CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS
ESTACIÓN EXPERIMENTAL DEL ZAIDÍN

UNIVERSIDAD DE GRANADA

The Reverse Transcriptases associated with CRISPR-Cas systems: Phylogenetic relationships and functional characterization.

TESIS DOCTORAL
Programa de Doctorado en Biología Fundamental y de Sistemas

Alejandro González Delgado
2021

The Reverse Transcriptases associated with CRISPR-Cas Systems: Phylogenetic relationships and functional characterization.

Memoria que presenta el graduado en Biología D. Alejandro González Delgado, como aspirante al título de Doctor con mención internacional.

Fdo: Alejandro Gonález Delgado

Vº Bº de los directores de Tesis Doctoral

Fdo: Dr. Nicolás Toro García
Doctor en Ciencias Biológicas
Profesor de investigación del CSIC

Fdo: Dr. Francisco Martínez-Abarca Pastor
Doctor en Ciencias Biológicas
Investigador Científico del CSIC

Universidad de Granada
2021

Editor: Universidad de Granada. Tesis Doctorales
Autor: Alejandro González Delgado
ISBN: 978-84-1306-822-0
URI: http://hdl.handle.net/10481/67921

El doctorando / The *doctoral candidate* **[Alejandro González Delgado]** y los directores de la tesis / and the thesis supervisor/s: **[Nicolás Toro García and Francisco Martínez-Abarca Pastor]**

Garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección de los directores de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

/

Guarantee, by signing this doctoral thesis, that the work has been done by the doctoral candidate under the direction of the thesis supervisor/s and, as far as our knowledge reaches, in the performance of the work, the rights of other authors to be cited (when their results or publications have been used) have been respected.

Lugar y fecha / Place and date:

Granada, December 15, 2020

Director/es de la Tesis / *Thesis supervisor/s;*

TORO GARCIA NICOLAS - 24155017A

Firmado digitalmente por TORO GARCIA NICOLAS - 24155017A
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-24155017A, givenName=NICOLAS, sn=TORO GARCIA, cn=TORO GARCIA NICOLAS - 24155017A
Fecha: 2020.12.15 10:57:49 +01'00'

MARTINEZ-ABARCA PASTOR FRANCISCO - 27440188S

Firmado digitalmente por MARTINEZ-ABARCA PASTOR FRANCISCO - 27440188S
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-27440188S, givenName=FRANCISCO, sn=MARTINEZ-ABARCA PASTOR, cn=MARTINEZ-ABARCA PASTOR FRANCISCO - 27440188S
Fecha: 2020.12.15 12:44:56 +01'00'

Firma / Signed

Doctorando / *Doctoral candidate*:

GONZALEZ DELGADO ALEJANDRO - 71032306A

Firmado digitalmente por GONZALEZ DELGADO ALEJANDRO - 71032306A
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-71032306A, givenName=ALEJANDRO, sn=GONZALEZ DELGADO, cn=GONZALEZ DELGADO ALEJANDRO - 71032306A
Fecha: 2020.12.17 18:25:46 +01'00'

Firma / Signed

La presente Tesis Doctoral ha sido realizada en el grupo de investigación de Ecología Genética de la Rizosfera perteneciente al Departamento de Microbiología y Sistemas Simbióticos de la Estación Experimental del Zaidín (CSIC).

Para la realización de este trabajo, el doctorando Alejando González Delgado ha sido financiado por las siguientes fuentes:

- Ayuda para contrato predoctoral para la Formación del Profesorado Universitario (FPU, ref. FPU15/02714), concecido por el Ministerio de Educación Cultura y Deporte (2016-2021).
- Ayuda para la realización de estancias breves asociadas al contrato FPU (ref. EST18/00395), en School of Life Sciences de la University of Nottingham (Reino Unido bajo la supervisión del Dr. Edward L. Bolt (Abril-Julio 2019).

Parte de los resultados presentados en esta Tesis Doctoral han sido publicados en revisas internacionales:

- Toro, N., Martínez-Abarca, F., **González-Delgado, A**. (2017). The reverse transcriptases associated with CRISPR-Cas systems. *Sci Rep* 7, 7089.

- Toro, N., Martínez-Abarca, F., **González-Delgado, A**. Mestre, M. R. (2018). On the origin and evolutionary relationships of the reverse transcriptases associated with type III CRISPR-Cas systems. *Front Microbiol* 9, 1317.

- Toro, N., Martínez-Abarca, F., Mestre, M. R. **González-Delgado, A**. (2019). Multiple origins of reverse transcriptases linked to CRISPR-Cas systems. *RNA Biol* 16, 1486–1493 (2019).

- **González-Delgado, A.**, Mestre, M. R., Martínez-Abarca, F. Toro, N. (2019). Spacer acquisition from RNA mediated by a natural reverse transcriptase-Cas1 fusion protein associated with a type III-D CRISPR-Cas system in *Vibrio vulnificus*. *Nucleic Acids Res* 47, 10202–10211.

- Toro, N., Mestre, M. R., Martínez-Abarca, F., **González-Delgado, A**. (2019). Recruitment of reverse transcriptase-Cas1 fusion proteins by type VI-A CRISPR-Cas systems. *Front Microbiol* 10, 2160.

Adicionalmente, durante el periodo de Tesis Doctoral el doctorando ha colaborado en las siguientes **publicaciones** y participado en **congresos**:

Publicaciones:

- Mestre, M. R., **González-Delgado, A.**, Gutiérrez-Rus, L. I., Martínez-Abarca, F., Toro, N. (2020). Systematic prediction of genes functionally associated to bacterial retrons and classification of the encoded tripartite systems. *Nucleic Acids Res,* gkaa1149.
- González-Delgado, A., Mestre M.R., Martínez-Abarca, F., Toro, N. (2020). Prokaryotic Reverse Transcriptases: from Retroelements to Specialized Defense Systems. *Submitted.*

Participación en **Congresos:**

- **González-Delgado, A.** Las reverso transcriptasas asociadas a los sistemas CRISPR-Cas. Póster de las II Jornadas de Investigadores en formación: Fomentando la Interdisciplinariedad (JIFFI). Granada, 17-19/05/2017.
- **González-Delgado, A.**, Martínez-Abarca, F. Toro, N. The Reverse Transcriptases associated with CRISPR-Cas systems. Poster en el 7º Congreso Europeo de la FEMS. Valencia, 09-13/07/2017
- **González-Delgado, A.**, Martínez-Abarca, F. Toro, N, Mestre, M.R. On the origin(s) and evolutionary relationships of the reverse transcriptases associated with type III CRISPR-Cas systems. Poster en el 12º CRISPR Internacional Meeting. Vilnius (Lituania), 20-23/06/2018.
- **González-Delgado, A.**, Mestre, M. R., Martínez-Abarca, F. Toro, N. 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.* Poster en el 25º Annual Meeting of the RNA society. On-line, 26/05 al 02/06/2020.
- **González-Delgado, A.**, 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.* Primer Premio MicroMolecular Granada-Octubre 2020. Exposición on-line, 21/10/2020.

A mis padres

Acknowledgements/ Agradecimientos

Cuando echo la vista atrás, a mi llegada a Granada, mi mente toma conciencia del gran vuelco que ha experimentado mi vida en los últimos 5 años y medio. Esta etapa predoctoral ha sido dura, con momentos difíciles, pero finalmente ha desembocado en un gran crecimiento, científico, sí, pero sobre todo a nivel personal. Es por ello que no hay mejor manera de empezar esta tesis que expresando mi eterna gratitud a todas las personas que han formado parte de este camino.

En primer lugar, me gustaría dar las gracias a mis directores de tesis, el Dr. Nicolás Toro García y el Dr. Francisco Martínez-Abarca, que me han brindado la posibilidad de iniciarme en el mundo de la ciencia trabajando en una temática que realmente me apasiona.

Especialmente me gustaría dar las gracias a Paco por su total implicación y por su disponibilidad en todo momento, aunque las aguas estuvieran revueltas. Tras todas y cada una de las intensas horas de discusiones en tu despacho, muchas veces que se alargaban durante el café e incluso las interminables llamadas telefónicas durante la pandemia que terriblemente nos ha tocado vivir, puedo afirmar con total rotundidad que no he podido tener mejor mentor en esta etapa. Gracias de todo corazón.

I would also like to express my sincere gratitude to Dr. Ed L. Bolt for giving me the opportunity to work in his lab during my stay at the Nottingham University, as well as the rest of lab members for their help at all times.

A la Dra. Carmen Amaro y a la Dra. Eva Sanjuán, que desde el primer momento estuvieron dispuestas a escuchar nuestras ideas y que nos ayudaron a completar los ensayos de interferencia con *Vibrio.*

También me gustaría agradecer a todos los miembros del tribunal de la presente tesis (el Dr. Luis Menéndez-Arias, el Dr. Roberto de la Herrán, la Dra. Isabel Chillón, el Dr. Antonio Sanchéz-Amat y la Dra. Maria Trinidad Gallegos) por su disposición a formar parte de este proceso, así como a los dos expertos que se han encargado de validar internacionalmente la investigación aquí presentada (el Dr. José R. Penadés y el Dr. Franklin L. Nobrega).

A todos y cada uno de los miembros del grupo de Genómica de Rizobacterias (Sensi, Rafa, Chema, Natalia, Tita, José Ignacio, Marta, Pablo, Manolo, Hurry, Anita, Nuria), sin los cuáles todo el trabajo realizado en esta tesis habría sido imposible y que desde el primer momento habéis hecho que un zamorano se sintiera casi como en casa, pese a los 700 km de distancia. Especialmente a Fernando, por adentrarme en el increíble (y tortuoso) mundo de las proteínas y estar siempre dispuesto a resolver mis innumerables dudas. Y, sobre todo, a Lola, la mejor compañera de lab que se he podido tener. Echaré de menos nuestro pequeño reducto, el laboratorio rojo, hogar de charlas científicas y no tan científicas, pero donde siempre has estado para escucharme y ayudarme a ver el vaso medio lleno.

A todo el personal, tanto científico como no científico, de la Estación Experimental del Zaidín. que siempre me lo han puesto todo fácil. Especialmente a mi tutora durante esta etapa, la Dra. Socorro Mesa, que ha resuelto con celeridad todas las cuestiones que han surgido durante estos años.

A mi compañero de beca y amigo, Sergio Parejo, ya que juntos hemos luchado contra las trabas de la burocracia y con otros problemas que nos ha puesto delante la vida estos años y que, poniéndole una pizca de alegría a la vida, hemos sabido superar. A Mary que, ya fuera para una Delirium o para una ruta, siempre ha estado al pie del cañón.

Pero sin duda tengo que agradecer a la mejor persona que me ha dado la ciencia, Mario (a.k.a M.R. Mestre). Gracias por tu ímpetu, por ser una fuente inagotable de ideas, en definitiva, por ser como eres, un diamante en bruto. Espero que nuestro camino científico siga muchos años de la mano, por que como amigo ya me tienes para toda la vida.

Gracias a todas las personas que me ha dado Granada. A mi gente del máster, que hicisteis que mi primer año aquí estuviera cargado de felicidad. Muchos de los recuerdos más bonitos que me llevo son con vosotros. Al único e inigualable Camacho, que hizo que no me volviera loco durante el confinamiento, con ese peculiar humor manchego impredecible y que me enseño a disfrutar aún más del séptimo arte. A Lydia, que con su luz vino a iluminar mi momento de mayor oscuridad y que se ha convertido en uno de los mayores pilares de mi vida. Gracias amiga por estar siempre cuando te necesito.

A mis biólogos. Hace cinco años escribí unas palabras que ahora, si cabe, tienen más peso para mí: "Si me preguntaran que es lo mejor de la carrera, no tengo ni una sola duda en la respuesta, vosotros. La Biología es la ciencia de la vida y vosotros me habéis enseñado a vivir". Gracias por las innumerables visitas y por las mil y una historias que hemos vivido. Especialmente a Deju, la persona más mítica y carismática que conozco, mi compañero de cervezas y de fiestas hasta al amanecer. Y por supuesto a mi extremeño, mi Splash Brother, Roberto. Contigo he reído, he llorado, y he sentido todo con más intensidad. Con vosotros todo es mejor.

A los de toda la vida. Mis zamoranos, quizá cada vez nos veamos menos, pero eso solo hace que los momentos que pasamos juntos sean aún mas especiales. Todos los viajes, anécdotas y risas de estos años han sido el motor que me ha dado la fuerza para seguir. En especial a Alberto, el impacto que has tenido en mi vida no se mide con palabras, gracias por haber crecido junto a mi. A mi leonés pelirrojo, Mario, contigo de la plazuela al cielo. A Naza, mi best, 20 años desde que empezamos a compartir la banda izquierda y desde entonces siempre a mi lado, siendo una constante, manteniéndome en pie pese a todos los baches.

A mi familia, de la que no puedo sentirme más orgulloso. Y, por encima de todo, a mis padres. Vosotros, que tanto habéis luchado por que no me falte de nada, que nunca habéis dejado que me sintiera solo, que me habéis levantado cuando me he caído. Todo lo que soy, todo lo que he conseguido, es gracias a vosotros. Sois la gran suerte de mi vida.

Summary

Reverse transcriptases (RTs) are enzymes capable of synthesizing DNA from an RNA template. RTs were discovered in 1970 in RNA tumor viruses, although they were not found in prokaryotes until 1989. Currently, prokaryotic RTs display an extraordinary diversity with most of them classified into three main groups: group II introns, retrons and diversity-generating retroelements (DGRs). Phylogenetic analysis has revealed a group of RTs closely related to CRISPR-Cas systems (Clustered Regularly Interspaced Short Palindromic Repeats and CRISPR-associated proteins), an adaptive immune system harbored by prokaryotes to protect against bacteriophages and other mobile genetic elements.

The immunity mediated by CRISPR-Cas system present three stages: adaptation, expression and interference. The adaptive stage involves the integration of small DNA fragments, known as spacers, from invader elements (phages and plasmids) within the so-called CRISPR Arrays, keeping this information as a memory to facilitate future defense. In the expression stage, the inserted DNA fragments are transcribed and processed to generate a mature RNA which act as guide for the degradation of the invader genetic material by the effector machinery during the interference stage.

Analysis of the genomic context of the RTs associated with CRISPR-Cas systems has revealed that these enzymes are located upstream of *cas1* and *cas2* genes, which encodes the proteins responsible for the adaptive stage. Thus, a plausible hypothesis is that these RTs facilitates the acquisition of novel spacers from RNA phages or highly transcribed regions. With this background, the objective of this thesis consisted of the characterization of novel groups of RTs linked to CRISPR-Cas systems.

An initial analysis of the different distribution of RTs in bacteria and archaea indicates that specific prokaryotic groups have recruited particular types of RTs in order to improve their responses in the environmental conditions of their ecological niche. Then, further analysis focused on the RTs related to CRISPR-Cas systems clustering them in 15 phylogenetic clades which at least present three evolutionary origins supporting a “multiple origins” model. These RTs are found alone, fused to the C-termini of the Cas1 domain (RTCas1) or as a multidomain protein (Cas6RTCas1). Furthermore, most RTs are associated with type III CRISPR-Cas systems, whereas a few examples are related to types I-A and VI.

The first criterion for the functional characterization of RT-containing CRISPR-Cas systems was to determine the exogenous RT activity *in vitro* of several RT homologs. Then, the study focused on two systems: the adaptive operon from the Cyanobacterium *Scytonema hofmanni* PCC 7110, which harbor an RT alone, and that from the Proteobacterium *Vibrio vulnificus* YJ016, that contain an RTCas1 fusion protein. In both systems, the purification of the different proteins of the adaptive operon allows to study their interactions, demonstrating that the two different *V. vulnificus* Cas2 proteins, Cas2A and Cas2B, form a stable heterodimer complex.

In vivo assays revealed that the RTCas1-Cas2A-Cas2B adaptive operon from *V. vulnificus* YJ016 is able to acquire spacers in a heterologous host *(E. coli)* in a process in which the two different Cas2 are required. Moreover, mutation in the RT active site strongly impaired novel acquisition events. It has been also demonstrated that this system is able to acquire spacers directly from RNA. In addition, the analysis of the acquired spacers and their sequence revealed particular features that may be related to a specific recognition by the adaptive operon. Finally, assays in the natural host have confirmed that the *V. vulnificus* type III-D CRISPR-Cas interference module is functional as long as their DNA target is transcriptionally active.

Resumen

Las transcriptasas inversas (RTs, del inglés: Reverse Transcriptases) son enzimas capaces de sintetizar ADN a partir de un molde de ARN. Estas enzimas fueron descubiertas por primera vez en virus de ARN en 1970, pero no fueron descritas en organismos procariotas hasta 1989. Actualmente se conocen gran diversidad de RTs en los genomas de bacterias y arqueas, perteneciendo la mayoría a tres grupos principales: los intrones del grupo II, los retrones y los retroelementos generadores de diversidad (DGRs). Análisis filogenéticos han revelado la existencia de un grupo de RTs estrechamente asociado a los sistemas CRISPR-Cas (de las siglas en inglés: Repeticiones Cortas Palindrómicas Agrupadas y Regularmente Interespaciadas – y proteínas Cas asociadas), un sistema inmune adaptativo presente en procariotas para defenderse frente a fagos y otros elementos genéticos móviles.

La inmunidad mediada por los sistemas CRISPR-Cas presenta tres etapas: adaptación, expresión e interferencia. La etapa adaptativa consiste en la integración de pequeños fragmentos de ADN, conocidos como espaciadores, de los elementos invasores (fagos o plásmidos) en los denominados CRISPR Arrays, que almacenan en el genoma la información de los diferentes eventos a los que se ha enfrentado la bacteria. Durante la etapa de expresión, los fragmentos de ADN insertados en el array son transcritos y procesados generando un ARN que servirá de guía para que, en la etapa interferencia, la maquinaria efectora de estos sistemas degrade el material genético del agente en futuras invasiones.

El análisis del contexto genómico de las RTs asociadas a estos sistemas ha puesto de manifiesto que se encuentran justo aguas arriba de los genes *cas1* y *cas2*, que codifican las proteínas responsables de la etapa adaptativa de los sistemas CRISPR-Cas. Entonces, una hipótesis plausible es que estas RTs participen en la adquisición de nuevos espaciadores procedentes de fagos de ARN o de regiones altamente transcritas. Con estos antecedentes, el objetivo planteado en esta tesis doctoral consistió en la caracterización de nuevos grupos de RTs haciendo especial hincapié en aquellos asociados a los sistemas CRISPR-Cas.

Un primer análisis de la distribución diferencial de estas enzimas entre bacterias y arqueas reveló que determinados grupos de procariotas han reclutado ciertos tipos de RTs para mejorar su respuesta a las condiciones ambientales de su nicho ecológico. Posteriormente, el análisis se focalizó en las RTs asociadas a los sistemas CRISPR-Cas clasificándolos en 15 clados filogenéticos con al menos tres orígenes evolutivos diferentes apoyando un modelo de "múltiples orígenes" para esta asociación. Estas RTs se pueden encontrar solas, fusionadas en su extremo C-terminal a un dominio Cas1 (RTCas1) o formando parte de una proteína multidominio (Cas6RTCas1). Además, la mayoría de estas enzimas forman parte de sistemas CRISPR-Cas de tipo III, aunque también existen ejemplos minoritarios asociados a tipo I-A y a tipo VI.

Para la caracterización funcional de los sistemas CRISPR-Cas que contienen RTs, primero se determinó la actividad RT exógena *in vitro* de diversos candidatos, centrando finalmente el estudió en dos sistemas: el operón adaptativo de la Cianobacteria *Scytonema hofmanni* PCC 7110, que contiene una RT sin fusionar, y el de la Proteobacteria *Vibrio vulnificus* YJ016, que presenta una fusión RTCas1. En ambos sistemas, la purificación de las distintas proteínas pertenecientes al operón adaptativo ha permitido estudiar su interacción, demostrando como resultado más relevante que las proteínas Cas2A y Cas2B del sistema de *V. vulnificus* forman un complejo heterodimérico estable.

Ensayos *in vivo* demostraron que el operón adaptativo (RTCas1-Cas2A-Cas2B) de *V. vulnificus* YJ016 es capaz de adquirir espaciadores en un hospedador heterólogo (*E. coli*) en un proceso que requiere la presencia de las dos Cas2. Además, la adquisición se reduce drásticamente mediante mutaciones en el sitio activo de la RT. También se verificó que este sistema es capaz de adquirir espaciadores directamente de moléculas de ARN. Por otro lado, el análisis de los espaciadores adquiridos y de su secuencia reveló características particulares que podrían estar relacionadas con su reconocimiento específico por parte del operón adaptativo. Finalmente, ensayos en el hospedador natural han revelado que el módulo interferente de este sistema CRISPR-Cas de tipo III-D es activo, requiriendo para ello la transcripción de su ADN diana

Table of contents

Table of contents

List of figures

Introduction

Material and Methods

Results

Chapter 1

Chapter 2

Chapter 3

List of tables

Introduction

Material and Methods

Results

Chapter 1

Appendix

Introduction

I.1. Prokaryotic Reverse Transcriptases

Enunciated by Francis Crick in 1958, the Central Dogma of molecular biology once stated that the genetic information flows unidirectionally from DNA to RNA to protein molecules (Crick, 1958). However, the discovery of enzymes capable of synthesizing DNA from an RNA template challenged this over-simplified dogma, demonstrating that the flow of information can be reversed. 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 their existence reshaped existing views on all forms of life function.

