs of this type in the adaptation of these organisms to their biological niches through changes in membrane proteins facilitating interactions with the host (Castelle et al. 2018).

Biotechnological applications of prokaryotic RTs

RTs, particularly those from eukaryotes and viruses, have been exploited as biotechnological tools in a wide range of fields (Martín-Alonso, Frutos-Beltrán and Menéndez-Arias 2020). The unique properties of several prokaryotic RTs groups have enabled their use in a wide variety of genome editing applications, including gene targeting, high-throughput site-directed mutagenesis technologies, libraries of highly mutagenized genes or with high-throughput transcriptional characterization purposes (Table 1).

For gene targeting the specific interactions between intron and exon-binding sites (IBS-EBS) required for group II intron retroromobility was harnessed to target introns to preprogrammed positions (Guo et al. 2000; Mohr et al. 2000; Karberg et al. 2001; Enyeart et al. 2014). Thus, group II introns have been exploited as the first RNA-guided gene targeting tools in a wide variety of bacterial genomes in the so-called targetron knockout technology (Table 1). With an algorithm designed to identify optimal matches, targetrons are programmed to insert into desired genes by modifying the sequence elements in the intron RNA to base-pair to the IBS in the DNA target site (Zhuang et al. 2009; García-Rodríguez et al. 2014; Gwee et al. 2019; Wen et al. 2020). The coupling of this technique with CRISPR-Cas9 counter-selection has recently increased the chances of finding clones in which the intron is integrated into the target sequence in a recombination-independent fashion (Velázquez, Lorenzo and Al-Ramahi 2019). Group II introns have also been programmed to deliver lox sites which are used by Cre recombinase to generate insertions, inversions, deletions and one-step cut-and-paste operations (Table 1; Enyeart et al. 2013). On the other hand, the ability of retrons to produce msDNA at high copy number has allowed the use of these elements with gene silencing purposes (Table 1; Mao et al. 1995; Simon, Ellington and Finkelstein 2019). This approach is based on the fact that the *msd* region can be modified with a random sequence without disturbing the production of msDNA. Thus, retrons could be engineered to synthesize specific DNA molecules that works as antisense cDNAs to knockdown the mRNA levels of a target gene (Mao et al. 1995).

The particularities of retrons have also been used as an interesting alternative in high-throughput directed mutagenesis applications (Simon, Ellington and Finkelstein 2019). The first approach, referred to as Synthetic Cellular Recorders Integrating Biological Events (SCRIBE), was based on engineered variants of Retron-Eco1 with a modified *msd* region (Table 1; Farzadfar and Lu 2014). The production of the modified msDNA, which is inducible by arbitrary transcriptional signals (light and chemical inducers), acts as a genomic 'tape recorder' of the presence of these signals at population level. To increase efficiency, the retron unit is co-expressed in combination with the

Beta recombinase of bacteriophage λ , leading to the recombination of the modified msDNA into a specific DNA locus in a magnitude- and duration-dependent way of the signals (Farzadfar and Lu 2014). However, due to the moderate recombination level ($\sim 10^{-4}$ recombination events per generation) observed in this system, several works have studied the factors that limits retron recombineering (Simon, Morrow and Ellington 2018; Farzadfar et al. 2020; Schubert et al. 2020). The first improvements in efficiency were carried out by optimizing the promoter and by knocking out genes that reduces the accumulation of msDNA, specifically the mismatch repair protein MutS and the single-stranded exonuclease ExoX (Simon, Morrow and Ellington 2018). This system yields recombination efficiencies 2 order of magnitude higher and allows continuous evolution of target loci by expressing the retron unit under the control of an error-prone RNA polymerase, which generates random mutations in the *msd* region that are introduced into the target DNA (Simon, Morrow and Ellington 2018). Furthermore, the deletion or silencing of other host factors, mainly exonucleases such as *recJ* and *sbcB* (*xonA*), significantly increased the recombination frequency achieving $>99\%$ editing efficiency (Table 1; Farzadfar et al. 2020; Schubert et al. 2021). Both works circumvent some limitations of previous genome-editing technologies in bacteria, including the need to electroporate with mutagenic oligonucleotides or the requirements of specific cis-encoded elements (i.e. PAM motif in CRISPR-Cas9 counterselection). Thus, Farzadfar and colleagues (2020) introduce HiSCRIBE (High-efficiency SCRIBE) and demonstrate their use in recording transient spatial information and continuous genome evolution. On the other hand, Schubert and collaborators (2021) create RLR (Retron Recombination Library) useful to prepare pooled, barcoded mutant libraries for high-throughput screens of allelic variants and uniquely suit to utilize non-designed sources of variation. Interestingly, a recent work has also shown that modifications in the ncRNA structure and in the architecture of the retron operon lead to a more efficient reverse transcription reaction and, thus, in a higher retron recombineering efficiency in bacteria, yeast and human cells (López et al. 2021). Moreover, retron recombineering technologies have also been used in combination with CRISPR-Cas systems for multiplex gene editing in bacteria (Lim et al. 2020) and eukaryotes (Sharon et al. 2018). This last study uses a chimeric RNA comprised by a modified retron ncRNA joined to a CRISPR guide in a technology called Cas9 Retron precISE Parallel Editing via homologY (CRISPEY) ideal to identify causal variants for polygenic traits (Table 1).

Continuous target evolution could also be performed using DGRs, thanks to their ability to generate multiple variants of a protein domain, which can be useful in biotechnological applications such as in phage therapy. However, only one preliminary study has been able to demonstrate the ability of the phage BPP-1 DGR to create variants of tropism proteins likely to bind the T4 lysozyme (Table 1; Yuan et al. 2013). The exploration of novel DGRs with advantageous properties, such as thermostability (Handa, Shaw and Ghosh 2019), could provide promising new tools. In addition, characterization of the mutagenic retrohoming mechanism has revealed that the RT and Avd proteins of the BPP-1 DGR work as a complex and that both are required for cDNA synthesis from both DGR and non-DGR templates with an oligodeoxynucleotide (ODN) primer (Handa et al. 2018). cDNAs synthesized from non-DGR templates also present adenine mutations, revealing this to be an intrinsic feature of the RT-Avd complex. This capacity can be used to create libraries of hypermutated cDNAs, to address the issue of sequence variability when searching for protein variants with significantly higher

Table 1. Biotechnological applications of different prokaryotic RTs.