In viruses, RTs play a crucial role in the replication of different families such as *Retroviridae, Metaviridae, Pseudoviridae, Hepadnaviridae* and *Caulimoviridae* (Menéndez-Arias *et al.*, 2017). RTs have also been found a variety of eukaryotic elements, including long terminal repeat (LTR) and non-LTR retroelements, Penelope-like elements and telomerase (Eickbush and Jamburuthugoda, 2008; Finnegan, 2012). Prokaryotic RTs also 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. 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).

Interestingly, amino acid sequence alignments of RTs from the three domains of life have revealed that all of them are phylogenetically related sharing a common domain architecture characterized by a series of motifs, typically from RT0 to RT7 (Xiong and Eickbush, 1990; Zimmerly and Wu, 2015). Despite the lack or presence of the motifs that characterize every group, all RTs clearly align across RT motifs 3, 4 and 5, which forms the core polymerase structure corresponding to the palm and finger domains. This core includes the three aspartate residues necessary for the catalytic activity of the RTs (Figure I.1). Most RTs also contain a recognizable motif 6, with a conserved lysine also required for the RT activity (Castro *et al.*, 2009).

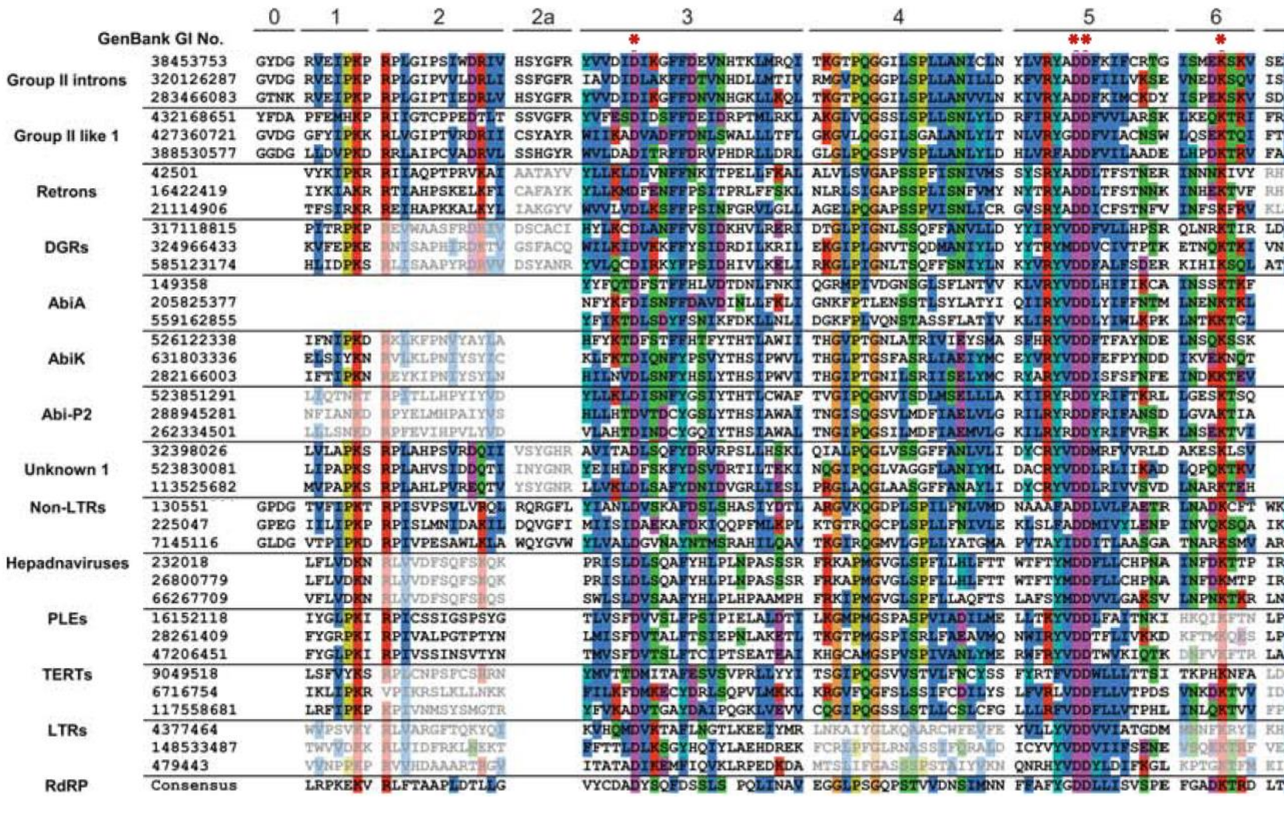


Figure I.1 Amino acid alignment of RT0-7 motifs for different groups of RTs in all domains of life. Three example sequence for each group. Sequences in black lettering and bold color shading are clearly aligned, while sequences in gray and light co ambiguously aligned. Positions with >30% identity across the entire alignment are back-shaded in colors to highlight the most con across RT classes. The consensus sequence for the Pfam group, RNA-dependent RNA polymerase (RdRP) is also indicated. Aste alignment mark the three catalytic aspartate residues in motifs 3 and 5 and the active site lysine in motif 6. TR, long terminal repeat; P like elements; TERT, telomerase reverse transcriptases (Adapted from Zimmerly and Wu, 2015).

While the RTs from eukaryotes and viruses have been extensively characterized, the function and mechanism of the great variety of RTs in prokaryotic organisms is only beginning to be understood. In fact, more recent investigations have shown that many prokaryotic RT lineages have evolved to provide new ways of facing phages. Thus, together with Abi-RTs (Fortier *et al.*, 2005; Odegrip, *et al.*, 2006; Durmaz and Klaenhammer, 2007) and CRISPR-RTs (Kojima and Kanehisa, 2008; Toro and Nisa-Martínez, 2014), the search for new antiviral immunity systems in defense islands had led to the discovery that some retron types along with several RTs from uncharacterized groups (UG) have anti-phage properties (Gao *et al.*, 2020; Millman *et al.*, 2020).

Group II introns RTs are the best-characterized in prokaryotes (Ferat and Michel, 1993; Lambowitz and Zimmerly, 2011), however, their ecological role is barely understood. Historically considered a selfish unit, a few studies provide evidence that these mobile retroelements are capable of disrupting other mobile genetics elements (MGE), participating in host defense against potentially harmful elements (Chillón *et al.*, 2011; Qu *et al.*, 2018). In the instance of DGRs, they present the ability to cause high-sequence variation through a directed-mutagenesis reaction providing an adaptive advantage in their host (Liu *et al.*, 2002; Wu *et al.*, 2018). With an open range of functions yet to be elucidated, DGRs participate in tropism switching and signaling pathways that could be involved in virus-host interactions as well. The relevance of prokaryotic RTs also lies in the fact that different RT lineages are being used as promising tools in distinct fields including genome editing, genetic engineering applications and recording biological information in bacterial genomes (Belfort and Lambowitz, 2019; Simon *et al.,* 2019; Schmidt *et al.*, 2018). These potential uses have boosted interest in these prokaryotic enzymes and their associated systems.

Therefore, the current knowledge of prokaryotic RTs suggests that most groups has been domesticated to perform immunity functions in the host cell showing a wide range of biological mechanisms that requires in-depth research to be fully understood. In the next sections, it will be illustrated the biology of the different types of prokaryotic RTs and their role within specialized systems as well as a summary of the use of these enzymes in cutting-edge technologies. Special focus will be performed in the particular association between RTs and CRISPR-Cas systems as the main subject of this study.

I.1.1 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, being the most numerous of RT types in bacteria representing almost 50% of total RT diversity (Toro *et al.*, 2019a). The fact that bacterial group II Introns have generally inserted outside of essential genes or into non-essential genes presumably reflects that introns are deleterious to the host (Leclercq and Cordaux, 2012) in fact albeit several exceptions, introns are typically found only at one or two copies per genome (Dai and Zimmerly, 2002).

This tendency for group II introns to be deleterious may drive to domesticate the RT-containing genes, and their conversion into non-mobile RTs, presumably with novel functions (Figure I.2). Indeed, 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 (Lambowitz and Belfort, 2015; Novikova and Belfort, 2017, Hack and Toor., 2020). Moreover, it has also been suggested that group II intron proliferation in primitive eukaryotic cells could stimulate the formation of a nuclear envelope in order to separate splicing from translation (Martin and Koonin, 2006).

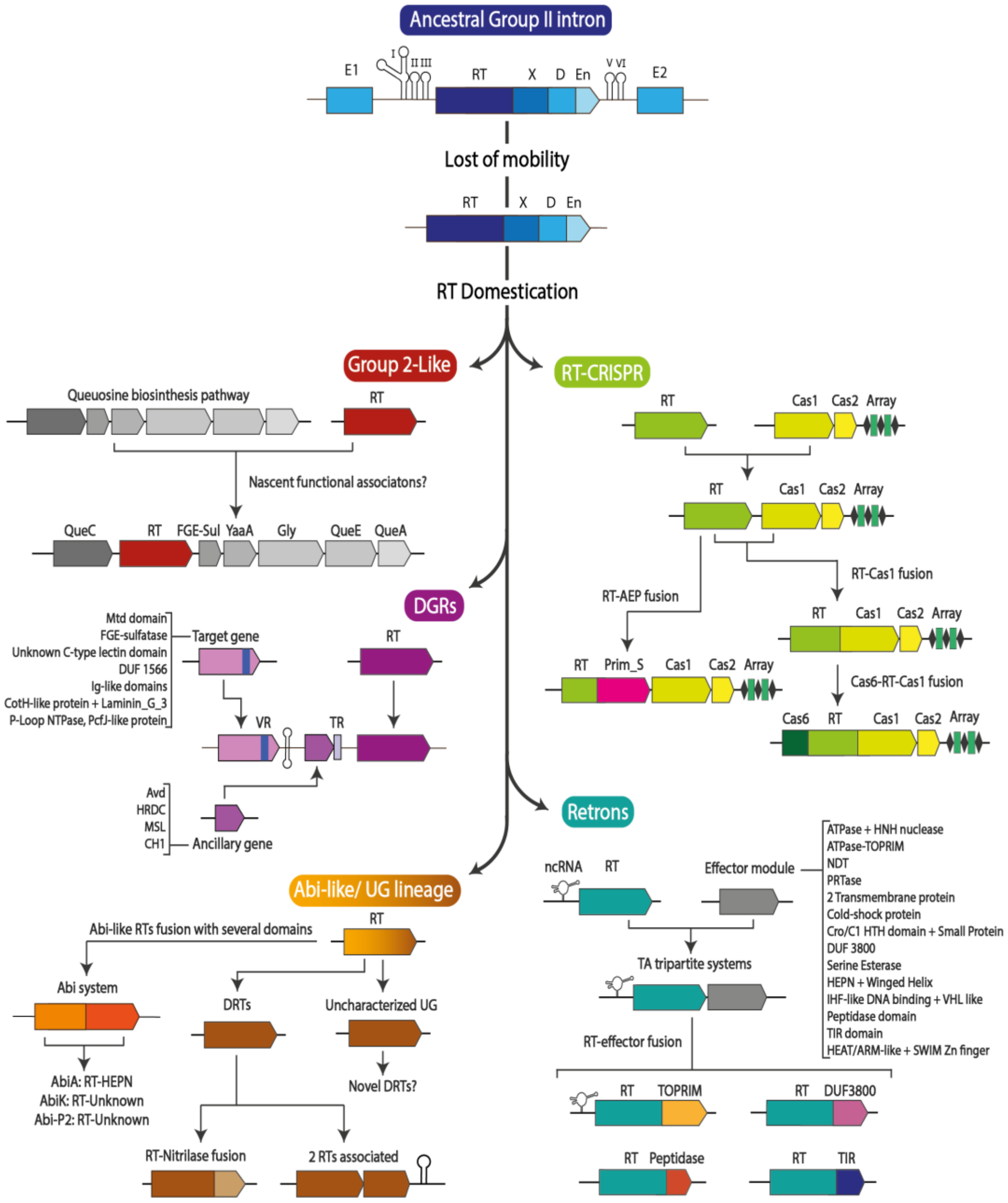
Group II introns consist of a catalytic RNA, which has a characteristic conserved 5′- and 3′-end sequences, GUGYG and AY, respectively, resembling those of spliceosomal eukaryotic introns, and an Intron-encoded Protein (IEP) (Lambowitz and Zimmerly, 2011). The catalytic intron RNA presents a conserved secondary structure, ranging from 400 to 800 nts which is organized into six domains, DI-VI, radiating from a central “wheel” (Michel *et al.*, 2009). The IEP presents different domains implicated in group II Intron retromobility (Lambowitz and Zimmerly, 2004): a N-terminal RT domain, an X domain (maturase) involved in facilitating RNA splicing, a D domain involved in DNA binding and, in some cases, an extra 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 lacking the endonuclease domain. The best-studied of the latter is *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 2 transesterification reactions, the first of which starts with the nucleophilic attack of 2’-OH of a bulged adenosine in DVI, with the second reaction resulting in the genesis of an excised intron lariat and the ligation of exons (Figure I.3A). To facilitate this stage, the IEP functions as a maturase, with the RT and X 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 synthetize the intron cDNA of the reverse-spliced intron RNA into 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 groups II introns, the ecological implications of these retroelements in their host cell are poorly understood, with only a few studies published to date. Group II introns, which are considered as selfish elements, tend to localize at higher densities on plasmids than 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 acquisition of the RmInt1 group II intron by conjugative transfer the colonization of the homing sites, typically insertions sequences (IS*Rm2011-2* and closed 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 control the spread of other MGEs (Chillón *et al.*, 2011). These findings suggest that group II introns may have an evolutionary role in circumventing efficient splicing and preventing the mobility of harmful elements in the bacterial cell becoming a particular defense system.

Consistent with these results, another group II intron integrated into a relaxase gene on a conjugative plasmid have been shown to inhibit its host gene expression and restrain the naturally cohabiting mobile element from the conjugative horizontal transfer by decreasing the levels of spliced mRNA level (Qu *et al.*, 2018). This process seems to function as a defense barrier, limiting the spread of other mobile elements acting as general inhibitors of gene expression. However, the relaxase stimulates intron dispersion by nicking the conjugative plasmid and the chromosome (Novikova *et al.*, 2014). Thus, the relaxase facilitate plasmid dispersion and retrotransposition events, whereas the group II intron regulates relaxase expression, maintaining a balance that may be positive for the host.

A recent study shows that group II introns can increase genetic diversity creating chimeric relaxases variants through the shuffling coding sequences at RNA and DNA level, thereby showing that these retroelements could be beneficial to the conjugative elements that harbor them and to their bacterial host (LaRoche-Johnston *et al.*, 2020). Moreover, the existence of host factors that act as global regulators of intron mobility, some of them as depressors, such as RNAse E, whereas others as stimulators, including alarmones ppGpp and cAMP, demonstrate the role of nutritional stress in the activation of these retroelements (Coros *et al.*, 2008; Coros *et al.*, 2009; Nisa-Martínez *et al.*, 2016). Thus, Group II introns may act by preventing the damage produced by others MGEs activated by stress conditions.



(Figure Legend in the next page)

Figure I.2. RT domestication from an ancestral selfish retroelement. The scheme depicts here a hypothetical scenario of the domestication of the different RT lineages from an autonomous mobile retroelement, probably an ancestral group II intron. 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 introns and a domestication of these RTs leading to nascent functional associations. 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 target genes with a specific fold domain and are typically assisted by various ancillary proteins. In the case of retrons, RTs have become associated with small ncRNAs and an effector module, 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; 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 it this lineage may represent an old domestication event, creating a new mechanism of defense against phages. In the Abi lineage, 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.

Although the findings shown in this section support the hypothesis 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.

.

I.1.2. Retrons

In 1984, a small DNA, known as multicopy single-stranded DNA (msDNA) was found to accumulate to high levels in the bacterium *Mixococcus xanthus* (Yee *et al.*, 1984). Subsequently, it was demonstrated that the 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). Although the biochemical characterization of retrons has been deeply studied (Lampson, *et al.*, 2005; Simon *et al.*, 2019), the biological role of retron has remained unknown for over 30 years after their discovery. 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 reports show that the msDNA molecule is crucial for the antiviral activity of retron systems.

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 *et al.*, 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, to facilitate the 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 *et al.*, 2005; Simon *et al.*, 2019).

The recent expansion of the diversity of known retrons based on genome survey analyses has made it possible to increase from tens to thousands of the number of putative retrons, most of them containing the characteristic 'VTG' signature in the RT 7 motif (Figure I.1; Toro *et al.*, 2019a). About a third of annotated retrons were thought to encode an ancillary gene (Simon *et al.*, 2019), a computer pipeline designed for the systematic prediction of genes specifically associated with retrons has revealed that most of them present an additional component, as an independent gene or a RT-fused domain. Thus, retrons should be considered as tripartite systems (Figure I.2; Mestre *et al.*, 2020).

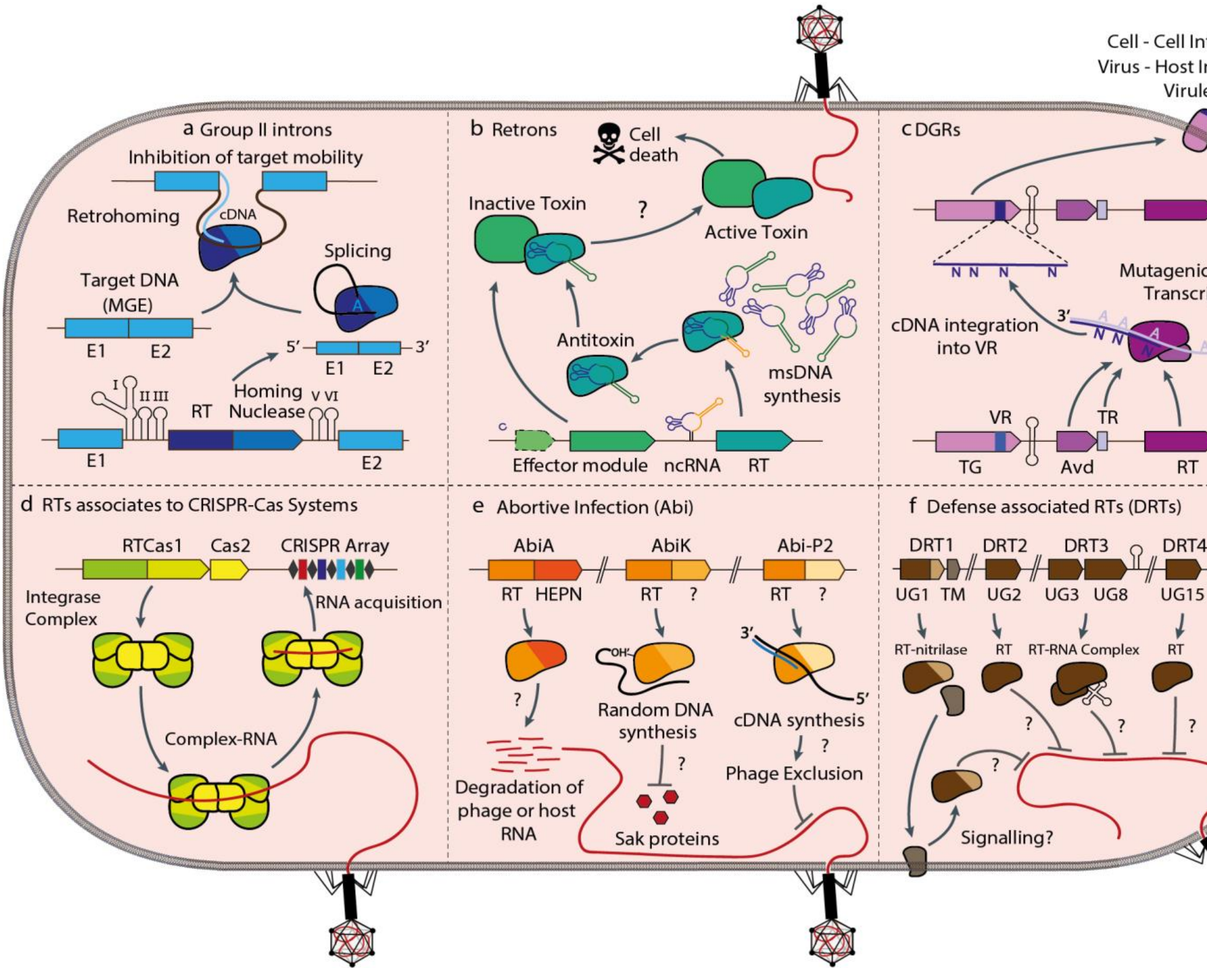

Cell - Cell In
Virus - Host In
Virul
a Group II introns
Inhibition of target mobility
Retrohoming
cDNA
Splicing
Target DNA
(MGE)
E1
E2
A
5′
3′
E1
E2
I
II III
RT
Homing
Nuclease
V VI
E1
E2
b Retrons
Cell
death
Inactive Toxin
?
Active Toxin
Antitoxin
msDNA
synthesis
Effector module
ncRNA
RT
c DGRs
N N
N
N
Mutagenic
Transcr
cDNA integration
into VR
3′
A
A
N N
A
N
VR
TR
TG
Avd
RT
d RTs associates to CRISPR-Cas Systems
RTCas1
Cas2
CRISPR Array
Integrase
Complex
RNA acquisition
Complex-RNA
e Abortive Infection (Abi)
AbiA
AbiK
Abi-P2
RT
HEPN
RT
?
RT
?
?
OH
3′
5′
Random DNA
synthesis
cDNA synthesis
?
?
Degradation of
phage or host
RNA
Sak proteins
Phage Exclusion
f Defense associated RTs (DRTs)
DRT1
DRT2
DRT3
DRT4
UG1
TM
UG2
UG3
UG8
UG15
RT-nitrilase
RT
RT-RNA Complex
RT
?
?
?
?
Signalling?

Figure I.3. Role of reverse Transcriptases in prokaryotic immunity. (a) Group II introns life cycle. Splicing step is initiatec
adenosine forming a lariat intermediate with the resulting binding of exons. In retrohoming step the introns can reverse splice into t
frequently a mobile genetic element preventing their mobility. (**b)** Retrons are tripartite systems in which RT produces a high-copy nu
which remains bound to the RT forming the antitoxin unit. The effector module, formed by one or more genes with different enzyr
constituted the toxin effector is inhibited by direct contact with the antitoxin. Through a mechanism that remains uncovered,
processed/degraded allowing the toxin cause cell death to protect cell population. (**c)** Diversity Generating Retroelements (DGRs) in
mutations in the template repeat (TR) in a reaction known as mutagenic reverse transcription carried out by the RT with the help
gene (typically Avd). Then, the mutated cDNA is integrated in the variable region (VR) of a target gene (TG), generating a great seque
TG generally presents protein-protein or surface displays activities enabling adaptability of host cell to different conditions such as
phages. **(d)** RTs associated with CRISPR-Cas system are part of the integrase complex together with Cas1 and Cas2 facilitating th
RNA molecules that are integrated in the CRISPR Array as a new spacer. (**e)** RTs involved in Abortive Infection (Abi) systems ar
AbiK and Abi-P2, however, their mechanism remains unknown. In AbiA, RT is fused to a HEPN domain and its thought to degrac
RNA to confer resistance. In AbiK, a random DNA remains attached covalently to a OH-group of a tyrosine residue in the RT do
fused to a unknown domain. Through an unravelling mechanism AbiK blocks phage Sak proteins conferring immunity. Abi-P2, wit
transcriptase activity, is formed by an RT and a domain of uncovered function able to perform phage exclusion in a way yet to be
Defense-associated RTs (DRTs) are novel antiviral systems where RT activity is necessary to confer resistance using an unknov
These systems are constituted by RTs from different UG groups which act alone (DRT2, DRT4, DRT5) together with small mer
(DRT1) or in case of DRT3 where RTs from two different UG groups with a ncRNA are necessary to have anti-phage properties.

The use of covariance models and consensus structure detection allows the identification of putative ncRNA consensus structures even in groups where there were no experimentally validated representatives. A comparison of the phylogenies of the three retron components suggests that not only retrons present high modularity, with the same type of RT associated with different domains or vice-versa, but also that they have co-evolved, evidencing that they could act as a functional unit. Moreover, due to the high diversity of putative enzymatic activities present in the genes or domains associated, 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 (Figure I.4; Mestre *et al.*, 2020).

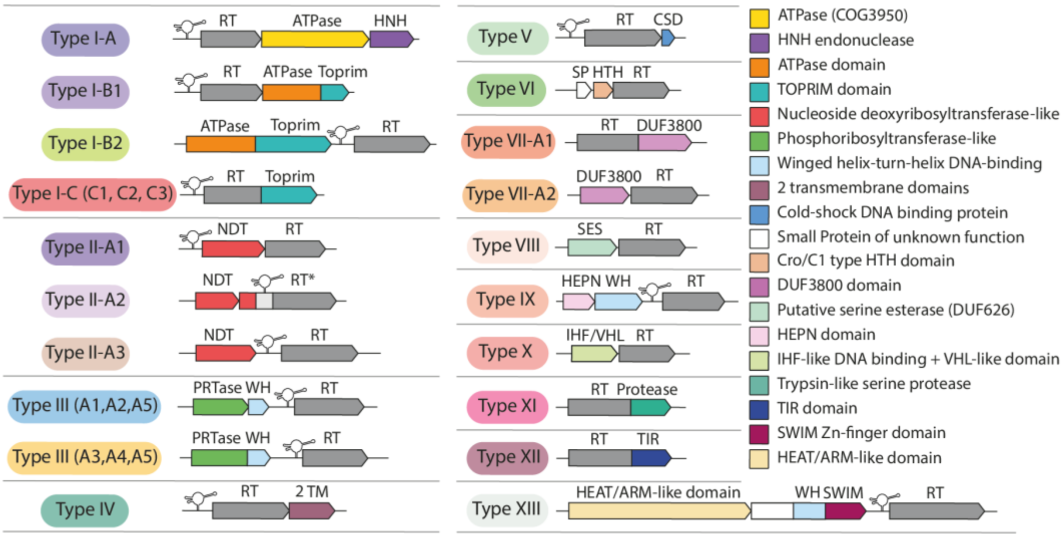


Figure I.4. Classification of retron systems. Schematic diagram of the genomic organization of the different types/variants of retron systems (Adapted from Mestre *et al.*, 2020)

The raised interest in prokaryotic defense mechanisms against phages has led to different strategies for looking for novel immunity systems. The most successful has been to look for clusters of antiviral systems in defense islands (Doron *et al.*, 2018). This approach has led to some retrons types being identified as abundant in these islands, suggesting a role of retrons in anti-phage defense (Gao *et al.*, 2020; Millman *et al.*, 2020). Both reports show that some retrons systems confer resistance against a wide range of phages, with different retron types protecting against different phages. Additionally, mutations in the three components of the system abolished immunity, indicating that all are required for correct activity (Figure I.3B). Isolating phages able to overcome resistance conferred by retron-Eco6 (Ec48), a type IV retron system according to the recent classification (Figure I.4), mutations in genes that inhibits the bacterial complex RecBCD were detected. This could suggest 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 2 transmembrane domain protein, that causes cell death (Millman *et al.*, 2020). However, Retron-Eco8 (type I-B2 retron system; Figure I.4) 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 (Figure I.4), has been shown to act as a novel type of tripartite toxin/antitoxin (TA) system, in which RT and msDNA form the antitoxin that directly blocks the toxin unit constituted by an ATPase-TOPRIM protein (RcaT), whose toxicity increase at low temperature or in anaerobiosis conditions (Bobonis *et al.*, 2020a). An RT-RcaT complex has been shown to be the active toxin, but the presence of the RT-msDNA complex binding the effector protein provides the antitoxin specificity (Figure I.3B). Phage-origin triggers and blockers of this novel TA system have also been identified, suggesting an extensive arms-race between retron system and phages (Bobonis *et al.*, 2020b). Some of the detected triggers (Dam and RecE) have anti-restriction properties and could lead to abortive infection mediated by RcaT by inactivating the RT-msDNA antitoxins suggesting crosstalk between innate/adaptive immunity systems and this tripartite TA system.

Despite the great progress made in the field of retrons thanks to these latest discoveries, multiple biological questions remain unanswered such as the way in which different types of retron systems sense phages, how the antitoxin is inactivated, or how the toxin performed the final step in the immunity. Moreover, there are new types of retrons with new ncRNA structures that have not yet been experimentally characterized (Mestre *et al.,* 2020)

I.1.3. 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 were 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.*, 207; Wu *et al.*, 2018; 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 comprise at least three essential components: a reverse transcriptase (RT), a template repeat (TR) and a target gene (TG) with a variable region (VR) displaying ≈90% sequence identity to the TR (Figure I.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. 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; 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 C-terminal 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 (Paul *et al.*, 2017; Yan *et al.*, 2019). Furthermore, bioinformatics tools have been developed to identify and characterize DGRs. These tools include DiGReF (Schillinger *et al.*, 2012), DGRscan (Ye *et al.*, 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 1,318 clusters (≥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 (Figure I.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). 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 (Figure I.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).

I.1.4. 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 stopping host metabolism or driving to cell death. Therefore, Abi systems avoid multiplication of the phage and protect the rest of the population (Bernheim and Sorek, 2020). A vast variety of Abi genes have been described, in fact, only in *Lactococcus* spp. are present about 20 different Abi systems. Another indicator of the heterogeneity of Abi systems is that three of them contain an RT domain: AbiA, AbiK and Abi-P2 (Fortier *et al.*, 2005; Odegrip *et al.*, 2006; Durmaz and Klaenhammer, 2007). However, the role of a putative reverse transcriptase activity in these systems remains still uncharacterized. Only in the case of AbiK, it has been demonstrated that the RT domain *per se* is enough for conferring resistance against the phage (Fortier, *et al.,* 2005; Wang *et al.*, 2011). For AbiA and Abi-P2, it remains unclear whether other genetic elements contained in the same loci could be required for the defense activity.

Each of these systems has a different mode of operation, but the N-terminal RT domain present in Abi proteins shares several features like conserving a potentially active site with a Y(R/V)DD sequence as well as lacking domains 0, 2a and 7 (Simon and Zimmerly, 2008; Toro and Nisa-Martínez, 2014). Besides, fused to this domain a C-terminal domain of variable length and unknown function is presented in all types of RT-based Abi systems (Figure I.2). Only the AbiA C-terminal domain has been proposed as a novel version of higher eukaryote and prokaryote nucleotide-binding (HEPN) domain (Anantharaman *et al.*, 2013). In multiple defense systems such as restriction-modification (R-M), toxin-antitoxin (TA), CRISPR-Cas, or in other Abi systems, HEPN domains represent a key component because of its RNAse activity that could directly attack viral RNAs or induce host suicide or dormancy attacking self-RNAs. The presence of RT and HEPN domains in AbiA could indicate that the phage multiplication is inhibited through interaction between phage-encoded proteins and a DNA molecule covalently linked to the RT and parallelly the HEPN domain degrades host RNA driving to cell suicide and protecting the surrounding community (Figure I.3E). Indeed, the best-characterized AbiA protein, found in a lactoccocal plasmid (Hill *et al.*, 1989), stops phage replication by targeting a viral recombinase (Dinsmore and Klaenhammer, 1997). Loci containing AbiA have been proved to confer immunity against a wide set of lactococcal phages (Hill *et al.*, 1990; Tangney and Fitzgerald, 2002). Additionally, lactococcal AbiA shows a versatile activity conferring resistance to a *Streptococcus thermophilus* strain against several phages (Tangney and Fitzgerald, 2002).

The best-known RT-based Abi system is AbiK, which have also been discovered in a native plasmid of a lactococcal strain (Emond *et al.*, 1997). As well as AbiA, this system provides resistance against a broad range of lactococcal phages (936, c2 and P335), reducing infectivity by six orders of magnitude (Fortier *et al.*, 2005). AbiK protein possesses an active polymerase activity which diverges from canonical RTs and acts analogously to a terminal transferase since polymerize DNAs of random sequence (Wang *et al.*, 2011). Moreover, the synthesized product remains covalently attached to the enzyme, possibly via a hydroxide group of a tyrosine located in the C-terminal domain which served as a primer during a reaction that would be similar to that observed in hepadnavirus self-priming (Wang and Seeger, 1992). On the other side, studies of phage mutants able to escape AbiK have identified the viral proteins targeted by 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, AbiK system works similarly to AbiA, preventing phage maturation by direct interaction with Sak proteins and provoking cell death by an unknown mechanism (Figure I.3E).

Unlike previously described systems, Abi-P2 was found in a highly variable region of several P2 prophages contained in different *E. coli* strains, with a higher AT content in comparison with genome host, suggesting that this region have an HGT origin (Odegrip *et al.*, 2006). Furthermore, the Abi-P2 encoding gene protein presents reverse transcriptase activity. Loci harboring the Abi-P2 protein has been shown to exclude phage T5 by reducing more than 10^7-fold the plating efficiency of this phage (Figure I.3E). Moreover, the deletion of the gene region containing the putative active site of Abi-P2 (YRDD) abolishes resistance to the phage (Odegrip *et al.*, 2006). A recent genome survey analysis enlarged the number of known RT-based Abi systems by an order of magnitude, with Abi-P2 type accounting for 75% of these systems (Toro *et al.*, 2019a).

I.1.5. Other groups of prokaryotic RTs

In addition to the former RT groups mentioned above, novel phylogenetic groups appear as a consequence of different types of domestication of probably ancient retroelements which function still remains to be elucidated but that could provide evolutionary advantages to the host. Over the years, novel uncharacterized RTs clustered in the so-called groups unknown groups (UG) have been appearing as new analyses increased the amount of RT sequences (Kojima and Kanehisa, 2008; Simon and Zimmerly, 2008; Toro and Nisa-Martínez, 2014) until a total of 28 different groups covering 10% of prokaryotic RT diversity (Toro *et al.*, 2019a). There is considerable sequence diversity in the UG-RTs. For example, UG1 and UG5 present a C-terminal nitrilase domain (Simon and Zimmerly, 2008), whereas in UG6, this domain is downstream in an adjacent ORF (Gao *et al.*, 2020). UG3 and UG8 represent a unique case in which RT sequences are always next to each other, suggesting that they act as a functional unit (Figure I.2; Kojima and Kanehisa, 2008).

UG-RTs have recently been found in defense islands and have been experimentally validated as new anti-phage systems called Defense-associated RTs (DRTs), in which immunity is dependent on RT activity (Gao *et al.*, 2020). In DRT type 1 (UG1), both the C-terminal nitrilase domain and a small membrane protein are required for defense (Figure I.3F). In DRT type 3, formed by the UG3 and UG8 reverse transcriptases, a structured ncRNA downstream of the UG8 gene is required for immunity. By contrast, DRT type 2 (UG2), type 4 (UG15) and type 5 (UG16) the RT alone confer resistance against phages. However, the mechanism of these novel systems remains uncovered and new experimental data are required to elucidate how these systems provide protection against different types of phages. However, not all UG groups have been tested, and only some of those investigated display antiviral properties (Gao *et al.*, 2020).

Another group of RTs were found closely related phylogenetically to Group II introns and were therefore named “group II-like” (G2L). 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). It is important to point out that RT groups classified as G2L were found to be associated with CRISPR-Cas systems (Kojima and Kanehisa, 2008, Simon and Zimmerly, 2008, Toro and Nisa-Martínez, 2014). In this way, G2L might 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 (Figure I.2).The present thesis will be focus on these particular association between RTs and CRISPR-Cas systems, a defense system that protects prokaryotic organisms from phages and other invading agents (see sections I.3 to I.6 for a detailed description of CRISPR-Cas systems).

I.2 Biotechnological application of Prokaryotic Reverse Transcriptases

RTs have been exploited as biotechnological tools in a wide range of fields, particularly those from eukaryotes and virus (Martín-Alonso *et al.*, 2020). The various functions of prokaryotic RTs, starting with group II introns, have been exploiting as genome editing over the last few decades, in the form of the first RNA-guided gene targeting tools (Table I.1; Guo *et al.*, 2000; Mohr *et al.*, 2000; Karberg *et al.*, 2001; Enyeart *et al.*, 2014). 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). Mobile group II introns were then used for a range of genome editing applications in bacteria known as targetron knockout technology (Gwee *et al.* 2019; Wen *et al.* 2020). The coupling of these techniques with CRISPR/Cas9 counterselection has recently increased the chances of finding clones that integrated the intron into the target *lacZ* sequence in a recombination-independent fashion (Velázquez *et al.*, 2019). Group II introns have also been used in genome editing to generate insertions, inversions, deletions, and one-step cut-and-paste operations in combination with Cre recombinase (Enyeart *et al.* 2013). The properties of 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 *et al.*, 2018), have also been exploited for many different high-throughput RNA characterization purposes.

The ability of retrons to produce a high copy number of msDNA has made them an interesting alternative and an useful addition to the genome-editing toolbox (Simon *et al.*, 2019). Moreover, the msd region can be modified with a random sequence without disturbing the production of msDNA, retrons could be engineered to synthesize a specific DNA molecule with different purposes. One of them is to produce antisense cDNAs to knockdown the mRNA of the target gene (Mao *et al.*, 1995). More recently, retrons has been used as genome editing tools, known as Synthetic Cellular Recorders Integrating Biological Events (SCRIBE), in which the expression of retron ncRNA, with the msd sequence modified to contain the desire sequence, is used to alter the target region (Table I.1; Farzadfard and Lu, 2014). The expression of retrons under an error-prone RNA polymerase generates random mutations in the msd region that are introduced in the target gene, enabling continuous *in vivo* evolution of the desire loci (Simon *et al.*, 2018). These applications could be used in combination with CRISPR-Cas systems enabling multiplex gene editing (Lim *et al.*, 2020). However, optimization of efficient and continuous genomic edition enables multiplexed applications using only the retron unit, in a technology termed Retron Library Recombineering (RLR) (Schubert *et al.*, 2020). In eukaryotes, retrons have been used to homology-directed repair of Cas9-targeted breaks, in a technology called Cas9 Retron precISe Parallel Editing via homologY (CRISPEY) (Sharon *et al.*, 2018). Additionally, the recent finding that retrons consist in novel tripartite TA systems with an effector protein associated with anti-phage activity could lead to new biotechnological applications such as in phage therapy.

The ability of DGRs to generate multiple variants of a protein domain has been used in synthetic biology for continuous target evolution useful in biotechnological applications such as in phage therapy. However, only a preliminary study has used the capacity of phage BPP-1 DGR to create variants of tropism proteins that likely bind to T4 lysozyme (Table I.1; Yuan *et al.,* 2013).

Table I.1: Biotechnological applications of different prokaryotic RT.

RT Type	Technology	Applications	General Description
Group II Introns	Targetron	Specific knockout	The specificity of the IBS–EBS interactions was harnessed to target introns to preprogrammed positions in a wide variety of bacterial genomes
	Targetron plus CRISPR-Cas9 counter selection	Specific knockout	CRISPR/Cas9 counterselection increased the chances of finding clones that integrated the intron into the target *lacZ* sequence in a recombination-independent fashion
	GETR	Genome editing	Group II introns deliver new *lox* sites allowing the recombinase Cre to produce insertions, inversions and deletions and one-step cut-and-paste operations
	TIGRT	RNA-seq and epitranscriptomics	Thermostable properties of RTs used in different high-throughput RNA characterization purposes
	Marathon RT	RNA-seq	Ultraprocessive and accurate properties of *E. rectale* Group II intron RT used in different high-throughput RNA characterization purposes
Retrons	Antisense cDNA gene regulation	Gene knockdown	Retron engineered to produce a msDNA which contains an antisense cDNA to knockdown a target gene
	SCRIBE	Genome editing	Retron engineered to produce a msDNA with a desire sequence to modify a target region after recombination
	Multiplex gene editing	Genome editing	Combination of retron and CRISPR-Cas9 to enable multiplex gene editing
	Continuous gene evolution	Continuous Genome editing	Expression of retron under an error-prone RNA polymerase that generates random mutations in the msd region which later is introduce in the desire loci.
	RLR	Continuous Genome editing	Optimization of retron-based genome editing t to increase efficiency applicable to multiplexed technologies.
	CRISPEY	Genome editing in Eukaryotes	Retron homology-direct reparation of CRISPR-Cas9 double-strand breaks in yeast
DGRs	Variants of tropism proteins	Phage Therapy	Continuous target evolution trough mutagenic retrohoming reaction carried out by Bordetella-phage1 DGR to create variants of tropism proteins that bind T4 lysozyme
RT-CRISPR-Cas	Record-seq	Record transcriptional events	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

The exploration of novel DGRs with advantageous properties such as thermostability (Handa *et al.,* 2019) could provide promising systems to work with. In addition, characterization of the mutagenic retrohoming mechanism has revealed that RT and Avd protein in BPP-1 DGR work as a complex and both are necessary to synthesize cDNA from both DGR and non-DGR templates, thanks to the addition of an oligodeoxynucleotide (ODN) primer (Handa *et al.*, 2018). cDNAs synthesized from non-DGRs templates also presented adenine mutations showing that this fact is an intrinsic feature of the RT-Avd complex. This ability can be used to create libraries of hypermutated cDNA to address sequence variability searching for protein variants that significantly improve native protein activity. Theoretically, mutagenic retrohoming is the biological process that creates more sequence variability, potentially about 10^{30} protein variants, several orders of magnitude above eukaryotic systems (Wu *et al.*, 2018). Altogether, these observations show that DGRs can be a powerful tool for protein engineering.