Applications	Technology	RT type	General description	References
Specific gene knockout-knockdown-deletion	Targetron	Group II introns	The specificity of the IBS-EBS interactions was harnessed to target introns to preprogrammed positions in a wide variety of bacterial genomes.	Zhuang et al. (2009); García-Rodríguez et al. (2014); Gwee et al. (2019); Wen et al. (2020)
	Targetron plus CRISPR-Cas9 counter selection	Group II introns	CRISPR/Cas9 counterselection increased the chances of finding clones that integrated the intron into the target lacZ sequence in a recombination-independent fashion.	Velázquez, Lorenzo and Al-Ramahi (2019)
	GETR	Group II introns	Group II introns deliver new lox sites allowing the recombinase Cre to produce insertions, inversions and deletions and one-step cut-and-paste operations.	Enyeart et al. (2013)
	Antisense cDNA gene regulation	Retrons	Retron engineered to produce a msDNA which contains an antisense cDNA to knockdown a target gene.	Mao et al. (1995)
High-throughput site directed mutagenesis	SCRIBE	Retrons	Retron engineered to produce a msDNA with a desire sequence to modify a target region after recombination which acts as a genomic 'tape recorder' of the presence of regulatory signals at population level.	Farzadfar and Lu (2014)
	Continuous gene evolution	Retrons	Expression of retron under an error-prone RNA polymerase that generates random mutations in the msd region which later is introduced in the desired loci.	Simon, Morrow and Ellington (2018)
	HiSCRIBE	Retrons	Efficient retrov recombineering to recording transient spatial information and continuous genome evolution.	Farzadfar et al. (2020)
	RLR	Retrons	Optimization of retrov-based genome editing to prepare pooled, barcoded mutant libraries for high-throughput screens of allelic variants and uniquely suit to utilize non-designed sources of variation.	Schubert et al. (2021)
	Multiplex gene editing	Retrons	Combination of retrov and CRISPR-Cas9 to enable multiplex gene editing in bacteria.	Lim et al. (2020)
Phage therapy	CRISPEY	Retrons	Retron homology-direct reparation of CRISPR-Cas9 double-strand breaks in yeast.	Sharon et al. (2018)
	Variants of tropism proteins	DGRs	Continuous target evolution through mutagenic retrohoming reaction carried out by <i>Bordetella-phage1</i> DGR to create variants of tropism proteins that bind T4 lysozyme.	Yuan et al. (2013)
High-throughput transcriptional technologies	TIGRT	Group II introns	Thermostable properties of RTs used in different high-throughput RNA characterization purposes.	Mohr et al. (2013)
	Marathon RT	Group II introns	Ultraprocessive and accurate properties of <i>E. rectale</i> Group II intron RT used in different high-throughput RNA characterization purposes.	Zhao, Liu and Pyle (2018)
	Record-Seq	CRISPR-RTs	Using RTCas1-Cas2 integrase complex to storage transcriptional information into CRISPR Array as DNA, describing specific and complex cellular behaviours assessing the cumulative gene expression.	Schmidt, Cherepkova and Platt (2018); Tanna et al. (2020)

levels of activity than the native protein. Theoretically, mutagenic retrohoming is the biological process capable of creating the highest levels of sequence variability, potentially about 10^{30} protein variants, several orders of magnitude more than is possible with eukaryotic systems (Wu et al. 2018). These observations show that DGRs are a powerful potential tool for protein engineering, for which detailed studies are required to determine their usefulness.

Prokaryotic RTs have also been exploited for different high-throughput RNA characterization purposes (Martín-Alonso, Frutos-Beltrán and Menéndez-Arias 2020). The features of specific Group II intron-RTs, such as some thermostable group II intron RTs (TGIRTs; Mohr et al. 2013) and other from *Eubacterium rectale* (Marathon RT; Zhao and Pyle 2018), which present ultra-processor and accurate properties, have been used in RNA-seq and epitranscriptomics technologies. Interestingly, the RT-containing CRISPR loci have also been used with similar applications to record transcriptional events (Schmidt, Cherepkova and Platt 2018), also expanding the CRISPR-Cas toolbox (Jinek et al. 2012; Komor, Badran and Liu 2017; Pickar-Oliver and Gersbach 2019; Vigouroux and Bikard 2020). The adaptive operon of *F. saccharivorans*, a complex formed by the RTCas1 fusion protein and Cas2, has been used to develop Record-seq technology, which records transcriptional events in a CRISPR array as DNA, describing specific and complex cellular behaviors and assessing cumulative gene expression (Table 1; Schmidt, Cherepkova and Platt 2018; Tanna et al. 2020). However, the use of the *F. saccharivorans* system as an RNA-recording tool presents the disadvantage of apparent skewing towards AT-rich regions at the ends of transcripts. It will be important to analyze more CRISPR-Cas system harboring RTs, such as the *V. vulnificus* RTCas1-Cas2A-Cas2B system, which can acquire spacers regardless of their 'GC' content and from any point in the coding sequence (González-Delgado et al. 2019), to overcome this limitation. Record-Seq-derived methods could also be used to improve current technologies for archiving real data for populations of living cells (Shipman et al. 2017), thanks to the ability of systems of this type to store highly transcribed regions potentially containing the information of interest. Moreover, these systems could be used as highly scalable bacterial biosensors to report on gut function (Tanna, Ramachanderan and Platt 2021). In addition, the characterization of singular RTs linked to CRISPR-Cas systems could provide useful tools, such as groups of RTs fused to a putative primase (AEP) domain that might not require priming for cDNA reactions.