I.3. Discovery and biological role of CRISPR-Cas Systems

At the beginning of 90's last century, in the Santa Pola salt marshes (Alicante, Spain) a series of intriguing repetitive sequences were found in the genome of the archaea *Haloferax mediterranei* (Mojica *et al.*, 1993). Afterwards, similar repetition patterns were found in more distant archaea and also in eubacteria (Mojica *et al.*, 1995). These particular repeats, termed CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats), were frequently associated with specific genes, named *cas* (CRISPR-associated) (Jansen *et al.*, 2002). Despite the many proposed hypothesis, the role of CRISPR repeats remained unsolved for more than a decade until different research groups realized that some of the sequences present between the DNA repeats were identical to sequences of bacteriophages and plasmids, suggesting the idea that CRISPR-Cas systems might be a prokaryotic defense mechanism (Mojica *et al.*, 2005; Bolotin *et al.*, 2005; Pourcel *et al.*, 2005). This hypothesis was confirmed by Barrangou *et al.*, 2007, by demonstrating that a complete CRISPR–Cas system found in *Streptococcus thermophilus* provides adaptive immunity against foreign mobile genetic elements. In addition, these systems have a high ecological impact as they are present in almost all archaea and about 50% of all bacterial genomes (Grissa *et al.*, 2007a; Makarova *et al.*, 2015). A typical CRISPR-*cas* locus is defined by the presence of the following components: the CRISPR Array, the leader sequence, and a variable number of flanked *cas* genes (Figure I.5A).

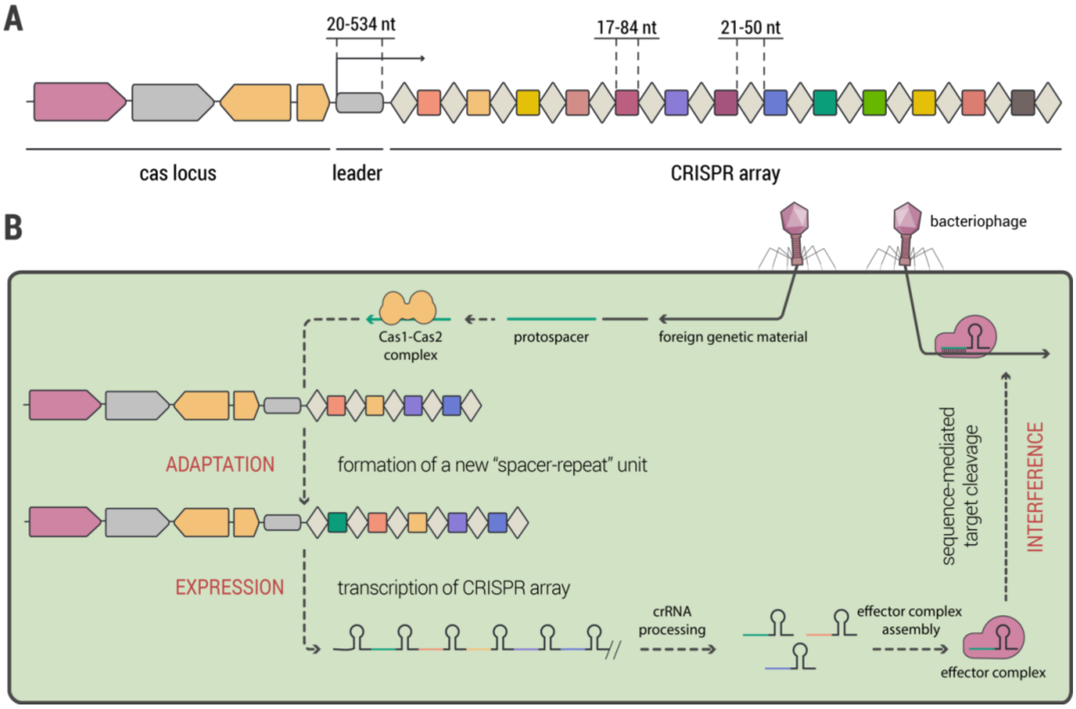


Figure I.5 Architecture and mechanism of CRISPR-Cas systems. (A) General CRISPR-*cas* locus. The different lengths of the leader sequence, direct repeats and spacers within the CRISPR Array are indicated above the partial genome. Arrows show *cas* genes flanking the CRISPR locus. Grey diamonds indicate the direct repeats and coloured squares highlight the different spacers found in the CRISPR Array. **(B)** CRISPR-Cas mediated immunity. Upon infection, new foreign DNA sequences are captured and integrated into the host CRISPR Array as new spacers during the adaptation stage. In the expression stage, the CRISPR locus is transcribed and processed to generate mature CRISPR RNAs (crRNAs), each encoding a unique spacer sequence which serves as guide in the last stage, the interference where each crRNA associates with Cas effector proteins to target foreign genetic elements that are complementary to the crRNA sequence.

The CRISPR Array is formed by repeat sequences, termed Direct Repeats (DRs) alternating with short DNA sequences originated from the invading agent, known as spacers. The number of DRs and spacers of a CRISPR locus is broadly variable (Kunin *et al.*, 2007). Upstream of the CRISPR Array is located the leader sequence, the regulatory promoter region with a high AT content, responsible for the transcription of the CRISPR Array (Pougach *et al.*, 2010). Finally, a set of extremely diverse Cas proteins provides the enzymatic machinery required for effective immunity (Makarova *et al.*, 2015; Makarova *et al.*, 2020).

The adaptive defense mediated by CRISPR-Cas systems takes place in three steps: adaptation, expression, and interference (Figure I.5B) (Deveau *et al.*, 2010; van der Oost *et al.,* 2014; Amitai and Sorek, 2016). During the adaptation stage, the integrase complex formed by Cas1 and Cas2 selects, processes and integrates short DNA sequences from the foreign nucleic acids into the CRISPR Array as a new spacer, providing a memory of infection (Yosef *et al.*, 2012; Nuñez *et al.*, 2014). Then, in the expression stage, the CRISPR Array is transcribed, usually by Cas6, to produce a long precursor RNA that is further processed within the repeat sequences to generate small RNA units knows as CRISPR RNA (crRNA) (Haurwitz *et al.*, 2010). Every crRNA contains a single spacer flanked by the processed DRs and this whole unit serves as a guide in the final stage, the interference. A complex formed by the crRNA and the effector Cas proteins acts as a cell guard scanning for target invading elements that are recognized by base-pairing with complementary crRNA sequences. Once the recognition is successful, the effector module cleaves the nucleic acid of the target, completing the immunity process (Brouns *et al.*, 2008).

I.4. Classification of CRISPR-Cas systems

According to the nature of the interference complex CRISPR-Cas systems have been assigned to 2 classes, which are further subdivided into 6 types and 33 subtypes that each possesses signature *cas* genes (Makarova *et al.*, 2020). In Class 1 CRISPR-Cas systems, the interference is carried out by a protein complex composed of multiple Cas proteins, whereas in Class 2 systems, a single effector protein with multiple domains is responsible for accomplishing interference (Figure I.6). All *cas* genes are involved in a total of four different functional modules: adaptation, expression, interference and signal transduction/ancillary (Makarova *et al.*, 2020).

The almost universal adaptation module is mainly comprised by *cas1* and *cas2*, which form a heterohexamer integrase complex responsible for the acquisition of novel spacers. Other proteins are implicated in adaptation in the different types and subtypes, including the Cas4 nuclease in some type I and V subtypes (Lee *et al.*, 2018; Kieper *et al.*, 2018), Csn2 in subtype II-A (Heler et al., 2015), and RTs in some types III CRISPR-Cas systems (Silas *et al.,* 2016). Despite the wide distribution of the adaptation module, genomic analysis has revealed many recombination events of this unit, which also shows no phylogenetic correlation with the genes of the interference module (Garrett *et al.*, 2011; Silas *et al.*, 2017a). Thus, the current classification of CRISPR-Cas system is based on the signature genes of the effector module (Makarova *et al.*, 2015). This classification employs a multipronged computational strategy that includes the identification of signature genes for CRISPR-Cas types and subtypes, comparison of gene repertoires, and genomic organizations and sequence similarity, and phylogenetic analysis of conserved genes.

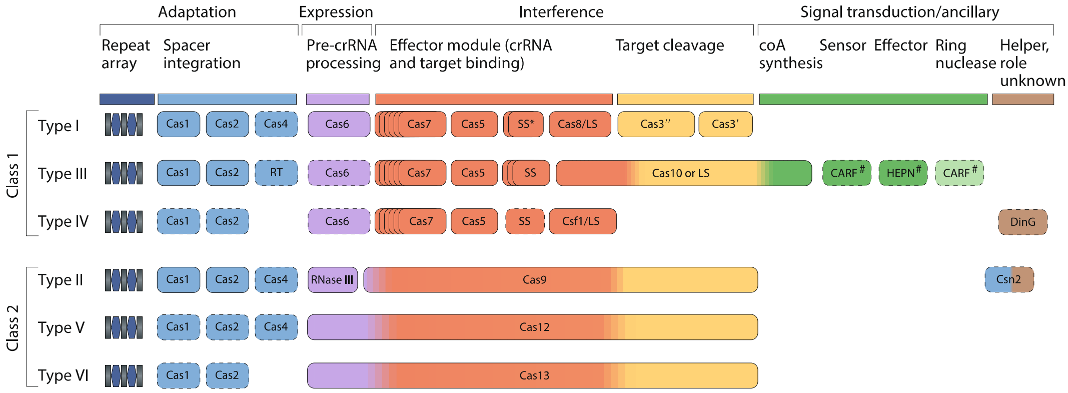


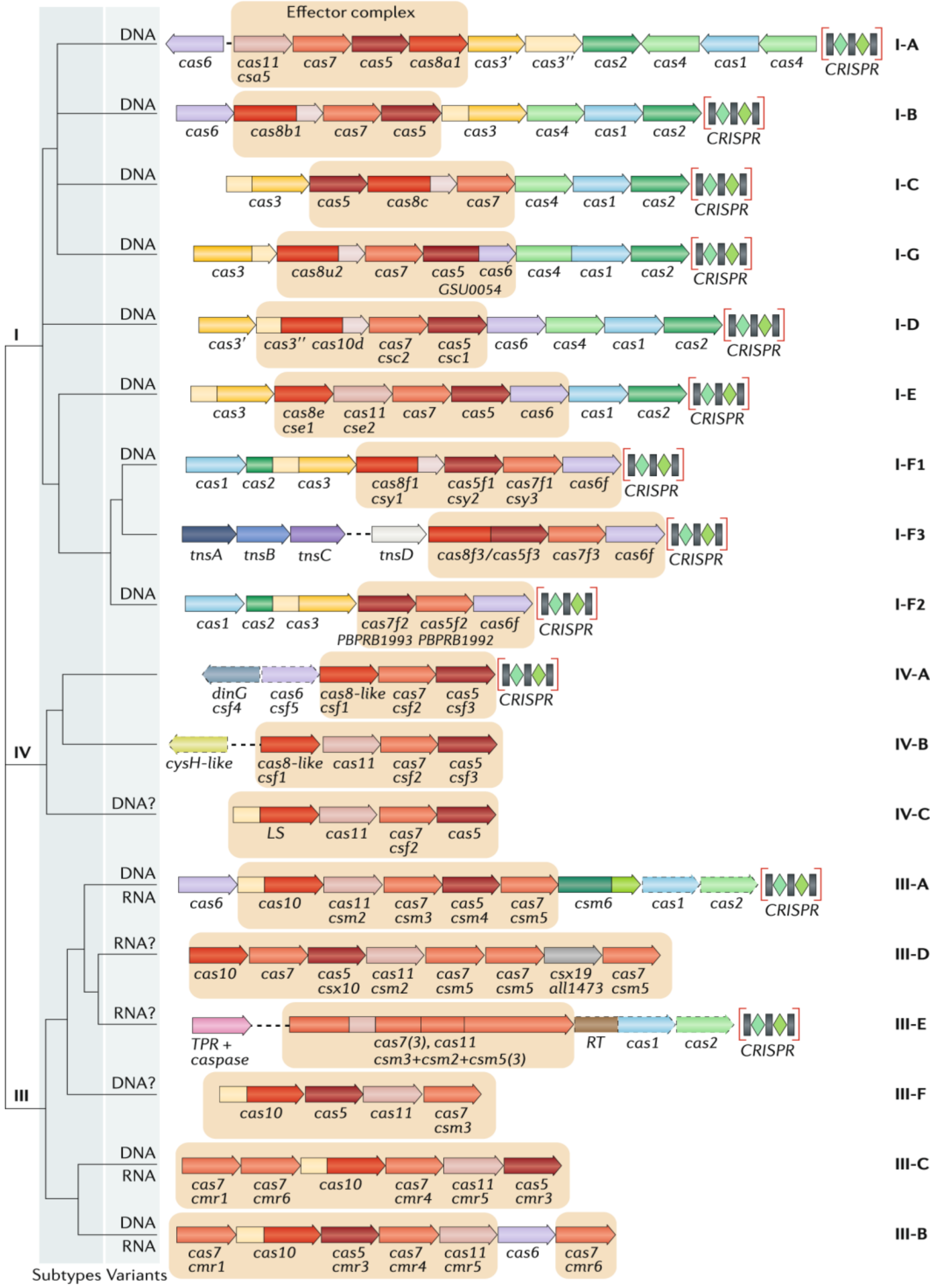
Figure I.6 Functional modules of CRISPR-Cas systems. Relationships between the genetic, structural, and functional organization of CRISPR-Cas systems. An asterisk indicates the putative small subunit that might be fused to the large subunit in several type I subtypes. The pound symbols indicate proteins families that could be involved in signal transduction. The CRISPR-associated Rossmann fold (CARF) and higher eukaryotes and prokaryotes nucleotide-binding (HEPN) are the common sensors and effectors, respectively. Dispensable (or missing) components are indicated by dashed lines. The three colours for Cas9, Cas10, Cas12 and Cas13 reflect that contribute to different stages of CRISPR-Cas immunity.LS, large subunit; SS, small subunit (Adapted from Makarova *et al.,* 2020).

I.4.1. Class 1 CRISPR-Cas Systems

Class 1 CRISPR-Cas systems are characterized by the presence of a multi-subunit crRNA-effector complex. This class includes types I, III and IV CRISPR-Cas system which are in turn subdivided into 16 subtypes (Figure I.7; Makarova *et al.*, 2020). In all class 1 types the backbone of the effector complexes is formed by the domain-containing proteins of the repeat-associated mysterious proteins (RAMPs) and Cas6, the RNAse in charge of processing the crRNA. The following is a summary of the main features that define the different types of the Class 1 CRISPR-Cas systems.

I.4.1.1 Type I CRISPR-Cas systems

Type I CRISPR-Cas systems are the broadest spread of all CRISPR-Cas systems and present a conserved architecture. However, its composition is variable among the different subtypes with effector subunits sharing functional and structural homology rather than sequence identity (Koonin *et al.*, 2017; Makarova *et al.*, 2015). The signature gene of this type is *cas3*, which encodes a single stranded DNA (ssDNA) helicase with capacity to unwind double-stranded DNA (dsDNA) and DNA-RNA hybrids (Sinkunas *et al.*, 2011; Gong *et al.*, 2014). Type I systems are currently divided into seven subtypes, I-A to I-G, defined by a combination of signature genes and operon organization derived from the ancestral type I gene arrangement (*cas1-cas2-cas3-cas4-cas5-cas6-cas7-cas8*) (Figure I.7; Makarova *et al.*, 2020). Furthermore, genomic surveys had let to the discover of several defective variants, type I-B and I-F encoded by Tn7-like transposons that lacks the helicase-nuclease Cas3, required for the interference stage. This "minimal" variants could perform other functions apart from immunity, such as the recently validated guide-RNA-mediated transposition (Klompe *et al.*, 2019). These defective variants are also encoded by phages, in which their function remains to be deciphered (Al-Shayeb *et al.*, 2020).


Effector complex
DNA
cas6 cas11 csa5 cas7 cas5 cas8a1 cas3' cas3'' cas2 cas4 cas1 cas4 CRISPR
I-A
DNA
cas6 cas8b1 cas7 cas5 cas3 cas4 cas1 cas2 CRISPR
I-B
DNA
cas3 cas5 cas8c cas7 cas4 cas1 cas2 CRISPR
I-C
DNA
cas3 cas8u2 cas7 cas5 cas6 GSU0054 cas4 cas1 cas2 CRISPR
I-G
I
DNA
cas3' cas3'' cas10d cas7 csc2 cas5 csc1 cas6 cas4 cas1 cas2 CRISPR
I-D
DNA
cas3 cas8e cse1 cas11 cse2 cas7 cas5 cas6 cas1 cas2 CRISPR
I-E
DNA
cas1 cas2 cas3 cas8f1 csy1 cas5f1 csy2 cas7f1 csy3 cas6f CRISPR
I-F1
tnsA tnsB tnsC tnsD cas8f3/cas5f3 cas7f3 cas6f CRISPR
I-F3
DNA
cas1 cas2 cas3 cas7f2 PBPRB1993 cas5f2 PBPRB1992 cas6f CRISPR
I-F2
dinG csf4 cas6 csf5 cas8-like csf1 cas7 csf2 cas5 csf3 CRISPR
IV-A
IV
cysH-like cas8-like csf1 cas11 cas7 csf2 cas5 csf3
IV-B
DNA?
LS cas11 cas7 csf2 cas5
IV-C
DNA
RNA
cas6 cas10 cas11 csm2 cas7 csm3 cas5 csm4 cas7 csm5 csm6 cas1 cas2 CRISPR
III-A
RNA?
cas10 cas7 cas5 csx10 cas11 csm2 cas7 csm5 cas7 csm5 csx19 all1473 cas7 csm5
III-D
RNA?
TPR + caspase cas7(3), cas11 csm3+csm2+csm5(3) RT cas1 cas2 CRISPR
III-E
III
DNA?
cas10 cas5 cas11 cas7 csm3
III-F
DNA
RNA
cas7 cmr1 cas7 cmr6 cas10 cas7 cmr4 cas11 cmr5 cas5 cmr3
III-C
DNA
RNA
cas7 cmr1 cas10 cas5 cmr3 cas7 cmr4 cas11 cmr5 cas6 cas7 cmr6
III-B
Subtypes Variants

(Figure Legend in the next page)

Figure I.7 Classification of Class 1 CRISPR-Cas systems. The scheme represents typical CRISPR-*cas* loci of each class 1 subtype, with the dendrogram on the left showing the likely evolutionary relationships between types and subtypes. Homologous genes are color-coded and identified by a family name. The legacy name is used under the systematic name. Dispensable genes are indicated with dashed lines. Gene regions coloured cream represent the HD nuclease domain; the HD domain in Cas10 is distinct from that in Cas3. The tan shading shows the effector module. Uncharacterized genes are shown in grey. CHAT, protease domains of the caspase family; RT, reverse transcriptase; TPR, tetratricopeptide repeat (Adapted from Makarova *et al.*, 2020)

The process of crRNA maturation in type I systems shows great similarity with type III systems, as both uses Cas6 family proteins to process the CRISPR Array transcript (Carte *et al.*, 2008; Haurwitz *et al.*, 2010; Sashital *et al.*, 2011). Note that the I-C subtype represents an exception which lacks a Cas6 homolog, being functionally replaced by Cas5d (Garside *et al.*, 2012; Nam *et al.*, 2012). Both enzymes cleave the pre-crRNA within the repeat regions, yielding matures crRNAs that include the spacer flanked by the repeat-derived 5' handle and the 3' stem loop. Thereafter, Cas6 remains bound to the crRNA in most type I CRISPR-Cas systems, acting as scaffold for the formation of the effector complex (Jore *et al.*, 2011). However, in subtype I-A and I-B, which present non-palindromic repeats, Cas6 does not bind the crRNA after cleavage (Charpentier *et al.*, 2015). Furthermore, the Cas6 of these subtypes acts as a dimer, which seems to be the responsible for the reshaped Cas6 activity (Reeks *et al.*, 2013; Shao and Li, 2013).

The type I effector complex is known as CRISPR-associated complex for antiviral defense (Cascade). To understand how the type I interference works, Well-characterized type I-E system from *Escherichia coli* can serve as a model to explain it. The type I-E Cascade complex displays as follow: $(Cas5e)_1$-$(Cas6e)_1$-$(Cas7e)_6$-$(Cas8e)_1$-$(Cas11e)_2$ (Brouns *et al.*, 2008). After crRNA maturation, Cas6 remains bound to the 3' region of the RNA guide and serves as scaffolding for the Cascade assembly into a seahorse-like structure (Gesner *et al.*, 2011; Sashital *et al.*, 2011). The crRNA is bound along the helical backbone of the complex, formed by the six Cas7 subunits, and capped by Cas5E at the 5' end. The structure of Cas7 subunits enables and efficient base pairing between the crRNA and the target DNA. Cas11e, the small subunit, and Cas8e, the large subunit, form the belly and the tail of the complex, respectively (Jackson *et al.*, 2014; Mulepati *et al.*, 2014).

Upon Cascade formation, Cas8e recognizes the Protospacer-Adjacent Motif (PAM) of the target DNA, a short sequence which is specifically recognized by both adaptive and interference complex preventing autoimmunity in most CRISPR-Cas systems. This recognition allows Cas8e to initiate the unwilling of DNA and the subsequent proof for complementarity with the crRNA. Cas11e binds the non-target strand, playing a crucial role in the formation and stabilization of the so-called R-loop structure, that together with major conformational changes in Cascade subunits allow the recruitment of the Cas3 nuclease/helicase for target degradation (Jore *et al.*, 2011; Jackson *et al.*, 2014; Mulepati *et al.*, 2014; Hayes *et al.*, 2016; Xiao *et al.*, 2017). Cas3 nicks the non-target DNA strand and further unwinds the DNA leading to a successive cleavage of non-target strand (Westra *et al.*, 2012; Redding *et al.*, 2015). The Cascade/Cas3 complex continues translocating along the DNA until finding DNA proteins that block their advance allowing Cas3 perform a double-strand break, which results in the whole DNA degradation (Dillard *et al.*, 2018).

I.4.1.2. Type III CRISPR-Cas systems

In type III systems *cas10* is the hallmark gene, which encodes a multidomain protein that contains a Palm domain, homologous to nucleic acid polymerases and cyclases, often fused to a nuclease domain and which represents the largest subunit of type III effector complex. Currently, type III systems are subdivided into six subtypes, III-A to III-F (Figure I.7; Makarova *et al.*, 2020), and are defined by their ability to target both DNA and RNA substrates. Moreover, DNA degradation strictly depends on transcription of the target sequence (Deng *et al.,* 2013; Elmore *et al.*, 2016; Estrella *et al.*, 2016; Kazlauskiene *et al.*, 2016). The composition of type III CRISPR-*cas* loci is more varied in comparison with type I systems due to domain insertion and gene duplications and deletions, most of them being poorly characterized. Furthermore, type III-CRISPR-Cas systems present the highest diversity of ancillary proteins, among which it is worth noting the RNAses of Csm6 and Csx1 family (Koonin *et al.*, 2017; Shmakov *et al.*, 2018; Shah *et al.*, 2019). Subtypes III-A and III-B mainly differ in the small subunit, while subtype III-C presents an inactivate Cas10 cyclase-like domain and subtype III-D usually lacks the Cas10 nuclease domain (Makarova *et al.*, 2015). The subtype III-E, recently discovered, is characterized by a large multidomain protein fusion comprising the crRNA-binding part of the interference complex, probably evolved from subtype III-D systems, whereas subtype III-F contains divergent features from the rest of subtypes suggesting a different interference mechanism (Makarova *et al.*, 2020).

Cas6 homologs encoded by type III systems share structural and functional features with those from subtypes I-A and I-B CRISPR-Cas systems such as not being part of the effector complex. Here, Cas6 helps to remodel the pre-crRNA due to the direct repeats are unable to form stable stem loop. After the first Cas6 cut, the crRNA undergoes 3’ end trimming, removing the stem-loop (Carte *et al.*, 2008; Carte *et al.,* 2010; Hatoum-Aslan *et al.*, 2011). In the systems lacking Cas6 homologs, as usually happens in subtypes III-C and III-D, Cas5 is assumed to be the responsible for crRNA maturation.

Type III CRISPR-Cas interference complexes are known as Csm and Cmr in subtypes III-A and III-B, respectively. Similar to Cascade, these effector modules assemble along the mature crRNA which is bound by the type III Cas5 homologs (Csm4/Crm3) with Cas7-family proteins (Csm3 and Csm5/Cmr1, Cmr4 and Cmr6) forming the complex backbone. Furthermore, Cas11 homologs (Csm2/Cmr5) represent the small subunit and Cas10 is the largest one (Staals *et al.*, 2014; Taylor *et al.*, 2015). The interference is initiated by crRNA-dependent complex recognition of the nascent target transcript (Figure I.8). Thereafter, Cas7 subunits degrade the transcript cutting every six nucleotides (Tamulaitis *et al.*, 2014; Taylor *et al.*, 2015), while Cas10 carry out a dual function. On one hand, Cas10 nuclease domain cleaves the DNA after target recognition (Samai *et al.*, 2015). In addittion, Cas10 Palm domain converts ATP to cyclic oligoadenylates which act as a second messengers binding the pocket of a CARF domain required for the activation of the unspecific RNAses Csm6 (subtype III-

A) and Csx1 (subtype III-B). These enzymes are responsible for the degradation of cell RNA providing an optimized defense mechanism (Kazlauskiene *et al.*, 2017; Niewoehner *et al.*, 2017).

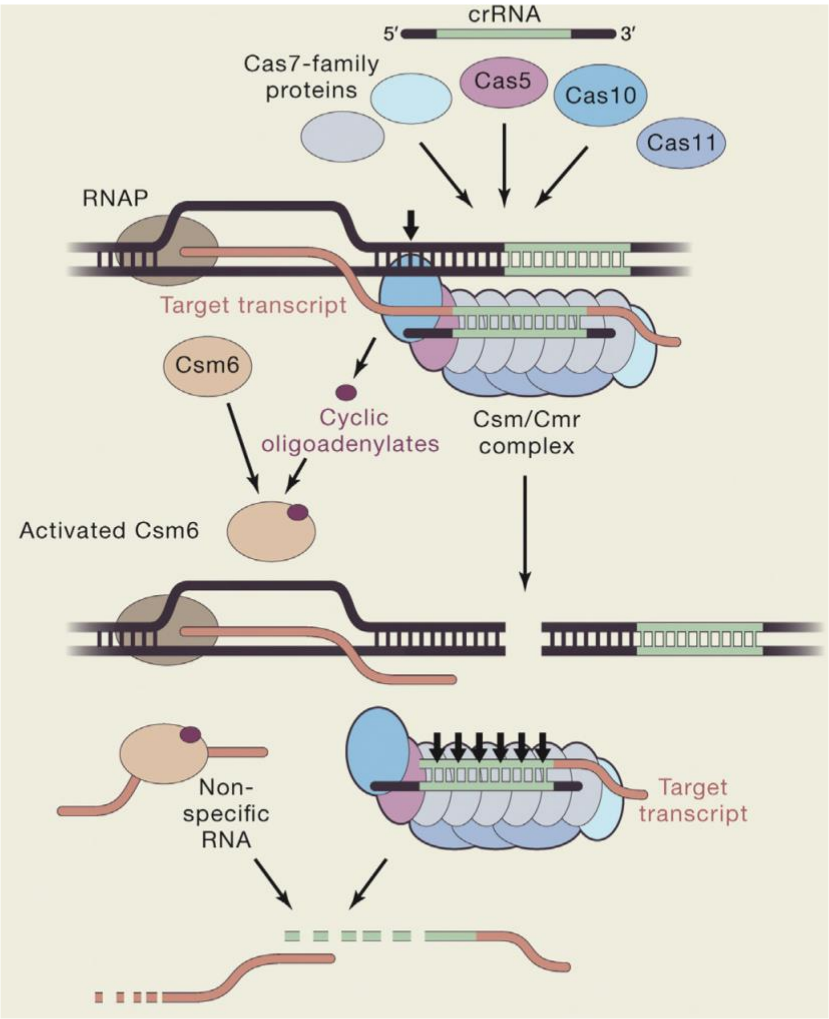


Figure I.8 Interference in type III-CRISPR-Cas systems. Type III CRISPR-Cas systems form interference complexes (Csm for type III-A and III-D, and Cmr for type III-B and III-C) using the crRNA as scaffold. Type III-A is shown here as an example. The crRNA generetad by Cas6 binds to complementary regions in target RNA transcripts. Binding triggers a Cas10-mediated double-strand break within the target DNA, after wich Cas7 cleaves the transcript RNA. Upon target binding, Cas10 produces cyclic oligoadenylates, which enable the non-specific RNAse activity of Csm6 (Adapted from Hille *et al.*, 2018).

The production of the second messenger implicates a regulation level of type III CRISPR-Cas systems due to cyclic oligoadenylates are only generated upon complex recognition of the target. Moreover, other ancillary proteins encoded by type III systems, known as Ring nucleases, catalyze the degradation of the cyclic oligoadenylates switching off the non-specific ribonuclease activity of Csm6 and Csx1, adding a novel layer of regulation to type III systems (Athukoralage *et al.*, 2018). It is important to highlight that self- versus no-self discrimination in type III CRISPR-Cas systems takes place in a PAM-independent way, consisting in the inhibition of target DNA degradation once it is bound to the 5' repeat region of the crRNA (Marraffini and Sontheimer, 2010).

I.4.1.3. Type IV CRISPR-Cas systems

Type IV CRISPR-Cas systems are poorly characterized. Historically, Csf1 protein has been used as signature for type IV CRISPR-Cas systems (Makarova *et al.*, 2015), a recent work uses the most conserved Cas7-like (Csf2) protein in order to finding novel type IV systems (Pinilla-Redondo *et al.,* 2020). Type IV classification reveals three subtypes, IV-A to IV-C (Figure I.7), characterized by encoding a minimal effector subunit which consist of Csf1 (the large subunit), Cas5, Cas7, and a putative small subunit in some cases (Makarova *et al.*, 2011). Subtype IV-A contains a DinG family helicase, whereas most subtypes IV-B loci present the small subunit and *cysH* as an ancillary gene. Subtype IV-C encodes a Cas10-like protein with a nuclease domain required for DNA cleavage but lacking the cyclase Palm domain, which suggest that this variant is not involved in oligoadenylate signaling as the Cas10 of type III systems (Makarova *et al.*, 2020; Pinilla-Redondo *et al.*, 2020).

In type IV systems, crRNA maturation is carried out similarly to other Class 1 systems by Csf5, a Cas6-like protein, which remains bound to the crRNA, laying the base for the formation of the effector complex (Özcan *et al.*, 2019). Despite their minimal configuration, type IV interference complex is similar to Cascade, with Csf2 forming the backbone (Özcan *et al.*, 2019). However, the complete mechanism of type IV CRISPR-Cas immunity remains mostly uncovered. Most type IV CRISPR-Cas systems are encoded by plasmids and, in contrast to others CRISPR-Cas systems, analysis of the spacers have revealed that type IV systems show a targeting bias towards plasmid-like elements (Pinilla-Redondo *et al.*, 2020). Indeed, one example of a type IV-A1 system has been shown to mediate RNA-guided interference against plasmids (Crowley *et al.*, 2019). These data suggest that type IV CRISPR-Cas systems present a specific anti-plasmid activity.

I.4.2 Class 2 CRISPR-Cas Systems

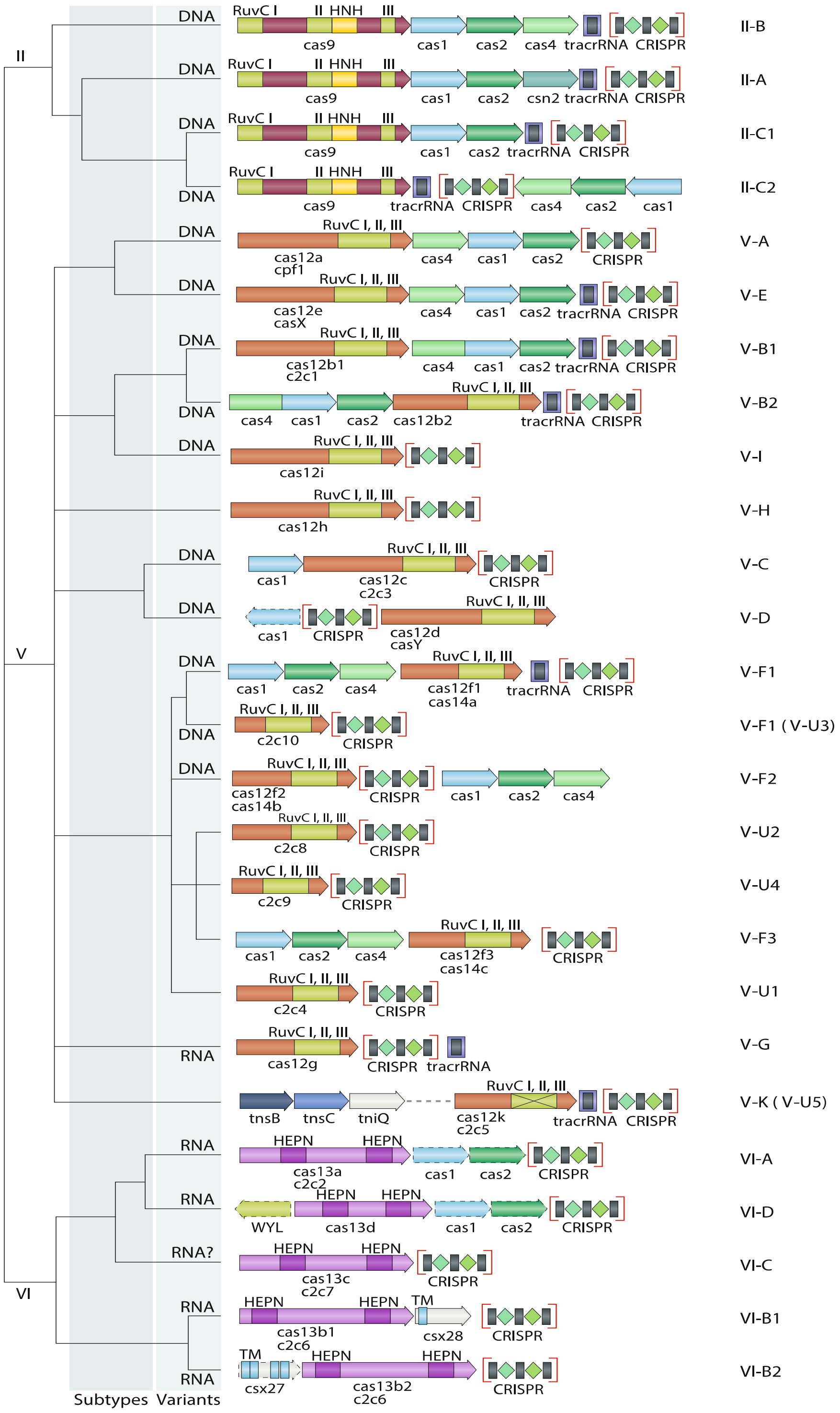
Class 2 CRISPR-Cas systems are defined by the presence of a single subunit crRNA-effector module. This class encompasses type II, V, and VI CRISPR-Cas systems with a total of 17 subtypes (Figure I.9; Makarova *et al.*, 2020). Another important difference respect to Class 1 systems is the mechanism of processing the crRNA, since they lack Cas6 homologs. In type II and several subtype V, the host RNAse III is the enzyme responsible for crRNA biogenesis, whereas in types VI and V-A this activity is also encoded by the large effector protein of CRISPR-Cas systems (Fonfara *et al.*, 2016; East-Seletsky *et al.*, 2016). The relevance of the studies of type II, V, and VI systems lies in their use as potential genome-editing tools (Hsu *et al.*, 2014; Cox *et al.*, 2017). The main features of each Class 2 CRISPR-Cas systems will be briefly highlighted below.

I.4.2.1 Type II CRISPR-Cas Systems

The hallmark gene in type II CRISPR-Cas system is *cas9*, that encodes a dual RNA-guided DNA endonuclease (Jinek *et al.*, 2012) and contributes to the adaptation stage as well (Heler *et al.*, 2015). All type II CRISPR-*cas* loci contain the adaptive machinery and a trans-activating crRNA (tracrRNA) partially complementary to the repeat region of the crRNA, which is required for Cas9 activity (Deltcheva *et al.*, 2011). Type II systems are currently classified into three subtypes, II-A to II-C, that differs in the Cas9 size and the presence of ancillary genes (Fonfara *et al.*, 2014; Makarova *et al.*, 2015). Subtype II-A is characterized by harboring the *csn2* gene, which is involved in spacer acquisition (Heler *et al.*, 2015). In subtypes II-B *csn2* is replaced by *cas4*, a gene typically found in type I CRISPR-Cas systems that ensures the acquisition of functional spacers (Lee *et al.*, 2018; Kieper *et al.*, 2018). Subtype II-C, the most widespread type II system, only contains *cas1, cas2* and *cas9* genes (Makarova *et al.*, 2015).

Type II crRNA maturation requires the binding of the repeat region of the CRISPR Array transcript to the tracrRNA. Cas9 stabilizes the tracrRNA:crRNA duplex leading to the recruitment of the host RNAse III, responsible for trimming the substrate within the double-stranded repeat sequence. A second cleavage takes place in the 5' end of the crRNA, removing the 5' repeat-derived tag. At this point, the tracrRNA:crRNA:Cas9 complex is ready for the interference stage (Deltcheva *et al.*, 2011). However, RNAse III is not a mandatory requirement since the repeat of type II-C CRISPR-Cas systems in *Neisseria* and *Campylobacter* species contain promoter sequences, leading to the transcription of individual crRNAs, which forms functional complexes along with Cas9 and the tracrRNA (Dugar *et al.*, 2013; Zhang *et al.*, 2013).

Upon formation of the effector complex, Cas9 undergoes structural changes to accommodate the dual RNA guide and, subsequently, the complex scans the DNA searching for target sequences through PAM recognition. With this aim, Cas9 unwinds the DNA of the non-target strand testing base pairing between the crRNA and the target DNA. As long as enough complementarity exists between the two sequences, the resulting stable R-loop structure allows Cas9 to produce a blunt, double-strand break just three nucleotides upstream of the PAM sequence (Garneau *et al.*, 2010; Jinek *et al.*, 2012). Cas9 protein presents a bilobed structure comprising a REC lobe (recognition) and a NUC lobe (target degradation), linked by an arginine-rich region, which acts as a bridge between the two lobes. The NUC lobe presents two different nuclease domains, the HNH domain responsible for the cut in the target strand, and the RuvC domain which breaks the non-target strand. This process leads to the double-strand break in the target DNA (Anders *et al.*, 2014; Jinek *et al.*, 2014; Nishimasu *et al.*, 2014).


II
V
VI
Subtypes
Variants
DNA RuvC I II HNH III cas9 cas1 cas2 cas4 tracrRNA CRISPR II-B
DNA RuvC I II HNH III cas9 cas1 cas2 csn2 tracrRNA CRISPR II-A
DNA RuvC I II HNH III cas9 cas1 cas2 tracrRNA CRISPR II-C1
DNA RuvC I II HNH III cas9 tracrRNA CRISPR cas4 cas2 cas1 II-C2
DNA RuvC I, II, III cas12a cpf1 cas4 cas1 cas2 CRISPR V-A
DNA RuvC I, II, III cas12e casX cas4 cas1 cas2 tracrRNA CRISPR V-E
DNA RuvC I, II, III cas12b1 c2c1 cas4 cas1 cas2 tracrRNA CRISPR V-B1
DNA cas4 cas1 cas2 cas12b2 RuvC I, II, III tracrRNA CRISPR V-B2
DNA RuvC I, II, III cas12i V-I
RuvC I, II, III cas12h V-H
DNA cas1 RuvC I, II, III cas12c c2c3 CRISPR V-C
DNA cas1 CRISPR RuvC I, II, III cas12d casY V-D
DNA cas1 cas2 cas4 RuvC I, II, III cas12f1 cas14a tracrRNA CRISPR V-F1
DNA RuvC I, II, III c2c10 CRISPR V-F1 (V-U3)
DNA RuvC I, II, III cas12f2 cas14b CRISPR cas1 cas2 cas4 V-F2
RuvC I, II, III c2c8 CRISPR V-U2
RuvC I, II, III c2c9 CRISPR V-U4
cas1 cas2 cas4 RuvC I, II, III cas12f3 cas14c CRISPR V-F3
RuvC I, II, III c2c4 CRISPR V-U1
RNA RuvC I, II, III cas12g CRISPR tracrRNA V-G
tnsB tnsC tniQ RuvC I, II, III cas12k c2c5 tracrRNA CRISPR V-K (V-U5)
RNA HEPN HEPN cas13a c2c2 cas1 cas2 CRISPR VI-A
RNA WYL HEPN HEPN cas13d cas1 cas2 CRISPR VI-D
RNA? HEPN HEPN cas13c c2c7 CRISPR VI-C
RNA HEPN HEPN TM cas13b1 c2c6 csx28 CRISPR VI-B1
RNA TM csx27 HEPN HEPN cas13b2 c2c6 CRISPR VI-B2

(Figure Legend in the next page)

Figure I.9 Classification of class 2 CRISPR-Cas systems. The scheme represents typical CRISPR-*cas* loci for each class 2 subtype, with the dendrogram showing the likely evolutionary relationships between types. Homologous genes are color-coded and are identified by the family name. The legacy name is given under the systematic name. Dispensable genes are shown by dashed lines. Additional genes encoding components of the interference module, such as tracrRNA are shown. The domains of the effector protein are colour-coded: RuvC-like nuclease, green; HNH nuclease, yellow; higher eukaryotes and prokaryotes nucleotide-binding (HEPN) RNAse, purple; transmembrane domains, blue (Adapted from Makarova *et al.*, 2020).