On the other hand, the recent finding that some retrons systems can constitute novel tripartite TA systems (Gao et al. 2020; Mestre et al. 2020; Millman et al. 2020; Bobonis et al. 2020a, b) containing an effector associated protein endowed with anti-phage activity could also lead to new biotechnological applications. For example, retrons could be engineered for their use in phage therapy or as riboswitches, applying certain conditions that trigger the desired response. Thus, the potential of these elements as biotechnological tools is only just beginning to be understood. Furthermore, the large RT proteins involved in Abi systems could also provide novel biotechnological tools, thanks to the enzymatic diversity provided by the domains fused to the C-terminus of the RT domain. These natural fusions are worth characterizing, particularly for AbiA, in which polymerase and putative RNase activities are combined in a single protein that may function differently from the RTs of retroviruses. Similarly, the properties of other RTs from other groups, such as the UG and G2L groups, some of which are involved in antiviral defenses (Gao et al. 2020), and those found in Archaea could supply novel

functions for use in the existing or new fields. We are currently witnessing the first uses of prokaryotic RTs in a broad range of biotechnology applications.

CONCLUDING REMARKS

This review describes how prokaryotic organisms have domesticated the different types of RTs to perform defense functions against phages and other MGEs. The outstanding research performed in the last few years has not only resolved the 30-year-old enigma of the biological role of retrons (Gao et al. 2020; Millman et al. 2020; Bobonis et al. 2020a, b), but also revealed why the presence of an RT is advantageous for some types of CRISPR-Cas systems (Silas et al. 2016; Toro et al. 2019b), and has shown that some uncharacterized groups of RTs confer resistance to a wide range of phages (Gao et al. 2020). However, several key questions remain unanswered. From an evolutionary point of view, it would be worthwhile improving our understanding of how group II introns evolved in their host genome and gave rise to other types of prokaryotic RTs. Interestingly, defense systems and MGEs have several characteristics in common, including being the only two functional categories of genes with a negative selection coefficient (Iranzo et al. 2017), implying that they can be deleterious to the host. These parallel evolutionary histories could provide the necessary substrate for mobile genetic elements and defense systems to become associated on multiple occasions throughout evolution, as demonstrated for CRISPR-Cas systems (Koonin and Makarova 2019). It would be reasonable to hypothesize that group II introns may have been the source of other domesticated RTs for various defense systems. Research on RTs phylogenetically related to group II introns, such as the G2L lineage, could shed light on this issue.

With our current knowledge of RTs, we are only just beginning to understand the impact of these enzymes in prokaryotes. Exploring new groups of recently expanded group II introns (Toro et al. 2019a) may shed light on how these selfish elements achieve a balance between their inherent need to spread and their ability to control MGEs, resulting in benefit to the host. The recently discovered anti-phage activity of retrons has opened up a whole new avenue of research, extending from clarification of the stages of retron-mediated immunity to the way in which phages overcome this defensive barrier. Given the modularity and diversity of retron systems (Mestre et al. 2020), a number of completely different biological mechanisms may be involved. A parallel track could be followed for UGs, in which computational analysis may also reveal the existence of more diversity, potentially associated with other proteins or ncRNAs and constituting novel DRTs. Interestingly, phylogenetic analysis (Toro and Nisa-Martínez 2014; Toro et al. 2019a) has shown that the Abi-RTs lie among UG-RTs, suggesting that DRTs may confer resistance similarly to Abi-RTs, leading the host cell to commit suicide. The presence of an RT domain provides an adaptive advantage to RNA-targeting CRISPR-Cas systems (types III and VI; Silas et al. 2016; González-Delgado et al. 2019), but the origin of the spacers in the CRISPR arrays of systems carrying RTs and the structural basis of the mechanism of RNA spacer integration into the array remain a mystery. Further studies are therefore required to improve our understanding of type III and VI CRISPR biology. Finally, in addition to their known functions, the other possible roles of DGRs should be studied, to shed light on the ways in which mutagenic reverse transcription helps the host to adapt to changing environmental conditions, particularly for CPR bacteria and DPANN archaea, in which DGRs are particularly abundant (Paul et al. 2017; Roux et al. 2020).

The recently proposed ‘pan-immune system’ model suggests that, even if a particular strain does not have genes encoding all types of defense systems, the pan-genome of a mixed population of strains potentially encodes a battery of defense systems that protects the whole population (Bernheim and Sorek 2019). According to the implications of this model, it would be worthwhile analyzing the possible co-occurrence of RT-containing defense systems with other immunity systems, to check for potential crosstalk between antiviral mechanisms. Such studies would help to unravel the complexity of prokaryotic immunity.

The answers to all these questions would have implications for the biotechnological applications of RT-containing systems in new fields, such as phage therapy. It might become possible to design phages or molecules triggering cell death mediated by a retror toxin or by DRT/Abi immunity. Furthermore, natural or engineered phages could be used to overcome RT-mediated immunity to kill a particular bacterial strain with precision. Moreover, the biochemical characterization of novel RTs may lead to the discovery of new functions useful in high-throughput RNA technologies, genome editing or continuous gene evolution, among others. This survey should facilitate the experimental characterization of novel RT groups and promote additional lines of research leading to a better understanding of the roles of these prokaryotic RTs and future applications.

BOX 1.

Role of group II introns in the origin of eukaryotic life

RTs may have played a key role in the transition from the simplest RNA molecules to the current DNA world (Iyer, Koonin and Aravind 2003; Mustafin and Khusnutdinova 2019). RNA-dependent RNA polymerases (RdRP) are thought to be the evolutionary ancestors of the RT protein family (Ellefson et al. 2016; Koonin et al. 2020). There has been considerable speculation about the evolution of RTs since the most recent common ancestor (Eickbush 1997; Curcio and Belfort 2007). In this scenario, group II introns are believed to be one of the drivers of eukaryotic evolution, as they are considered to be the ancestors of nuclear spliceosomal introns, non-LTR retrotransposons and telomerase after colonizing eukaryotes through bacterial endosymbionts that evolved into mitochondria and chloroplasts (Lambowitz and Belfort 2015; Novikova and Belfort 2017; Haack and Toor 2020). It has also been suggested that group II intron proliferation in primitive eukaryotic cells may have promoted the formation of a nuclear envelope to separate splicing from translation (Martin and Koonin 2006).

The spliceosome is an RNP complex, formed by five small nuclear RNAs (snRNAs) and about 80 proteins, which eliminates introns from pre-mRNA molecules (Wahl, Will and Lührmann 2009). Parallels between group II introns and spliceosomal introns suggest that the later ones evolved from group II introns by fragmentation and reassembly (Chalamcharla, Curcio and Belfort 2010; Qu et al. 2014). Functionally, both elements require two transesterification reactions for splicing, with a bulged adenosine involved in the first step (Lambowitz and Belfort 2015). Structurally, the organization of the active site of the group II intron RNP is similar to that of the yeast spliceosome (Agrawal, Wang and Belfort 2016; Qu et al. 2016), and the architecture of the maturase domain of various group II introns is similar to that of Prp8 in the spliceosome complex (Zhao and Pyle 2016). All these similarities suggest a strong relationship

between group II introns and the spliceosome. On the other hand, the RTs of group II introns, non-LTR retrotransposons and telomerase act through analogous mechanisms, suggesting that ancestral group II introns also gave rise to these other elements (Eickbush and Malik 2002). Once non-LTR-retrotransposons had evolved from group II introns, they may have given rise to other more complex retroelements, such as LTR-retrotransposons, through the incorporation of an integrase. Finally, retroviruses are believed to have evolved from LTR-elements by incorporating envelope genes from other viruses (Eickbush and Jamburuthugoda 2008; Finnegan 2012).

AUTHOR CONTRIBUTIONS

A.G.-D. and M.R. M. researched data for this article. All authors contributed to the discussion of the content and reviewed and edited the manuscript. A.G.-D. wrote the paper with contributions from all the authors.

ACKNOWLEDGMENTS

We thank F.L. Nobrega, J. Irazoqui, L. Menéndez-Arias and M. Redrejo-Rodríguez for their helpful comments and critical reading of the manuscript. This work was supported by the Spanish Ministerio de Ciencia, Innovación y Universidades; ERDF (European Regional Development Funds) research grant (BIO2017-82244-P). A.G.-D. was supported by a FPU predoctoral fellowship grant from the Ministerio de Economía y Competitividad (FPU15/02714). We acknowledge support of the publication fee by the CSIC Open Access Publication Support Initiative through its Unit of Information Resources for Research (URCI).

SUPPLEMENTARY DATA

Supplementary data are available at [FEMSRE](#) online.

Conflicts of interest. None declared.

REFERENCES

- Agrawal RK, Wang HW, Belfort M. Forks in the tracks: group II introns, spliceosomes, telomeres and beyond. *RNA Biol* 2016;13:1218–22.
- Alayyoubi M, Guo H, Dey S et al. Structure of the essential diversity-generating retroelement protein bAvd and its functionally important interaction with reverse transcriptase. *Structure* 2013;21:266–76.
- Anantharaman V, Makarova KS, Burroughs AM et al. Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing. *Biol Direct* 2013; 8:15.
- Arambula D, Wong W, Medhekar BA et al. Surface display of a massively variable lipoprotein by a *Legionella* diversity-generating retroelement. *Proc Natl Acad Sci* 2013;110: 8212–7.
- Baltimore D. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* 1970;226:1209–11.
- Barrangou R, Fremaux C, Deveau H et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 2007;315:1709–12.