I.4.2.2 Type V CRISPR-Cas Systems

The signature gene of type V systems is *cas12*, also known as *cpf1*, which differs from Cas9 in the domain architecture. Cas12 only contains a RuvC-like domain responsible for the double strand-break (Strecker *et al.*, 2019; Swarts and Jinek, 2019). Type V systems display an extraordinary heterogeneity and are subdivided into 10 subtypes, V-A to V-K and V-U (unknown) (Figure I.9; Makarova *et al.*, 2020). Interestingly, type V are thought to evolve from TnpB proteins, encoded by IS605-like transposons (Shmakov *et al.*, 2017). As example of this great diversity, Cas12f (subtype V-F), also known as Cas14, is able to perform the degradation of both ssDNA (Harrington *et al.*, 2018) and dsDNA (Karvelis *et al.*, 2020). Subtype V-G effector, termed Cas12g, is an RNA-guided RNAse with collateral RNAse and ssDNAse activities (Yan *et al.*, 2019). Subtype V-K performs site-directed transposition through their nearby Tn7-like transposase (Strecker *et al.*, 2019), as the case of defective type I CRISPR-Cas systems (subtype I-F).

Cas12 protein possess a dual nuclease activity for both crRNA maturation and target cleavage. In the expression stage of subtype V-A, Cas12a binds the hairpin structure of the direct repeat trimming within the repeat to produce the mature crRNA (Fonfara *et al.*, 2016). Then, the 3' end is further processed, probably by host RNases (Swarts *et al.*, 2017). The mechanism of crRNA maturation in other type V subtypes still requires further experimental research. For example, subtypes V-B and V-E also encode a tracrRNA, which could be involved in crRNA biogenesis (Shmakov *et al.*, 2015; Koonin *et al.*, 2017).

Despite the diversity of type V CRISPR-Cas systems, only subtypes V-A and V-B interference mechanisms has been studied in detail. The major difference between these subtypes is that Cas12b requires the tracrRNA for interference (Shmakov *et al.*, 2015; Fonfara *et al.*, 2016). The effector proteins of both systems present a bilobed structure similar to Cas9, comprising REC and NUC lobes (Dong *et al.*, 2016). Contrary to other CRISPR-Cas systems, Cas12 proteins recognize PAM motifs on both strands (Fonfara *et al.*, 2016; Yang *et al.*, 2016). Upon PAM recognition and R-loop formation, the RuvC domain of Cas12a and Cas12b cleaves both DNA strands, resulting in double-strand break (Fonfara *et al.*, 2016; Yang *et al.*, 2016). In the case of Cas12a, it remains catalytically active after the cleavage and is able to degrade trans-ssDNA substrates (Swarts and Jinek, 2019), whereas Cas12b homologs have shown variability in sequence requirements for target recognition revealing species-specific variations (Jain *et al.*, 2019).

I.4.2.3 Type VI CRISPR-Cas Systems

Type VI CRISPR-Cas systems are characterized by the fact that their signature protein, Cas13, possess two HEPN domains and, thereafter, is involved in RNA interference (Abudayyeh *et al.*, 2016; East-Seletsky *et al.*, 2016). These systems may have evolved from an ancestral *cas13* which emerged from the recombination of two distant HEPN domains, possibly, from abortive infection modules (Koonin and Makarova, 2019). Since then, type VI CRISPR-Cas systems have diversified to the 4 subtypes known to date, VI-A to VI-D (Figure I.9; Makarova *et al.*, 2020). RNA-targeting activity have been experimentally validated for all subtypes except VI-C (Abudayyeh *et al.*, 2016; East-Seletsky *et al.*, 2016; Smargon *et al.*, 2017; Yan *et al.*, 2018). Every subtype present particular features, such as the Type VI-B loci which encodes genes responsible for the regulation (up and down) of the Cas13b activity (Smargon *et al.*, 2017), or subtype VI-D, that encodes a smaller effector protein (Cas13d) which activity is stimulated by an ancillary WYL-domain-containing protein (Yan *et al.*, 2018).

As occurring in type V CRISPR-Cas systems, the large type VI effector protein (Cas13) is the enzyme responsible for crRNA maturation by recognizing the structure of the direct repeat and trimming just upstream of the hairpin (East-Seletsky *et al.*, 2016). Curiously, crRNA maturation is not a requirement in subtype VI-A as pre-crRNAs can be used as guides as well (East-Seletsky *et al.*, 2017). One specific feature of type VI-B systems is that direct repeats can vary in length within the same CRISPR Array, thus, the effector can be associated to different mature crRNAs and still promote the cleavage of the target (Smargon *et al.*, 2017). Cas13d lacks a counterpart of the crRNA processing domain present in the other subtypes and, even so is able to perform the crRNA maturation process that may imply substantial structural flexibility, resulting in generation of different intermediate forms of pre-crRNA (Yan *et al.*, 2018).

In Cas13-family proteins, the two HEPN domains carry out the RNA cleavage upon crRNA-mediated target binding. Upon activation, Cas13 degrades the target RNA and also collateral ssRNAs, as described for Csm6 and Csx1 proteins in type III CRISPR-Cas interference (Abudayyeh *et al.*, 2016; East-Seletsky *et al.*, 2016; Smargon *et al.*, 2017; Yan *et al.*, 2018). As other Class 2 effector proteins, Cas13 also presents a REC and NUC lobe conforming a bilobed structure (Liu *et al.*, 2017). Sequence complementarity in the central part of the binding region between crRNA and the target sequence is required, but Cas13 tolerates mismatches in the flanking regions, especially in the 3′ end of the protospacer which does not need to base pair with the repeat sequence in the guide RNA for optimal target degradation (Meeske and Marraffini, 2018). Furthermore, different type VI CRISPR-Cas systems are able to confer defense against ssRNA viruses, such as *Escherichia coli* MS2 phage (Abudayyeh *et al.*, 2016; Smargon *et al.*, 2017). Furthermore, the non-specific collateral RNA degradation carry out by Cas13 can induce cell dormancy under dsDNA viruses attack, halting the growth of the infected cells and, therefore, protecting the population (Meeske *et al.*, 2019).

I.5. CRISPR-Cas adaptation

The ability to acquire novel spacers is the reason why immunity mediated by CRISPR-Cas systems is adaptive and heritable. Together with the CRISPR Array, Cas1 and Cas2 proteins are the essential components in the adaptation stage (Yosef *et al.*, 2012). These proteins are nearly universal as they are present in most CRISPR-Cas systems (Makarova *et al.*, 2015). The integrase complex responsible for spacer acquisition consists of two distal Cas1 dimers bound by two central Cas2 units, assembled into a heterohexamer "butterfly-like" structure (Figure I.10; Nuñez *et al.*, 2014; Nuñez *et al.*, 2015a). In the integrase complex, Cas1 harbor the catalytic endonuclease activity, whereas Cas2 plays a structural function required for the formation and the stabilization of the complex (Wang *et al.*, 2015). The adaptation machinery enables the integration of novel spacers after the leader proximal repeat of the CRISPR Array (Yosef *et al.*, 2012; Díez-Villaseñor *et al.*, 2013), thereby, the latter incorporated spacer is the first to be expressed, enhancing CRISPR-Cas defense against recent attackers (McGinn and Marraffini, 2016). Thus, a CRISPR Array reproduces the chronological encounters between phages and prokaryotic organisms highlighting phage-host evolution and ecology (Andersson and Banfield, 2008).

From an evolutionary perspective, the components of the integrase complex present different origins. The ancestors of Cas1 are a family of transposons called casposons, self-replicating MGEs capable of spreading due to site-specific casposase activity that leads to transposon integration in a similar way to the adaptation mechanism (Krupovic *et al.*, 2014; Béguin *et al.*, 2016). Furthermore, the terminal inverted repeats flanking the casposon are believed to evolve to generate the CRISPR direct repeats (Krupovic *et al.*, 2016). On the other hand, Cas2 proteins belong to the VapD toxins family, thus, their origin may be related to a toxin-antitoxin system (Makarova *et al.*, 2006). Therefore, it is thought that the interaction between a casposase and a toxin leads to a functional association that was domesticated by CRISPR-Cas systems during evolution (Koonin and Makarova, 2019). To understand CRISPR-Cas adaptation, the steps required for the acquisition of new spacers will be briefly described below.

I.5.1 Origin of prespacers

In the adaptation process, the first step is the recognition and processing of foreign genetic elements to generate a short sequence, known as prespacer, that will be integrated into the CRISPR Array as a novel spacer. Cas1-Cas2 integrase does not present any inherent capacity to distinguish self- versus non-self prespacers. Indeed, the acquisition of host-derived spacers results in autoimmunity that in most cases leads to cell death (Bikard *et al.*, 2012). Nevertheless, several examples show that self-targeting spacers are involved in gene regulation and many of them map to prophage regions (Stern *et al.*, 2010; Nobrega *et al.*, 2020). In any case, prokaryotic organisms have developed different strategies to overproduce spacers from invading elements in comparison with genome-derived spacers in order to prevent cytotoxic self-targeting.

The well-known mechanism to avoid autoimmunity is the RecBCD DNA repair system in Gram-negative bacteria (AddAB in the Gram-positive). RecBCD complex is recruited to double-strand breaks (DSBs) in replication folks to unwind and degrade DNA until reaching a Chi site, in which the activity of the complex ceases (Wigley, 2013). During the process of reparation, RecBCD generates ssDNA fragments partially annealed to produce duplexed substrates that are suitable for Cas1-Cas2 spacer acquisition (Figure I.10). The success of this mechanism lies in the fact that Chi sites are more frequent in bacterial chromosomes than in phages and plasmids. Thereby, the RecBCD complex degrades larger DNA portions in these foreign sequences generating more extrachromosomal substrates for CRISPR-Cas adaptation (Figure I.10; Levy *et al.*, 2015). Besides, RecBCD reparation takes place preferentially over linear DNA, biasing the selection against dsDNA phages (Modell *et al.*, 2017), whereas the circular host chromosome is protected (Modell *et al.*, 2017). This mechanism occurs when the cell faces the invading element for the first time, leading to a novel spacer acquisition known as naïve CRISPR adaptation (Fineran and Charpentier, 2012). However, this is an inefficient event to compensate the inability of the Cas1-Cas2 complex to discriminate self- versus non-self-sequences.

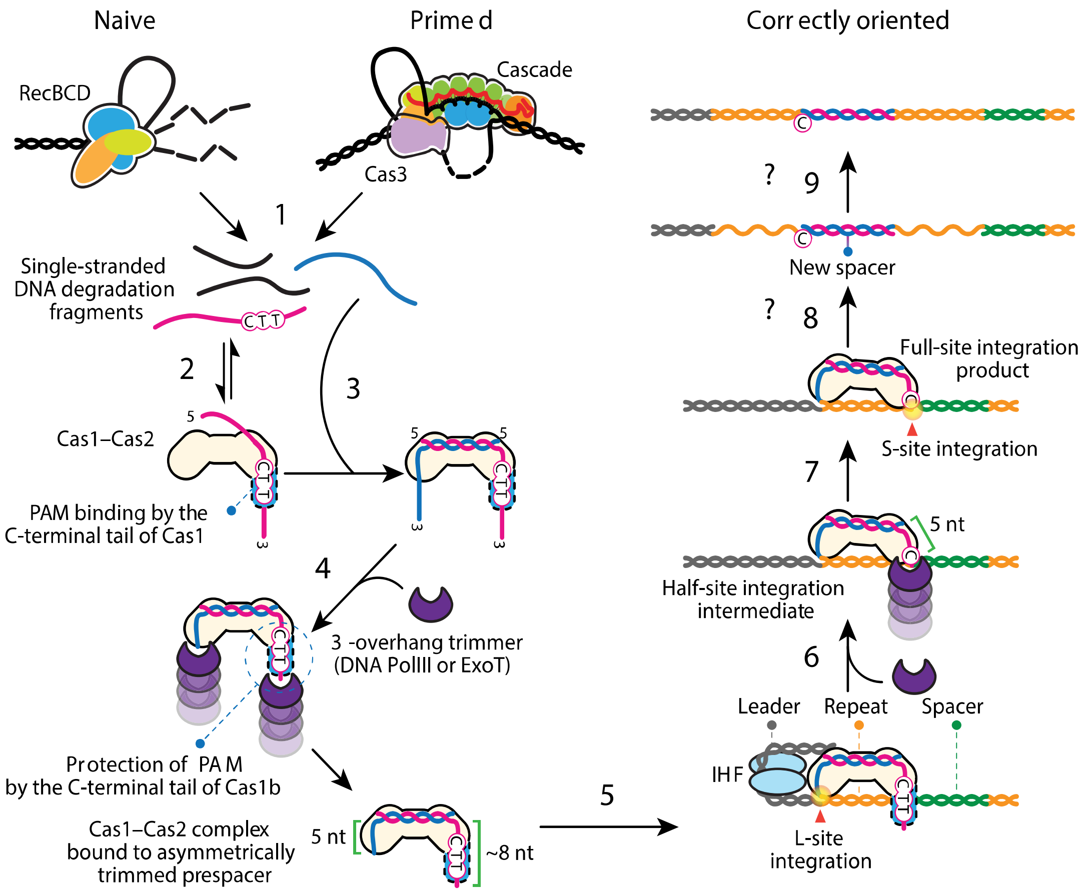


Figure I.10 CRISPR-Cas adaptation mechanism. (1) RecBCD and Cascade-Cas3 complexes generate ssDNA degradation fragments that can be used for naïve and primed adaptation, respectively**. (2)** Cas1-Cas2 complex capture PAM-containing sequences. **(3)** Complementary strands are annealed by Cas1-Cas2. Prespacers likely present 3'-overhangs at both ends. **(4)** DNAPolIII process the non-PAM strand, while PAM-containing strand is partially trimmed and protected by the C-terminal tail of Cas1. **(5)** The mature non-PAM-end is integrated at the leader side of the first repeat**. (6)** The PAM—derived-3'-overhang is released and further trimmed into the canonical size. **(7)** The mature PAM-end is integrated at the spacer side. (8) Cas1-Cas2 are released from the CRISPR locus. (9) DNA repair enzymes fill the gaps, duplicating repeats (Adapted from Kim *et al.*, 2020).

An additional strategy of prespacer selection is enabled when the CRISPR-Cas interference machinery is guided by a pre-existing spacer to a known invader. Target degradation by the effector complex generates DNA fragments that facilitate the rapid acquisition of additional spacers from the previously encountered genetic element in a process termed primed spacer acquisition (Figure I.10; Swarts *et al.*, 2012). This mechanism has been profoundly characterized in type I CRISPR-Cas systems, in which the Cascade complex promotes direct or Cas1-Cas2-stimulated recruitment of Cas3 after target binding (Xue *et al.*, 2016). The Cas3 nuclease activity produces multiple target degradation, resulting in DNA fragments that can be used by the integrase complex to increase the pool of integrated spacers against a specific foreign element ensuring an effective defense against viral escape mutants (Datsenko *et al.*, 2012; Fineran *et al.*, 2014). Indeed, in some type I-F CRISPR-Cas systems a Cas2-Cas3 natural fusion protein is present, presumably enhancing the selection of prespacers for new integrations events (Fagerlund *et al.*, 2017; Rollins *et al.*, 2017). Furthermore, a similar mechanism has been observed in type II-A CRISPR-Cas systems, in which the Cas9 activity increases the acquisition rate of spacers close to the target site (Nussenzweig *et al.*, 2019).

I.5.2 Cas1-Cas2 substrate capture and processing

After the generation of DNA fragments according to the mechanisms explained above, the selection of a prespacer compatible with spacer integration is a non-random event. In type I and type II CRISPR-Cas systems, Cas1-Cas2 integrase complex selects sequences with the protospacer adjacent motif (PAM), which is required for a functional interference stage (Swarts *et al.*, 2012; Datsenko *et al.*, 2012). In type I systems, the presence of a canonical PAM within the substrate enhances the affinity for Cas1-Cas2 binding but is not an indispensable requirement. The model of CRISPR-Cas adaptation is the type I-E Cas1-Cas2 complex from *E. coli*, which preferably selects dsDNA substrates with 3'-single-stranded-overhangs of at least 7 nucleotides in both strands (Wang *et al.*, 2015). In this complex, two tyrosine residues in the Cas1 subunits are the responsible for stabilizing the substrate. A recent model of type I-E adaptation stage shows that capture, maturation, and integration of the prespacer is a tightly coordinated process. Upon prespacer capture by the integrase complex, DNAPolIII and other DNAQ-like exonucleases trimming the substrate in order to generate an asymmetric prespacer with the correct features for integration and, meanwhile, the PAM motif is protected by the C-terminal tail of

Cas1 (Figure I.10; Kim *et al.*, 2020). Cas1-Cas2 structure also determine the correct spacer length by acting as a 'molecular ruler' (Wang *et al.*, 2015; Nuñez *et al.*, 2015b).

In other CRISPR-Cas systems, additional proteins are implicated in prespacer selection. In type II-A systems, Cas9 and Csn2 proteins are also required to ensure PAM-substrate selection (Wei *et al.*, 2015; Heler *et al.*, 2015). The Cas1-Cas2-Csn2 complex slides along the DNA until finding a Cas9 protein that remains bound to a PAM motif in the DNA. Through an unknown processing mechanism in which the Cas9 catalytic activity is dispensable, the PAM-terminal end of the protospacer is placed into the core of the Csn2 ring structure. Prior to integration, the Csn2 ring dissociates from DNA and allows the spacer integration by the Cas1-Cas2 complex (Wilkinson *et al.,* 2019). Moreover, in several subtypes of type I, II and V CRISPR-Cas systems, Cas4-family proteins are involved in PAM-dependent prespacer selection and processing (Lee *et al.,* 2018; Kieper *et al.,* 2018; Almendros *et al.,* 2019). In fact, *cas4* gene is present adjacent or fused to *cas1*, indicating a tight interaction between both domains (Makarova *et al.*, 2020). Cas4 is an endonuclease responsible for both, prespacer processing over a PAM substrate (Lee *et al.*, 2019) and the correct orientation of the spacer during the integration step (Shiimori *et al.*, 2018). Other CRISPR-Cas systems such as type III or type VI are characterized by a variable spacer length within a CRISPR Array, therefore, further research in the adaptation mechanisms of these systems is required to unravel how the substrate is captured in these systems and what proteins are involved in this step.

I.5.3. Spacer integration

Upon capture of a proper DNA substrate by the integrase complex, the recognition of the CRISPR Array is the next step in CRISPR-Cas adaptation. The integrase-DNA complex must be located near the leader-proximal repeat in order to integrate the spacer in the correct place. In type I-E complex from *E. coli*, this process is facilitated by the AT-rich leader sequence, which is recognized by a host protein, the integration host factor (IHF) (Nuñez *et al.,* 2015a; Nuñez *et al.,* 2016). IHF causes DNA bending in the leader region ensuring Cas1-Cas2 recognition of the first repeat, leading to a leader-polarized spacer integration (Wright *et al.*, 2017). The IHF requirement has been validated for type I-F system, and the presence of IHF binding sites in the leader sequence of multiple type I CRISPR-Cas systems suggest that this is a conserved mechanism in CRISPR adaptation (Nuñez *et al.*, 2016; Wright *et al.*, 2017; Yoganand *et al.*, 2017). Nevertheless, a large number of prokaryotic organisms harboring CRISPR-Cas system lack IHF, and recent surveys have confirmed that not yet identified host factors are required for the polarized acquisition of novel spacers in several type I systems (Rollie *et al.,* 2018; Grainy *et al.,* 2019).

For example, Type II-A CRISPR-Cas systems use an IHF-independent mechanism for specific leader-proximal repeat recognition based on an intrinsic affinity of the Cas1-Cas2 complex for the end of the leader sequence. In these systems, a short leader-anchoring site (LAS) just adjacent to the first repeat, and the approximately 6 first bases of the repeat are crucial for CRISPR-Cas adaptation (McGinn and Marraffini, 2016; Xiao *et al.*, 2017). Interestingly, mutation or deletion in the LAS sequence result in ectopic spacer integration in leader-distal repeats (McGinn and Marraffini, 2016), whereas the placement of an additional LAS before non-leader repeats leads to spacer acquisition in both sites (Wei *et al.,* 2015). The host factors or sequence requirements in other CRISPR-Cas systems for recognition of the CRISPR Array are mostly uncovered, thus, efforts in this direction are necessary to fully understand CRISPR adaptation.

Upon leader-proximal repeat recognition, the integration of the spacer into the CRISPR Array requires two transesterification reactions in both ends of the repeat in a mechanism similar to that of viral integrases and transposases (Nuñez *et al.*, 2015a). In type I-E CRISPR-Cas system from *E. coli*, the 3'-OH ends of the prespacer, which are generated during the spacer processing by DNAPolIII or homologous proteins, perform a nucleophilic attack on each strand of the leader-proximal repeat (Figure I.10; Nuñez *et al.,* 2015a; Rollie *et al.,* 2015; Kim *et al.,* 2020). The asymmetric maturation of the precursor carried out by these proteins leads to a bias for the first nucleophilic attack by the PAM-distal 3'-OH to occur in the leader-proximal end. Once the half-site intermediate is stabilized, the PAM in the other end is released from Cas1 and the substrate is processed to the final mature prespacer. Subsequently, the second nucleophilic attack takes place in the opposite strand of the leader-distal end of the repeat, resulting in a full-site integrated spacer (Figure I.10; Kim *et al.*, 2020). Moreover, DNAPolIII has also a DNA polymerase activity which may play a role in the restoration of the CRISPR Array to obtain the duplication of the directs repeats. Therefore, this recently proposed model further explains how is performed the integration of functional and correct oriented spacers within the CRISPR Array.

I.5.4 Other CRISPR adaptation mechanisms

Apart from the well-known CRISPR adaptation process in type I (especially I-E) and subtype II-A CRISPR-Cas systems, the acquisition of novel spacers has been studied in other systems. In subtypes V-C and V-D a mini-integrase complex comprising Cas1 alone catalyzes the integration of short DNA fragments (Wright *et al.*, 2019). In this variety, the integrase complex is formed by a Cas1 tetramer with an intrinsic ability to orient the spacers during integration. This smaller structure generates shorter spacers (18-20 nucleotides) in comparison with the system described above. Cas1 protein of these CRISPR-Cas systems could then represent an ancestral version that evolved from the casposase. This minimal integrase might later recruit Cas2 from a toxin-antitoxin system leading to the canonical Cas1-Cas2 complex present in most types which is able to acquire larger spacers increasing target specificity (Wright *et al.*, 2019).

In type III CRISPR-Cas systems, which target DNA transcriptionally active, a recent report shows that naturally acquired spacers in *Thermus thermophilus* after infection with phages are mainly originated from the early phage genome region and were complementary to phage transcripts (Artamonova *et al.*, 2020). Furthermore, the CRISPR-*cas* adaptation loci of some type III systems have recruited a reverse transcriptase gene, which could be involved in the acquisition of spacers from RNA phages or highly transcribed regions (Kojima and Kanehisa, 2008; Simon and Zimmerly, 2008; Toro and Nisa-Martínez, 2014). Interestingly, the role of reverse transcriptases in the adaptation process will be presented along this thesis and some of these RT-CRISPR-Cas systems has already been exploited for biotechnological application as it will be further explained in section below.

I.6. RT-CRISPR-Cas Systems

The relationship between RTs and these defense systems was described for the first time in 2008 (Kojima and Kanehisa, 2008; Simon and Zimmerly, 2008). Both works show that this group of RTs is found adjacent or fused to a *cas1* gene, which participates in the adaptive stage of CRISPR-Cas immunity as shown in the above sections. This observation was endorsed by broader studies of RTs linked to these systems (Toro and Nisa-Martínez, 2014). All the studies shown that this particular group of RTs is phylogenetically related to those from Group II introns, of which have likely evolved through a retrotransposition event (Figure I.2). The association between RTs and CRISPR–Cas systems has recently raised attention because of the experimental demonstration that the Cas6RTCas1 fusion protein from the type III-B system from *Marinomonas mediterranea* (MMB-1) is able to facilitate the RT-dependent acquisition of spacers directly from RNA molecules. Furthermore, this mechanism displays several similarities to group II intron retrohoming (Figure I.4D; Silas *et al.*, 2016).

Interestingly, the RT-containing CRISPR loci have been used to expand the CRISPR-toolbox as well. The adaptive operon of *Fusicatenibacter saccharivorans*, a complex formed by a RTCas1 fusion and a Cas2 protein, has allowed the development of Record-seq technology, a method for recording transcriptional events into CRISPR Array, describing specific and complex cellular behaviors assessing the cumulative gene expression (Table I.1; Schmidt *et al.*, 2018; Tanna *et al.*, 2020). Record-seq derived methods could also be used to improve the current technologies dedicate to archive real data in populations of living cells (Shipman *et al.,* 2017). Moreover, RT-containing CRISPR-Cas adaptive modules could be also used as highly scalable bacterial biosensors to report gut function (Tanna *et al.*, 2020). Nevertheless, the use of the *F. saccharivorans* system as RNA-recording tool present some disadvantages such as skewing to AT-rich regions at the ends of the transcripts. To overcome this limitation, result important to analyze other CRISPR-Cas system harboring RTs.

Since the role of RTs in CRISPR-Cas adaptation is only beginning to be understood, the aim of the present thesis is to analyze the phylogenetic distribution of RTs related to CRISPR-Cas systems in order to discover novel functional associations as well as the characterization of additional RT-containing CRISPR-Cas systems to better define the role of RTs in the immunity mediated by CRISPR-Cas systems.

Objectives

Objectives

Based on previous studies described above, the main objective of this thesis is deeping in our knowledge of prokaryotic RTs distribution focusing particularly on those RTs associated with CRISPR-Cas systems. This general objective has been developed by achieving the following specific objectives:

1. Establishing an update of the current diversity of RTs, focusing on the evolutionary origins and phylogenetic relationships of RTs associated with CRISPR-Cas systems.
2. Characterizing the formation of the adaptive complex in different examples of RT-containing CRISPR-Cas systems: those presenting a RT alone and those harboring a RTCas1 fusion.
3. Analyzing the *in vivo* implication of RTs in the acquisition of novel spacers in type III CRISPR-Cas systems.

Material and Methods

M.1. Phylogenetic analysis

M.1.1. Compilation of prokaryotic Reverse Transcriptases

M1.1.1. Compilation of bacterial Reverse Transcriptases

Different approaches were carried out to compile the bacterial RTs analyzed in Chapter 1 of the current thesis. Before a general compilation of total bacterial RTs, the known RTs associated with CRISPR-Cas system were compiled based on previous studies. The different steps performed for the generation of the RT dataset used to construct the phylogenetic tree of Figure R1.4 are shown in Figure M.1.

Firstly, 38 sequences corresponding with RTs associated with CRISPR-Cas systems identified by Makarova and co-workers (Makarova *et al.*, 2015) were compiled, including a unique RT associated with a CRISPR-Cas system found in archaea. To identify new sequences closely related to this archaeal RT, a search of the 262 complete archaeal genome sequence available from the PATRIC platform (Wattam *et al.*, 2014) was carried out. 120 sequences annotate as retron-type RNA-directed DNA polymerases (EC2.7.7.49) were retrieved. After removing duplicates, the unique RTs with ≥200 amino acid residues displaying ≤85% identity were 46 sequences. These were aligned with a previous RT dataset (RT 0-7 domains) of 742 sequences (Toro and Nisa-Martínez, 2014). With this preliminary phylogenetic analysis, 7 sequences were clustered in a well-supported clade. Using the CRISPR recognition tool implemented in Genious Pro software (Biomatter Ltd.; Bland *et al.*, 2007; Kearse *et al.*, 2012), we found that only 5 of the RTs had CRISPR Arrays associated. After a search for possible bacterial members of the archaeal clade by performing a BLAST search of 3043 annotated RTs with ≥200 amino acid residues from 6291 complete bacterial genome sequences available from the PATRIC platform (Wattam *et al.*, 2014) any new RT was added to this clade: whereas the significant e-values for archaeal RTs within the above-mentioned clade were in the range of 9.22e-97 to 1.39e-127, none of the 3043 analyzed sequences had an e-value ≤ e-75.

Besides the 38 RTs associated with CRISPR-Cas systems identified by Makarova *et al.,* 2015 and the 7 archaeal RTs, 14 RTs associated with CRISPR-Cas systems reported in Toro and Nisa-Martínez, 2014 together with 157 protein sequences that were obtained from the National Center for Biotechnology Information (NCBI) Conserved Domain Architecture Retrieval Tool (CDART; 24-Oct-2016) based on the presence of both an RT domain (pfam00078) and a Cas1 domain (pfam01867) were retrieved as well. Using the cutoffs indicated above (≥200 amino acid residues displaying ≤85% identity), we reduced the dataset to 118 unique RT sequences associated with CRISPR-Cas systems (Figure M.1).

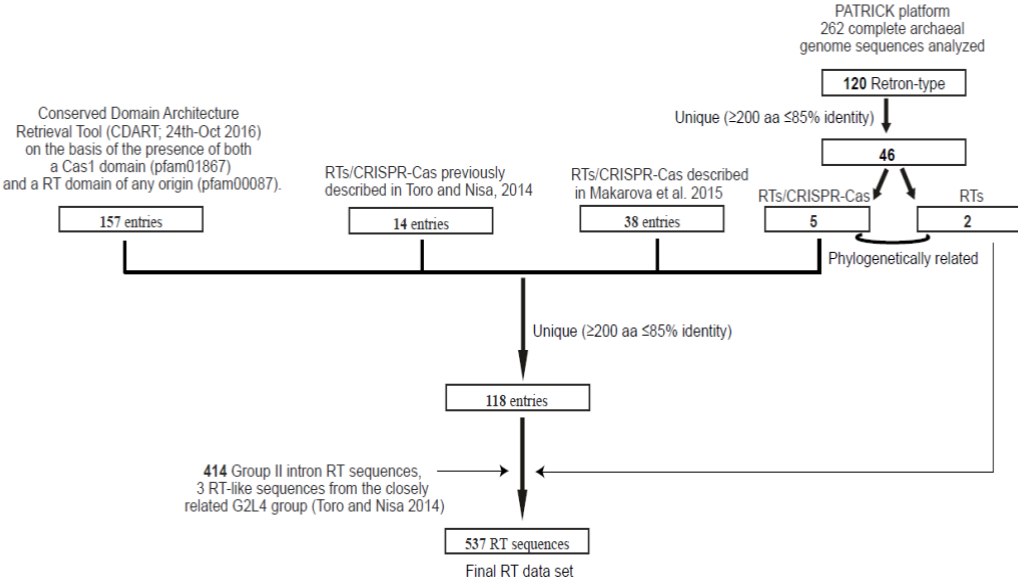

PATRICK platform
262 complete archaeal
genome sequences analyzed
120 Retron-type
Unique (≥200 aa ≤85% identity)
46
RTs/CRISPR-Cas
RTs
5
2
Phylogenetically related
Conserved Domain Architecture
Retrieval Tool (CDART; 24th-Oct 2016)
on the basis of the presence of both
a Cas1 domain (pfam01867)
and a RT domain of any origin (pfam00087).
157 entries
RTs/CRISPR-Cas previously
described in Toro and Nisa, 2014
14 entries
RTs/CRISPR-Cas described
in Makarova et al. 2015
38 entries
Unique (≥200 aa ≤85% identity)
118 entries
414 Group II intron RT sequences,
3 RT-like sequences from the closely
related G2L4 group (Toro and Nisa 2014)
537 RT sequences
Final RT data set

Figure M.1 Steps performed for the compilation of RTs associated to CRISPR-Cas systems and the generation of the dataset (Adapted from Toro *et al.*, 2017; details in the main text).

The RT dataset used to construct the phylogenetic tree shown in Figure R1.6, includes new 19 RT-like sequences that were added to the previous dataset, expanding to 137 RT sequences associated with CRISPR-Cas systems. These RTs were identified in branches 7, 8 and 9 reported by (Silas *et al.,* 2017a).

Finally, to generate the heterogenous RT dataset of 9141 RT sequences used to construct the phylogenetic trees of Figure R1.1A and R1.8, different approaches to encompass the wide diversity of prokaryotic RTs were used (Figure M.2). 133 new RTCas1 proteins were retrieved from CDART (23-Feb-2018) and were merged with RT sequences annotated as RNA-directed DNA polymerases (145.379) or reverse transcriptases (52.684) that were downloaded from the PATRIC webserver (Wattam *et al.*, 2014). Thereby, all these sequences were incorporated into the previously analyzed dataset of 558 RT sequences, including the 137 RT/RTCas1 associated with CRISPR-Cas systems described above. Additionally, 6 putative RT sequences from *Streptomyces* species, predicted to be linked to type I-E CRISPR-Cas systems (Silas *et al.*, 2017; Shmakov *et al.*, 2018), were included in the study. Filtering in multiple-step clustering with the cutoffs described above (≥200 amino acid residues displaying ≤85% identity) this procedure resulted in a final dataset of 9141 diverse unique RTs (Figure M.2).

M1.1.2. Compilation of archaeal Reverse Transcriptases

To build archaeal RT dataset to construct the phylogenetic tree of figure R1.1B two different approaches were used. RT sequences annotated as "RNA-directed DNA polymerases" (954 sequences) were downloaded from the PATRIC web server and 537 additional sequences (14-Apr-2020) were obtained from Uniprot database (https://www.uniprot.org/) using the hidden Markov model of Pfam00078 with the jackhammer tool from the HMMER suite (http://hmmer.org/). The resulted dataset of 1,491 protein sequences was then filtered by selecting the RT domain (RT0-7) of the proteins with a length ≥ 200 amino acids, in multiple-step clustering with a threshold of 85% sequence identity. This procedure reduced the final dataset to 411 diverse unique archaeal RTs.

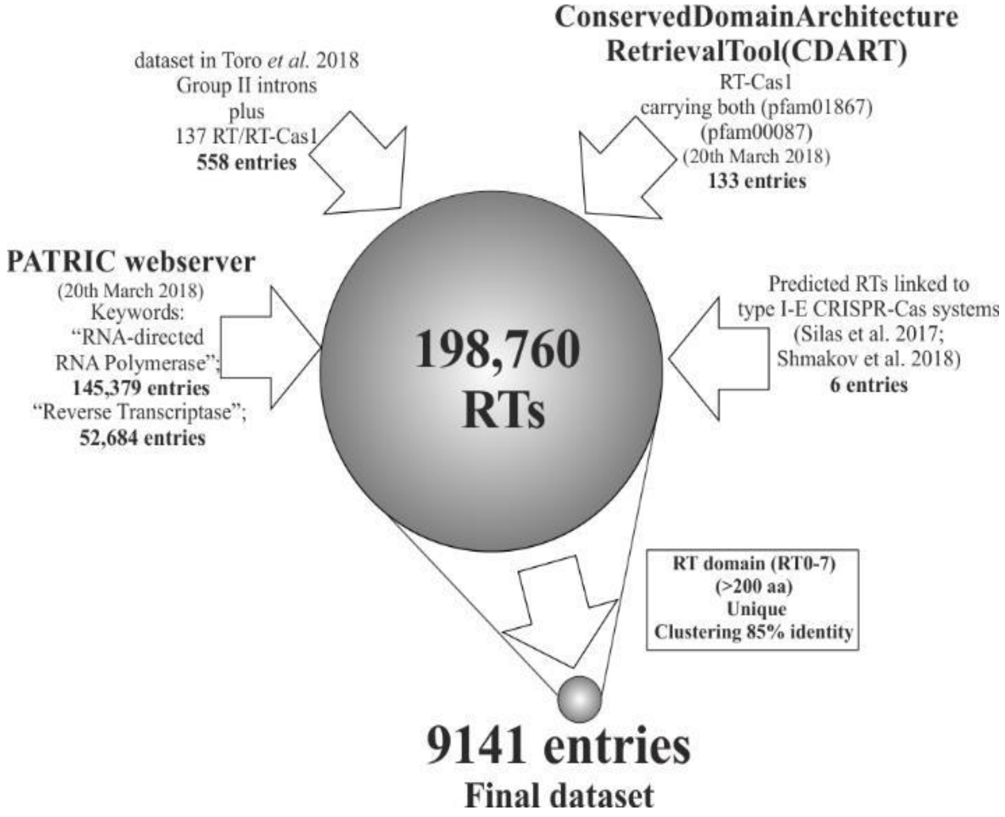


Figure M.2 Compilation of RTs from databases and generation of the dataset. The procedure depicted yields 9141 predicted unique sequences representative of the current diversity of RTs in prokaryotes (Adapted from Toro *et al.*, 2019a; details in the main text).

M.1.2. Generation of protein phylogenetic trees

To construct all the protein phylogenetic trees of this work, the standard protein sequences phylogenetic analysis was performed as follows: the sequences of each protein were aligned with MUSCLE (Edgar, 2004) or MAFFT (Katoh and Standley, 2013) software. Therefore, a phylogenetic tree was constructed with the FastTree program and the WAG evolutionary model, using pseudocounts (recommended for sequences containing large numbers of gaps) and a discrete gamma model with 20 rate categories.

In the trees shown in Appendix A3, *IQ-TREE* v. 1.6.1 (Nguyen *et al.*, 2015) was also used to infer phylogenetic trees using the amino acid substitution best-fit model (LG+G4) provided by ModelFinder selected on the basis of the Bayesian information criterion (BIC). Branch support was assessed by ultrafast bootstrap approximation (UFBoot), and the impact of severe model violations was reduced by using hill-climbing nearest interchange (NNI) search, SH (Shimodaira-Hasegawa)-aLRT (approximate likelihood ratio test, (Guindon *et al.*, 2010) with 1000 replicates, and standard non-parametric bootstrap (100 replicates).

M.1.2.1 RT sequences phylogenetic analysis

The 118 or 137 RT-CRISPR sequences encompassing RT motifs (RT 0-7) were aligned (250 positions) against 414 RT sequences representative of group II introns, 3 RT-like sequences from the closely related G2L4 group (Toro and Nisa-Martínez, 2014), and 2 RT sequences from archaea related to the archaeal RTs associated with CRISPR-Cas systems (Figure M.1). In all FastTree phylogenetic trees, the clades were assigned to the inner nodes showing a high local support value (≥0.9), and subclades were assigned when a large number of sequences were restricted to particular phyla. When IQ-Tree is used, the clades were designated whether they had a standard non-parametric bootstrap value ≥ 70% or SH-aLRT and Ufboot values ≥ 80%.

M.1.2.2. Cas1 sequences phylogenetic analysis

To construct the phylogenetic Cas1 tree shown in Figure R1.5, 148 unique Cas1 sequences, which were either separate or fused to the first 137 RT dataset associated with CRISPR-Cas systems described above, were aligned (329 positions) using as outgroup the Cas1 protein (unknown subtype) from *Arthospira platensis* (GI:479129287; (Makarova *et al.*, 2015). Therefore, the FastTree phylogenetic tree was constructed as described above.

M.1.3. Identification of RT related CRISPR-cas loci

The genomic neighborhoods (up to 50 kb in some cases) of the RT/RTCas1 sequences associated with CRISPR-Cas systems included in the final dataset were analyzed searching for CRISPR Arrays and *cas* genes encoding proteins.

M.1.3.1 CRISPR Array

In most cases, in the proximity (less than 1 kb) of the RT gene, a CRISPR Array was identified. The characteristics of each CRISPR Array were determined with CRISPRFinder (Grissa *et al.*, 2007b), CRISPRmap (Lange *et al.*, 2013), CRISPRDetect (Biswas *et al.*, 2016) and CRISPRstrand (Alkhnbashi *et al.*, 2014) tools, which provided information about the orientation, the number of spacers and their mean length among other properties. The correct orientation of the array was determined on the basis of the following criteria: (i) orientation predicted by the CRISPRDetect algorithm with a score of H or M if the flanking region of the array was available (>200 nt), (ii) for scores of L or NA, orientation was determined on the basis of the presence of Direct Repeats (DRs) in the CRISPRstrand database, and (iii) DR similarities between arrays of other members of the group with a predicted orientation.

M.1.3.2 CRISPR-Cas genes

The identification of *cas* genes was performed based on the consensus sequence from 395 profiles of CRISPR associated proteins described in Makarova *et al.*, 2015. A BLAST search was carried out and all the protein-coding genes present in the genomic region flanking the RT gene were annotated. An *e*-value of 0.01 was used, except for subtype specificity, for which an e-value threshold of 10^{-6} was used. The genomic regions containing all the identified *cas* genes and CRISPR Arrays were extracted and the region carrying the RT was trimmed to the first and last *cas* gene and/or the CRISPR Array carrying intervening sequences of less than 5 kb in length.

M.2. Bacterial strains

Bacterial strains and genomic DNAs used in this work are described in Tables M1 and M2 respectively.

Table M1. Bacterial strains used in this work.