- Belfort M, Lambowitz AM. Group II intron RNPs and reverse transcriptases: from retroelements to research tools. *Cold Spring Harb Perspect Biol* 2019;11:a032375.
- Benler S, Cobián-Güemes AG, McNair K et al. A diversity-generating retroelement encoded by a globally ubiquitous *Bacteroides* phage. *Microbiome* 2018;6:191.
- Bernheim A, Sorek R. The pan-immune system of bacteria: antiviral defence as a community resource. *Nat Rev Microbiol* 2020;18:113–9.
- Bobonis J, Mateus A, Pfalz B et al. Bacterial retrons encode tripartite toxin/antitoxin systems. 2020a. DOI: 10.1101/2020.06.22.160168.
- Bobonis J, Mitosch K, Mateus A et al. Phage proteins block and trigger retron toxin/antitoxin systems. 2020b. DOI: 10.1101/2020.06.22.160242.
- Callanan J, Stockdale SR, Shkoporov A et al. Expansion of known ssRNA phage genomes: from tens to over a thousand. *Sci Adv* 2020;6. DOI: 10.1126/sciadv.aay5981.
- Castelle CJ, Brown CT, Anantharaman K et al. Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat Rev Microbiol* 2018;16:629–45.
- Chalamcharla VR, Curcio MJ, Belfort M. Nuclear expression of a group II intron is consistent with spliceosomal intron ancestry. *Genes Dev* 2010;24:827–36.
- Chillón I, Martínez-Abarca F, Toro N. Splicing of the *Sinorhizobium meliloti* RmInt1 group II intron provides evidence of retroelement behavior. *Nucleic Acids Res* 2011;39:1095–104.
- Coros CJ, Piazza CL, Chalamcharla VR et al. A mutant screen reveals RNase E as a silencer of group II intron retromobility in *Escherichia coli*. *RNA* 2008;14:2634–44.
- Coros CJ, Piazza CL, Chalamcharla VR et al. Global regulators orchestrate group II intron retromobility. *Mol Cell* 2009;34:250–6.
- Crick FH. Central Dogma of Molecular Biology. *Nature* 1970;227:561–3.
- Crick FH. On protein synthesis. *Symp Soc Exp Biol* 1958;12:138–63.
- Curcio MJ, Belfort M. The beginning of the end: links between ancient retroelements and modern telomerases. *Proc Natl Acad Sci* 2007;104:9107–8.
- Dai L, Zimmerly S. Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res* 2002;30:1091–102.
- Dai L, Zimmerly S. ORF-less and reverse-transcriptase-encoding group II introns in archaeabacteria, with a pattern of homing into related group II intron ORFs. *RNA* 2003;9:14–9.
- Dinsmore PK, Klaenhammer TR. Molecular characterization of a genomic region in a *Lactococcus* bacteriophage that is involved in its sensitivity to the phage defense mechanism AbIA. *J Bacteriol* 1997;179:2949–57.
- Doron S, Melamed S, Ofir G et al. Systematic discovery of antiphage defense systems in the microbial pangenome. *Science*, 2018;359. DOI: 10.1126/science.aar4120.
- Doulatov S, Hodes A, Dai L et al. Tropism switching in *Bordetella* bacteriophage defines a family of diversity-generating retroelements. *Nature* 2004;431:476–81.
- Durmaz E, Klaenhammer TR. Abortive phage resistance mechanism AbiZ speeds the lysis clock to cause premature lysis of phage-infected *Lactococcus lactis*. *J Bacteriol* 2007;189:1417–25.
- East-Seletsky A, O'Connell MR, Knight SC et al. Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature* 2016;538:270–3.
- Eickbush TH, Jamburuthugoda VK. The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Res* 2008;134:221–34.
- Eickbush TH, Malik HS. Origins and Evolution of Retrotransposons. *Mobile DNA II Am Soc Microbiol*, 2002;1111–44.
- Eickbush TH. Telomerase and retrotransposons: which came first? *Science* 1997;277:911–2.
- Elfenbein JR, Knodler LA, Nakayasu ES et al. Multicopy single-stranded DNA directs intestinal colonization of enteric pathogens. *PLoS Genet* 2015;11:e1005472.
- Ellefson JW, Gollighar J, Shroff R et al. Synthetic evolutionary origin of a proofreading reverse transcriptase. *Science* 2016;352:1590–3.
- Emond E, Holler BJ, Boucher I et al. Phenotypic and genetic characterization of the bacteriophage abortive infection mechanism AbiK from *Lactococcus lactis*. *Appl Environ Microbiol* 1997;63:1274–83.
- Enyeart PJ, Chirieleison SM, Dao MN et al. Generalized bacterial genome editing using mobile group II introns and Cre-lox. *Mol Syst Biol* 2013;9:685.
- Enyeart PJ, Mohr G, Ellington AD et al. Biotechnological applications of mobile group II introns and their reverse transcriptases: gene targeting, RNA-seq, and non-coding RNA analysis. *Mobile DNA* 2014;5:2.
- Farzadfar F, Gharaei N, Citorik RT et al. Efficient retroelement-mediated DNA writing in bacteria. 2020. DOI: 10.1101/2020.02.21.958983.
- Farzadfar F, Lu TK. Synthetic biology. Genomically encoded analog memory with precise in vivo DNA writing in living cell populations. *Science* 2014;346:1256272.
- Ferat JL, Michel F. Group II self-splicing introns in bacteria. *Nature* 1993;364:358–61.
- Finnegan DJ. Retrotransposons. *Curr Biol* 2012;22:R432–7.
- Fortier LC, Bouchard JD, Moineau S. Expression and site-directed mutagenesis of the lactococcal abortive phage infection protein AbiK. *J Bacteriol* 2005;187:3721–30.
- Gao L, Altae-Tran H, Böhning F et al. Diverse enzymatic activities mediate antiviral immunity in prokaryotes. *Science* 2020;369:1077–84.
- García-Rodríguez FM, Hernández-Gutiérrez T, Díaz-Prado V et al. Use of the computer-retargeted group II intron RmInt1 of *Sinorhizobium meliloti* for gene targeting. *RNA Biol* 2014;11:391–401.
- García-Rodríguez FM, Neira JL, Marcia M et al. A group II intron-encoded protein interacts with the cellular replicative machinery through the β-sliding clamp. *Nucleic Acids Res* 2019;47:7605–17.
- González-Delgado A, Mestre MR, Martínez-Abarca F et al. 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 Res* 2019;47:10202–11.
- Guo H, Arambula D, Ghosh P et al. Diversity-generating retroelements in phage and bacterial genomes. *Microbiol Spectr* 2014;2. DOI: 10.1128/microbiolspec.MDNA3-0029-2014.
- Guo H, Karberg M, Long M et al. Group II introns designed to insert into therapeutically relevant DNA target sites in human cells. *Science* 2000;289:452–7.
- Guo H, Tse LV, Nieh AW et al. Target site recognition by a diversity-generating retroelement. *PLoS Genet* 2011;7:e1002414.
- Gwee CP, Khoo CH, Yeap SK et al. Targeted inactivation of *Salmonella Agona* metabolic genes by group II introns and in vivo assessment of pathogenicity and anti-tumour activity in mouse model. *PeerJ* 2019;7:e5989.
- Haack DB, Toor N. Retroelement origins of pre-mRNA splicing. *WIREs RNA* 2020;11:e1589.

- Handa S, Jiang Y, Tao S et al. Template-assisted synthesis of adenine-mutagenized cDNA by a retroelement protein complex. *Nucleic Acids Res* 2018;46:9711–25.
- Handa S, Paul BG, Miller JF et al. Conservation of the C-type lectin fold for accommodating massive sequence variation in archaeal diversity-generating retroelements. *BMC Struct Biol* 2016;16:13.
- Handa S, Shaw KL, Ghosh P. Crystal structure of a *Thermus aquaticus* diversity-generating retroelement variable protein. *PLoS One* 2019;14:e0205618.
- Hill C, Miller LA, Klaenhammer TR. Nucleotide sequence and distribution of the pTR2030 resistance determinant (hsp) which aborts bacteriophage infection in lactococci. *Appl Environ Microbiol* 1990;56:2255–8.
- Hill C, Pierce K, Klaenhammer TR. The conjugative plasmid pTR2030 encodes two bacteriophage defense mechanisms in lactococci, restriction modification (R+/M+) and abortive infection (Hsp+). *Appl Environ Microbiol* 1989;55:2416–9.
- Inouye M. The first demonstration of the existence of reverse transcriptases in bacteria. *Gene* 2017;597:76–7.
- Inouye S, Hsu MY, Eagle S et al. Reverse transcriptase associated with the biosynthesis of the branched RNA-linked msDNA in *Myxococcus xanthus*. *Cell* 1989;56:709–17.
- Iranzo J, Cuesta JA, Manrubia S et al. Disentangling the effects of selection and loss bias on gene dynamics. *Proc Natl Acad Sci* 2017;114:E5616–24.
- Iyer LM, Koonin EV, Aravind L. Evolutionary connection between the catalytic subunits of DNA-dependent RNA polymerases and eukaryotic RNA-dependent RNA polymerases and the origin of RNA polymerases. *BMC Struct Biol* 2003;3:1.
- Jackson SA, McKenzie RE, Fagerlund RD et al. CRISPR-Cas: adapting to change. *Science* 2017;356. DOI: 10.1126/science.aal5056.
- Jinek M, Chylinski K, Fonfara I et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 2012;337:816–21.
- Karberg M, Guo H, Zhong J et al. Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. *Nat Biotechnol* 2001;19:1162–7.
- Kojima KK, Kanehisa M. Systematic survey for novel types of prokaryotic retroelements based on gene neighborhood and protein architecture. *Mol Biol Evol* 2008;25:1395–404.
- Komor AC, Badran AH, Liu DR. CRISPR-based technologies for the manipulation of eukaryotic genomes. *Cell* 2017;169:559.
- Koonin EV, Krupovic M, Ishino S et al. The replication machinery of LUCA: common origin of DNA replication and transcription. *BMC Biol* 2020;18:61.
- Koonin EV, Makarova KS. Origins and evolution of CRISPR-Cas systems. *Philos Trans R Soc B Biol Sci* 2019;374:20180087.
- Lambowitz AM, Belfort M. Mobile bacterial group II introns at the crux of eukaryotic evolution. *Microbiol Spectr* 2015;3. DOI: 10.1128/microbiolspec.MDNA3-0050-2014.
- Lambowitz AM, Zimmerly S. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb Perspect Biol* 2011;3:a003616.
- Lambowitz AM, Zimmerly S. Mobile group II introns. *Annu Rev Genet* 2004;38:1–35.
- Lampson BC, Inouye M, Inouye S. Retrons, msDNA, and the bacterial genome. *Cytogenet Genome Res* 2005;110:491–9.
- Lampson BC, Sun J, Hsu MY et al. Reverse transcriptase in a clinical strain of *Escherichia coli*: production of branched RNA-linked msDNA. *Science* 1989;243:1033–8.
- LaRoche-Johnston F, Bosan R, Cousineau B. Group II introns generate functional chimeric relaxase enzymes with modified specificities through exon shuffling at both the RNA and DNA level. *Mol Biol Evol* 2020. DOI: 10.1093/molbev/msaa275.
- Leclercq S, Cordaux R. Selection-driven extinction dynamics for group II introns in Enterobacteriales. *PLoS One* 2012;7:e52268.
- Le Coq J, Ghosh P. Conservation of the C-type lectin fold for massive sequence variation in a *Treponema* diversity-generating retroelement. *Proc Natl Acad Sci* 2011;108:14649–53.
- Lim D, Maas WK. Reverse transcriptase-dependent synthesis of a covalently linked, branched DNA-RNA compound in *E. coli*. *B. Cell* 1989;56:891–904.
- Lim H, Jun S, Park M et al. Multiplex generation, tracking, and functional screening of substitution mutants using a CRISPR/Retron system. *ACS Synth Biol* 2020;9:1003–9.
- Liu M, Deora R, Doulatov SR et al. Reverse transcriptase-mediated tropism switching in *Bordetella* bacteriophage. *Science* 2002;295:2091–4.
- Lopes A, Amarir-Bouhram J, Faure G et al. Detection of novel recombinases in bacteriophage genomes unveils Rad52, Rad51 and Gp2.5 remote homologs. *Nucleic Acids Res* 2010;38:3952–62.
- López SC, Crawford KD, Bhattacharai-Kline S et al. Improved architectures for flexible DNA production using retrons across kingdoms of life. 2021. DOI: 10.1101/2021.03.26.437017.
- Makarova KS, Wolf YI, Iranzo J et al. Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol* 2020;18:67–83.
- Mao JR, Shimada M, Inouye S et al. Gene regulation by antisense DNA produced in vivo. *J Biol Chem* 1995;270:19684–7.
- Martín-Alonso S, Frutos-Beltrán E, Menéndez-Arias L. Reverse transcriptase: from transcriptomics to genome editing. *Trends Biotechnol* 2020. DOI: 10.1016/j.tibtech.2020.06.008.
- Martínez-Abarca F, Barrientos-Durán A, Fernández-López M et al. 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* 2004;32:2880–8.
- Martin W, Koonin EV. Introns and the origin of nucleus-cytosol compartmentalization. *Nature* 2006;440:41–5.
- McMahon SA, Miller JL, Lawton JA et al. The C-type lectin fold as an evolutionary solution for massive sequence variation. *Nat Struct Mol Biol* 2005;12:886–92.
- Medhekar B, Miller JF. Diversity-generating retroelements. *Curr Opin Microbiol* 2007;10:388–95.
- Mestre MR, González-Delgado A, Gutiérrez-Rus LI et al. Systematic prediction of genes functionally associated with bacterial retrons and classification of the encoded tripartite systems. *Nucleic Acids Res* 2020;48:12632–47.
- Michel F, Costa M, Westhof E. The ribozyme core of group II introns: a structure in want of partners. *Trends Biochem Sci* 2009;34:189–99.
- Michel F, Ferat JL. Structure and activities of group II introns. *Annu Rev Biochem* 1995;64:435–61.
- Millman A, Bernheim A, Stokar-Avihail A et al. Bacterial retrons function in anti-phage defense. *Cell* 2020;183:1551–61.e12.
- Mohr G, Silas S, Stamos JL et al. A reverse transcriptase-Cas1 fusion protein contains a Cas6 domain required for both CRISPR RNA biogenesis and RNA spacer acquisition. *Mol Cell* 2018;72:700–14.e8.
- Mohr G, Smith D, Belfort M et al. Rules for DNA target-site recognition by a lactococcal group II intron enable retargeting of the intron to specific DNA sequences. *Genes Dev* 2000;14:559–73.
- Mohr S, Ghanem E, Smith W et al. Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA* 2013;19:958–70.

- Mustafin RN, Khusnudinova EK. The role of reverse transcriptase in the origin of life. *Biochemistry (Moscow)* 2019;84:870–83.
- Naorem SS, Han J, Wang S et al. DGR mutagenic transposition occurs via hypermutagenic reverse transcription primed by nicked template RNA. *Proc Natl Acad Sci* 2017;114:E10187–95.
- Numkulrat S, Lee H, Doak TG et al. Genomic and metagenomic analysis of diversity-generating retroelements associated with *Treponema denticola*. *Front Microbiol* 2016;7:852.
- Nisa-Martínez R, Jiménez-Zurdo JI, Martínez-Abarca F et al. Dispersion of the RmInt1 group II intron in the *Sinorhizobium meliloti* genome upon acquisition by conjugative transfer. *Nucleic Acids Res* 2007;35:214–22.
- Nisa-Martínez R, Molina-Sánchez MD, Toro N. Host factors Influencing the retrohoming pathway of group II intron RmInt1, which has an intron-encoded protein naturally devoid of endonuclease activity. *PLoS One* 2016;11:e0162275.
- Novikova O, Belfort M. Mobile group II introns as ancestral eukaryotic elements. *Trends Genet* 2017;33:773–83.
- Novikova O, Smith D, Hahn I et al. Interaction between conjugative and retrotransposable elements in horizontal gene transfer. *PLoS Genet* 2014;10:e1004853.
- Nuñez JK, Kranzusch PJ, Noeske J et al. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat Struct Mol Biol* 2014;21:528–34.
- O'Connell MR. Molecular mechanisms of RNA targeting by Cas13-containing type VI CRISPR-Cas systems. *J Mol Biol* 2019;431:66–87.
- Odegrip R, Nilsson AS, Haggård-Ljungquist E. Identification of a gene encoding a functional reverse transcriptase within a highly variable locus in the P2-like coliphages. *J Bacteriol* 2006;188:1643–7.
- Park J, Zhang Y, Buboltz AM et al. Comparative genomics of the classical *Bordetella* subspecies: the evolution and exchange of virulence-associated diversity amongst closely related pathogens. *BMC Genomics* 2012;13:545.
- Paul BG, Burstein D, Castelle CJ et al. Retroelement-guided protein diversification abounds in vast lineages of Bacteria and Archaea. *Nat Microbiol* 2017;2:17045.
- Pickar-Oliver A, Gersbach CA. The next generation of CRISPR-Cas technologies and applications. *Nat Rev Mol Cell Biol* 2019;20:490–507.
- Ploquin M, Bransi A, Paquet ER et al. Functional and structural basis for a bacteriophage homolog of human RAD52. *Curr Biol* 2008;18:1142–6.
- Prahlad J, Yuan Y, Lin J et al. The DUF328 family member YaaA is a DNA-binding protein with a novel fold. *J Biol Chem* 2020;295:14236–47.
- Pyenson NC, Marraffini LA. Type III CRISPR-Cas systems: when DNA cleavage just isn't enough. *Curr Opin Microbiol* 2017;37:150–4.
- Qu G, Dong X, Piazza CL et al. RNA-RNA interactions and pre-mRNA mislocalization as drivers of group II intron loss from nuclear genomes. *Proc Natl Acad Sci* 2014;111:6612–7.
- Qu G, Kaushal PS, Wang J et al. Structure of a group II intron in complex with its reverse transcriptase. *Nat Struct Mol Biol* 2016;23:549–57.
- Qu G, Piazza CL, Smith D et al. Group II intron inhibits conjugative relaxase expression in bacteria by mRNA targeting. *Elife* 2018;7. DOI: 10.7554/eLife.34268.
- Roux S, Paul BG, Bagby SC et al. Ecology and molecular targets of hypermutation in the global microbiome. 2020. DOI: 10.1101/2020.04.01.020958.
- San Filippo J, Lambowitz AM. Characterization of the C-terminal DNA-binding/DNA endonuclease region of a group II intron-encoded protein. *J Mol Biol* 2002;324:933–51.
- Scaltriti E, Launay H, Genois MM et al. Lactococcal phage p2 ORF35-Sak3 is an ATPase involved in DNA recombination and AbiK mechanism. *Mol Microbiol* 2011;80:102–16.
- Scaltriti E, Moineau S, Launay H et al. Deciphering the function of lactococcal phage ul36 Sak domains. *J Struct Biol* 2010;170:462–9.
- Schillinger T, Lisfi M, Chi J et al. Analysis of a comprehensive dataset of diversity generating retroelements generated by the program DiGReF. *BMC Genomics* 2012;13:430.
- Schillinger T, Zingler N. The low incidence of diversity-generating retroelements in sequenced genomes. *Mob Genet Elem* 2012;2:287–91.
- Schmidt F, Cherepkova MY, Platt RJ. Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* 2018;562:380–5.
- Schubert MG, Goodman DB, Wannier TM et al. High throughput functional variant screens via in-vivo production of single-stranded DNA. *Proc Natl Acad Sci* 2021;118:e20181811182020.
- Sharifi F, Ye Y. MyDGR: a server for identification and characterization of diversity-generating retroelements. *Nucleic Acids Res* 2019;47:W289–94.
- Sharon E, Chen SA, Khosla NM et al. Functional genetic variants revealed by massively parallel precise genome editing. *Cell* 2018;175:544–557.e16.
- Shimamoto T, Inouye M, Inouye S. The formation of the 2',5'-phosphodiester linkage in the cDNA priming reaction by bacterial reverse transcriptase in a cell-free system. *J Biol Chem* 1995;270:581–8.
- Shipman SL, Nivala J, Macklis JD et al. CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature* 2017;547:345–9.
- Silas S, Makarova KS, Shmakov S et al. On the origin of reverse transcriptase-using CRISPR-Cas systems and their hyperdiverse, enigmatic spacer repertoires. *mBio* 2017;8. DOI: 10.1128/mBio.00897-17.
- Silas S, Mohr G, Sidote DJ et al. Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science* 2016;351:aad4234.
- Simon AJ, Ellington AD, Finkelstein IJ. Retrons and their applications in genome engineering. *Nucleic Acids Res* 2019;47:11007–19.
- Simon AJ, Morrow BR, Ellington AD. Retroelement-based genome editing and evolution. *ACS Synth Biol* 2018;7:2600–11.
- Simon DM, Zimmerly S. A diversity of uncharacterized reverse transcriptases in bacteria. *Nucleic Acids Res* 2008;36:7219–29.
- Tangney M, Fitzgerald GF. AbiA, a lactococcal abortive infection mechanism functioning in *Streptococcus thermophilus*. *Appl Environ Microbiol* 2002;68:6388–91.
- Tangney M, Fitzgerald GF. Effectiveness of the lactococcal abortive infection systems AbiA, AbiE, AbiF and AbiG against P335 type phages. *FEMS Microbiol Lett* 2002;210:67–72.
- Tanna T, Ramachanderan R, Platt RJ. Engineered bacteria to report gut function: technologies and implementation. *Curr Opin Microbiol* 2021;59:24–33.
- Tanna T, Schmidt F, Cherepkova MY et al. Recording transcriptional histories using Record-seq. *Nat Protoc* 2020;15:513–39.
- Temin HM, Mizutani S. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature* 1970;226:1211–3.
- Thiaville JJ, Kellner SM, Yuan Y et al. Novel genomic island modifies DNA with 7-deazaguanine derivatives. *Proc Natl Acad Sci* 2016;113:E1452–9.

- Toro N, Martínez-Abarca F, González-Delgado A. The reverse transcriptases associated with CRISPR-Cas systems. *Sci Rep* 2017;7:7089.
- Toro N, Martínez-Abarca F, González-Delgado A et al. On the origin and evolutionary relationships of the reverse transcriptases associated with type III CRISPR-Cas systems. *Front Microbiol* 2018;9:1317.
- Toro N, Martínez-Abarca F, Mestre MR et al. Multiple origins of reverse transcriptases linked to CRISPR-Cas systems. *RNA Biol* 2019a;16:1486–93.
- Toro N, Mestre MR, Martínez-Abarca F et al. Recruitment of reverse transcriptase-Cas1 fusion proteins by type VI-A CRISPR-Cas systems. *Front Microbiol* 2019b;10:2160.
- Toro N, Nisa-Martínez R. Comprehensive phylogenetic analysis of bacterial reverse transcriptases. *PLoS One* 2014;9:e114083.
- Toro N. Bacteria and Archaea group II introns: additional mobile genetic elements in the environment. *Environ Microbiol* 2003;5:143–51.
- Vallotta-Eastman A, Arrington EC, Meeken S et al. Role of diversity-generating retroelements for regulatory pathway tuning in cyanobacteria. *BMC Genomics* 2020;21:664.
- van der Oost J, Westra ER, Jackson RN et al. Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nat Rev Microbiol* 2014;12:479–92.
- Velázquez E, Lorenzo V, Al-Ramahi Y. Recombination-Independent genome editing through CRISPR/Cas9-enhanced TargeTron delivery. *ACS Synth Biol* 2019;8:2186–93.
- Vigouroux A, Bikard D. CRISPR tools to control gene expression in bacteria. *Microbiol Mol Biol Rev* 2020;84. DOI: 10.1128/MMBR.00077-19.
- Wahl MC, Will CL, Lührmann R. The spliceosome: design principles of a dynamic RNP machine. *Cell* 2009;136:701–18.
- Waldern J, Schiraldi NJ, Belfort M et al. Bacterial group II intron genomic neighborhoods reflect survival strategies: hiding and hijacking. *Mol Biol Evol* 2020;37:1942–8.
- Wang C, Villion M, Semper C et al. A reverse transcriptase-related protein mediates phage resistance and polymerizes untemplated DNA in vitro. *Nucleic Acids Res* 2011;39:7620–9.
- Wang GH, Seeger C. The reverse transcriptase of hepatitis B virus acts as a protein primer for viral DNA synthesis. *Cell* 1992;71:663–70.
- Wang JY, Hoel CM, Al-Shayeb B et al. Structural coordination between active sites of a Cas6-reverse transcriptase-Cas1Gas2 CRISPR integrase complex. *Nat Commun* 2021;12:2571.
- Wen Z, Lu M, Ledesma-Amaro R et al. TargeTron technology applicable in Solventogenic Clostridia: revisiting 12 years' advances. *Biotechnol J* 2020;15:1900284.
- Wolf YI, Silas S, Wang Y et al. Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nat Microbiol* 2020;5:1262–70.
- Wu L, Gingery M, Abebe M et al. Diversity-generating retroelements: natural variation, classification and evolution inferred from a large-scale genomic survey. *Nucleic Acids Res* 2018;46:11–24.
- Yan F, Yu X, Duan Z et al. Discovery and characterization of the evolution, variation and functions of diversity-generating retroelements using thousands of genomes and metagenomes. *BMC Genomics* 2019;20:595.
- Yee T, Furuichi T, Inouye S et al. Multicopy single-stranded DNA isolated from a gram-negative bacterium, *Myxococcus xanthus*. *Cell* 1984;38:203–9.
- Ye Y. Identification of diversity-generating retroelements in human microbiomes. *Int J Mol Sci* 2014;15:14234–46.
- Yosef I, Goren MG, Qimron U. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res* 2012;40:5569–76.
- Yuan TZ, Overstreet CM, Moody IS et al. Protein engineering with biosynthesized libraries from *Bordetella bronchiseptica* bacteriophage. *PLoS One* 2013;8:e55617.
- Zhao C, Liu F, Pyle AM. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA* 2018;24:183–95.
- Zhao C, Pyle AM. Crystal structures of a group II intron maturase reveal a missing link in spliceosome evolution. *Nat Struct Mol Biol* 2016;23:558–65.
- Zhuang F, Karberg M, Perutka J et al. EciS, a group IIB intron with high retrohoming frequency: DNA target site recognition and use in gene targeting. *RNA* 2009;15:432–49.
- Zimmerly S, Wu L. An unexplored diversity of reverse transcriptases in bacteria. *Microbiol Spectr* 2015;3:MDNA3-0058-2014.