Bacterial Strains	Characteristics	Reference
Escherichia coli DH5*α*	*F*$^{-}$*,Ø80dlacZΔM15, Δ(lacZya-argF)U169, deoR, recA1, endA1, hsdr17(rK*$^{-}$*, mK*$^{+}$*), phoA, supE44, λ*$^{-}$*, thi-1, gyrA96, relA1.*	Bethesda Research Lab
E. coli HMS 174 (DE3)	*F- recA1 hsdR(rK12- mK12+) (DE3) (Rif R)*	Novagen
E. coli JM109 (DE3*)*	*endA1 glnV44 thi-1 relA1 gyrA96 recA1 mcrB+ Δ(lac-proAB) e14- [F' traD36 proAB+ lacIq lacZΔM15] hsdR17(rK-mK+) λ(DE3)*	Promega Corporation
E. coli BL21 (DE3)	*F– ompT gal dcm lon hsdSB(rB–mB–) λ(DE3 [lacI lacUV5-T7p07 ind1 sam7 nin5]) [malB+]K-12(λS)*	New England Biolabs
E. coli BL21 (AI)	*F– ompT gal dcm lon hsdSB(rB–mB–) [malB+]K-12(λS) araB::T7RNAP-tetA*	Thermo Fisher Scientific
E. coli Rosetta™ 2(DE3)pLysS	*F– ompT gal dcm lon? hsdSB(rB–mB–) λ(DE3 [lacI lacUV5T7p07 ind1 sam7 nin5])[malB+]K12(λS)pL ysSRARE[T7p20 ileX argU thrU tyrU glyT thrT argW me tT leuW proL orip15A](CmR)*	Novagen
E. coli CC118 λpyr	*Δ(ara-leu) araD ΔlacX74 galE galK phoA20 thi-1 rpsE rpo argE* *(Am) recA1 λpir phage lysogen*	Herrero *et al.,* (1990)
Vibrio vulnificus YJ016	Strain harboring RT-CRISPR-Cas type III-D system.	Chen *et al.,* (2003)
V. vulnificus R99	RT-CRISPR-Cas System Type III-D-less strain.	Amaro *et al.,* (1990)

Table M2. Source of Genomic DNA for RT/RTCas1 sequences amplification. Unless specified, DNA was ordered to DSMZ collection (https://www.dszm.de/).

Phylum/Sub	Specie	Strain	Clade^{a}	Accesion number (NCBI)	Reference
δ-proteobacteria	*Desulfobacca acetoxidans*	DSM 11109	4	NC_015388.1	Oude-Elferink *et al.,* (1999)
Chlorobi	*Cholorobium limícola*	DSM 245	4	NC_010803.1	Imhoff, (2003)
Cyanobacteria	*Scytonema hofmanni*	PCC 7110^{b}	5	KQ976354.1	Rippka *et al.,* (1979)
γ-proteobacteria	*Vibrio vulnificus*	YJ016^{c}	6	NC_005139.1 NC_005140.1 NC_005128.1	Chen *et al.,* (2003)

Chloroflexi	*Roseiflexus castenholzii*	DSM 13941	9	NC_009767.1	Hanada *et al.*, (2002)

[a]RT/RTCas1 clade according to results shown in chapter 1
[b]kindly provided by Dr Agustin Vioque (IBVF-CSIC, Seville)
[c]kindly provided by Dr Carmen Amaro (University of Valencia)

M.3. Bacterial cultures

M.3.1. Culture media

E. coli strains were grown in Luria-Bertani (LB) medium at 37ºC (Sambrook *et al.*, 1989): 10g/L tryptone, 5g/L yeast extract, 5g/L NaCl made with deionized water (MQ water) and adjusted to pH 7. The solid medium was supplemented with 1.6% w/v agar (PANREAC). It was sterilized by autoclaving at 121ºC for 20 minutes.

The cyanobacteria *Scytonema hofmanni* PCC 7110 was obtained from the cyanobacterial collection from Agustin Vioque group in IBVF-CSIC (Seville). The culture was grown in BG11O± N source (nitrate or ammonium) at room temperature under environmental light (Rippka *et al.*, 1979).

Vibrio vulnificus strains YJ016 and R99 were grown by Carmen Amaro group in Valencia, in Tryptone Soy agar (St) or Thiosulfate-citrate-bile salts agar (TCBS) mediums at 28ºC (Miller, 1972; Pfeffer and Oliver, 2003).

M.3.2. Antibiotics

Antibiotics were made as 100x stock solution in MQ water and filter sterilized through 0.2 µm membrane Minisart® NML (Sartorius). Antibiotics and concentrations used in this work are listed in Table M3.

Table M3. Antibiotics used in this work.

Antibiotics	Stock Concentration (100x)
Ampicilin (Ap, Sigma)	200 g/L
Chloramphenicol (Cm, Sigma)	50 g/L
Tetracycline (Tc, Sigma)	10 g/L

M.3.3. Storage of bacterial cultures

Freezing was used for prolonged preservation of bacterial cultures. This method is based on the paralysis of cellular metabolism to decrease water availability. In order to maintain cell viability during the preservation period, cryoprotectants are required, which avoids the damage that water crystals could cause to cell walls. In our case, the cryoprotectant chosen was glycerol. A concentration of approximately 25% (v/v) was used, so that, at 250µl of glycerol, arranged in cryotubes, 750 µl of a bacterial culture grown in liquid medium until stationary phase was added. The vials were quickly frozen and stored at -80 ° C per duplicate.

M.4. Plasmids and cloning vectors

M.4.1. Basic plasmids and vectors used in this work

Basic vectors and plasmids used in this thesis are listed in Table M4.

Table M4. Basic plasmids and vectors used in this work.

Name	Characteristics	Reference
pGEM-T Easy	Vector for the cloning of PCR products. High copy number. Recombinant cloned identified by color screening. 3015 bp. Ap^R	Promega
pMal-c5X	Vector designed to produce maltose-binding protein (MBP) fusions, where the protein of interest can be cleaved from MBP with the specific protease Factor Xa. 5677 bp. Ap^R	NEB
pMal-Flag-IEP	Plasmid derived from pMal-c5X where is cloned as *Not*I a fragment that content the IEP (Intron Encoded Protein) from RmIntI labelled with the epitope Flag derived from pCEP4-Flag-IEP plasmid (Reinoso-Colacio *et al.,* 2015). This plasmid expresses the IEP as a fusion protein fused to MBP. 7048 bp. Ap^R	(García-Rodríguez *et al.,* 2019)
pET16b	Bacterial vector for inducible expression of N-terminally 10xHis-tagged proteins with a Factor Xa site. 5711 bp. Ap^R	Novagen
pMP220	Broad-host range transcriptional fusion vector. Precursor of pCA plasmids. It contains a multicloning sequencing site, a Shine-Dalgarno sequence from *E. coli* CAT gene and the β-galactosidase *lacZ* gene of *E. coli*. 10.5 kbp. Tc^R	(Spaink et al., 1987)
pVSV-105	A shuttle vector carrying an origin of replication derived from a *Vibrio fischeri*-specific plasmid pES213. Cloning purposes with this vector were carried out in *E. coli* CC118 λpyr. 6.5 kbp. Cm^R	(Dunn et al., 2006)

M.4.2. Plasmids for protein expression and purification

All plasmids used in this thesis for protein expression and purification and biochemical purposes are listed in Table M5. To measure reverse transcriptase activity the pMal-Flag backbone was used. pMal plasmids containing point mutations in the RT (YADD to YAAA at amino acid position 220 to 223) or the Cas1 domain (E517A and E597A) of the RT*cas1* gen of *V. vulnificus* YJ016 were generating through double PCR (section M.8.1.) with oligonucleotides containing the mutations. All plasmids were verified by sequencing. The oligonucleotides used to make plasmids for different proteins expression and purification are listed in Table M6.

Table M5. Plasmids for protein expression and purification.

Plasmid	Description
pMal-Flag	pMal-Flag-IEP derivative where the IEP was escinded as *Bam*HI and the plasmid was recirculated.
pMal-Flag-418	pMal-Flag derivative containing RT-418 from *Roseiflexus castenholzii* DSM 13941 as a *Bam*HI DNA fragment. Of the amplicon using primers 418a/418b.
pMal-Flag-432N2	pMal-Flag derivative containing RT-432N2 from *Scytonema hofmanni* PCC 7110 as a *Bam*HI/*Bgl*II DNA fragment using primers 432N2a/432N2b.
pMal-Flag-439	pMal-Flag derivative containing RTCas1-439 from *Vibrio vulnificus* YJ016 as a *Bgl*II DNA fragment using primer 439a/439b.
pMal-Flag-439-YAAA	pMal-Flag derivative containing the point mutation in the RT domain of RTCas1-439 from *Vibrio vulnificus* YJ016 as a *Bgl*II DNA fragment.
pMal-Flag-439-E517A	pMal-Flag derivative containing the point mutation E517A in the Cas1 domain of RTCas1-439 from *Vibrio vulnificus* YJ016 as a *Bgl*II DNAfragment.
pMal-Flag-439-E597A	pMal-Flag derivative containing the point mutation E597A in the Cas1 domain of RTCas1-439 from *Vibrio vulnificus* YJ016 as a *Bgl*II DNA fragment.
pMal-Flag-441N1	pMal-Flag derivative containing RTCas1-441N1 from *Chlorobium limicola* DSM 245 as a *Bgl*II DNA fragment using primers 441N1a/441N1b.
pMal-Flag-443	pMal-Flag derivative containing RT-443 from *Desulfobacca acetoxidans* DSM 11109 as a *Bgl*II DNA fragment using primers 443a/443b.
pET16b-Cas2A-439	pET16b derivative containing the *cas2A* gen from *Vibrio vulnificus* YJ016 as a *Nde*I/*Bam*HI DNA fragment.
pET16b-Cas2A-104-439	pET16b derivative containing the *cas2A* gen from *Vibrio vulnificus* YJ016 with 10 more amino acids at the N-Terminal end cloned as a *Nde*I/*Bgl*II DNA fragment.
pET16b-Cas2B-439	pET16b derivative containing the *cas2B* gen from *Vibrio vulnificus* YJ016 as a *Nde*I/*Bam*HI DNA fragment.
pMal-Op439	pMal-Flag derivative containing the whole adaptive operon of *Vibrio vulnificus* YJ016 with the *RTCas1* gen fused to the MBP and the two Cas2 as *Bgl*II DNA fragment.
pMal-C5x-439	pMal-C5x derivative containing RTCas1-439 from *Vibrio vulnificus* YJ016 as a *Bgl*II DNA fragment.
pMal-Cas2A-439	pMal-C5x derivative containing *cas2A* gen from *Vibrio vulnificus* YJ016 as a *Bgl*II DNA fragment.
pMal-Cas2B-439	pMal-C5x derivative containing *cas2B* gen from *Vibrio vulnificus* YJ016 as a *Bgl*II DNA fragment.

pMal-OpHisCas2B-439	pMal-C5x derivative containing the whole adaptive operon of *Vibrio vulnificus* YJ016 with the *RTCas1* gen fused to the MBP, *cas2A* gen, and *cas2B* gen with 6x His-tag at the N-terminal end cloned as *Nde*I/HindIII DNA fragment.
pMal-C5x-3C	pMal-C5x derivative in which the Xa Factor recognition site has been replaced by the recognition site of the 3C protease.
pMal-3C-RTCas1-439	pMal-C5x-3C derivative containing RTCas1-439 from *Vibrio vulnificus* YJ016 as a *Bgl*II DNA fragment.
pMal-3C-Cas2A	pMal-C5x-3C derivative containing *cas2A* gen from *Vibrio vulnificus* YJ016 as a *Bgl*II DNA fragment.
pMal-3C-cas2B	pMal-C5x-3C derivative containing *cas2B* gen from *Vibrio vulnificus* YJ016 as a *Bgl*II DNA fragment.
pMal-Xa-Cas2A-HisCasB	pMal-C5X derivative containing *cas2A* gen, and *cas2B* gen with 6x His-tag at the N-terminal end from *Vibrio vulnificus* YJ016 cloned as *Nde*I/*Hind*III DNA fragment.
pMal-3C-Cas2A-HisCasB	pMal-C5X-3C derivative containing *cas2A* gen, and *cas2B* gen with 6x His-tag at the N-terminal end from *Vibrio vulnificus* YJ016 cloned as *Nde*I/*Hind*III DNA fragment.
pET16b-Cas1-432N2	pET16b derivative containing the *cas1* gen from *Scytonema hofmanni* PCC 7110 as a *Nde*I/*Bam*HI DNA fragment.
pET16b-Cas2-432N2	pET16b derivative containing the *cas2* gen from *Scytonema hofmanni* PCC 7110 as a *Nde*I/*Bam*HI DNA fragment.
pMal-WYL-432N2	pMal-Flag derivative containing the WYL domain containing protein gen from the RT operon of Scytonema hofmanni PCC 7110 as a *Bam*HI/*Bgl*II DNA fragment.
pET16b-WYL-432N2	pET16b derivative containing the WYL domain containing protein gen from the RT operon of Scytonema hofmanni PCC 7110 as a *Bam*HI/*Bgl*II DNA fragment.
pMal-Ph-432N2	pMal-Flag derivative containing the putative phosphohydrolase gen from the RT operon of Scytonema hofmanni PCC 7110 as a *Bam*HI/*Bgl*II DNA fragment.
pET16b-Ph-432N2	pET16b derivative containing the putative phosphohydrolase gen from the RT operon of Scytonema hofmanni PCC 7110 as a *Bam*HI/*Bgl*II DNA fragment.
pVSV-105-OmpC-FlagRTCas1	pVSV-105-OmpC derivative containing FlagRTCas1-439 from *Vibrio vulnificus* YJ016 in the same ORF that the ATG from *ompC* gene cloned in *Sph*I restriction site.
pVSV-105-OmpC-RTCas1	pVSV-105-OmpC derivative containing RTCas1-439 from *Vibrio vulnificus* YJ016 in the same ORF that the ATG from *ompC* gene cloned in *Sph*I restriction site.

Table M6. Oligonucleotides used for the construction of plasmids for expression and purification.

Primer	Sequence (5´-3´)	Description
418a	GGGGGATCCG**ATG**CCGCTTTTTCCCCGTT	F RT-418
418b	GGGAGATCT**TCA**ACCTTCAACTTTCCCAC	R RT-418
432a	GGGGGATCCC**ATG**GTAAATATGAGTGATATTG	F RT-432
432b	GGGAGATC**TTA**TCTGGGAGAGAGATTCTT	R RT-432
439a	GGGAGATCTT**ATG**GCTATTCACCTTACGATT	F RTCas1-439
439b	GGGAGATC**TTA**GCGGATCCTCTGCTGAG	R RTCas1-439
441N1a	GGGGGATCCT**ATG**GGATGGCTCTACAACCA	F RTCas1-441N1
441N1b	GGGAGATCT**TCA**CCATACCAGCTTGTACG	R RTCas1-441N1
443a	GGGGGATCCA**ATG**ATCTCTGAAATAACAGTCA	F RT-443
443b	GGGAGATC**TTA**TACATCAATGTTATCATTACC	R RT-418
Cas2A-439f	GGGCAT**ATG**AGCCGTGAACATTACGTG	F Cas2A-439

Cas2A-439r	GGGGGATCC**TAA**TCAATATAATCAAAAGGTTG	R Cas2A-439
Cas2A-104-439f	GGGCAT**ATG**GTCTCTCACTCAGCAGA	F Cas2A with 10 more aa' at N-terminal end
Cas2A-104-439r	GGGAGATCT**TAA**TCAATATAATCAAAAGGTTG	R Cas2A with 10 more aa' at N-terminal end
Cas2B-439f	GGGCAT**ATG**CTTTGGTTGATCAGCTTTG	F Cas2B-439
Cas2B-439r	GGGGGATCC**TTA**TTCGTCAATCAAATACAGTG	R Cas2B-439
Cas2B-439-HindIIIr	GGGAAGC**TTA**TTCGTCAATCAAATACAGT	R Cas2B-439 *Hind*IIII
HisCas2Bf	AAGAAGGAGATATACCATGGGCCATCATCATCATCAT CACAGCAGCGGCC	F 6xHis for Op439-HisCas2B
HisCas2Br	CCATGGTATATCTCCTTCTTGTGTACCTCTAATCAA	R 6xHis for Op439-HisCas2B
3Cf	CCGCGGCTCTTGAGGTGCTCTTTCAGGGACCCGGGTA CCAGCA	F 3C Protease
3Cr	TATGCTGGTACCCGGGTCCCTGAAAGAGCACCTCAAG AGCCGCGGAGCT	R 3C Protease
VvFlag-RTCas1-SpHI	GGGGC**ATG**CAAGACTACAAAGACCATGACGGT	F Flag-RTCas1 for cloning in pVSV-105-OmpC
VvRTCas1-SphI	GGGGC**ATG**CAAATGGCTATTCACCTTACGATT	F RTCas1 for cloning in pVSV-105-OmpC
Cas1-432f	GGGCAT**ATG**GCAGATATTGCGACTTTAC	F Cas1-432
Cas1-432r	GGGGGATCC**TTA**CACTGCTCTTAAAAAAGGTT	R Cas1-432
Cas2-432f	GGGCAT**ATG**CTAGTGCTTGTAGTGTATG	F Cas2-432
Cas2-432r	GGGGGATCC**TTA**GATAACATAATACTTTGGCG	R Cas2-432
WYL-432f	GGGGGATCCG**ATG**AAGAGAGAAGTTTTTAACT	F WYL domain containing protein 432
WYL-432r	GGGGGATCC**TTA**TTTATAAAGTTTGTAAGTTG	R WYL domain containing protein 432
Ph-432f	GGGGGATCCG**ATG**AACTTAAACGCCACGG	F Phosphohydrolase 432
Ph-432r	GGGGGATCC**TTA**CTGGTCACTGTTAAACTT	R Phosphohydrolase 432
FlagRTCas1-SphI	GGGGCATGCAAGACTACAAAGACCATGACGGT	F pVSV-105-FlagRTCas1
RTCas1-SphI	GGGGCATGCAA**ATG**GCTATTCACCTTACGATT	F pVSV-105-RTCas1

*F = Forward; R = Reverse
*The designed restriction sites are underlined.
*The start or end codons of the gene are shown in bold.

M.4.3. *Plasmids for in vivo spacer acquisition assays in* Escherichia coli *host*

Plasmids for inducible overexpression of the adaptive operon from *V. vulnificus* YJ016 (RTCas1-Cas2A-Cas2B) or *S. hofmanni* PCC7110 (RT-Cas1-Cas2) were built in the pGEM-T Easy backbone (Promega). The two CRISPR Arrays associated with both systems complete or only containing the first DR and the first spacer were cloned into pMP220 backbone. All plasmids were verified by sequencing and listed in table M7. The oligonucleotides used to construct the plasmids for *in vivo* assays are presented in table M8.

Tabla M7. Plasmids for *in vivo* spacer acquisition assays in *E. coli* host.

Plasmid	Description
pAGDt-439	pGEM-T Easy (Promega) derivative containing the adaptive operon from *V. vulnificus* YJ016 (RTCas1-Cas2A-Cas2B) under the control of T7 promoter.
pAGDt-439-YAAA	pAGDt-439 derivative with a point mutation in the RT domain (YADD to YAAA at amino acid position 220 to 223).
pAGDt-439-E517A	pAGDt-439 derivative with a point mutation in the Cas1 domain (E517A).
pAGDt-439-E597A	pAGDt-439 derivative with a point mutation in the Cas1 domain (E597A).
pAGDt-439-ΔCas2B	pAGDt-439 derivative with a deletion of *cas2B* gene.
pAGDt-439-ΔCas2A	pAGDt-439 derivative with a deletion of *cas2A* gene.
pAGDt-439-ΔCas2AB	pAGDt-439 derivative with a deletion of both Cas2 sequences.
pAGDt-439-tdI	pAGDt-439 derivative with tdI sequence and its native exons cloned as a *Sal*I/*Xho*I DNA fragment in the same orientation under the control of the T7 promoter.
pAGDt-432	pGEM-T Easy (Promega) derivative containing the adaptive operon from *S. hofmanni* YJ016 (WYL-Ph-RT-Cas1-Cas2 and the CRISPR Array 1).
pAGDt-432-ΔCA	pAGDt-432 derivative without CRISPR Array 1.
pAGDt-RTCas-432	pAGDt-432-ΔCA derivative containing the RT, *cas1* and *cas2* genes from *S. hofmanni* PCC7110.
pAGDt-Cas-432	pAGDt-432-ΔCA derivative containing *cas1* and *cas2* genes from *S. hofmanni* PCC7110.
pCA1s-439	pMP220 derivative with the entire CRISPR Array 1 of *V. vulnificus* YJ106 cloned as an *Eco*RI DNA fragment in the same orientation as *lacZ* transcription.
pCA1as-439	Similar to pCA1-439.1 but inserted in the opposite orientation to *lacZ* transcription.
pCA2s-439	pMP220 derivative with the entire CRISPR Array 2 of *V. vulnificus* YJ016 cloned as an *Eco*RI DNA fragment in the same orientation as *lacZ* transcription.
pCA2as-439	Similar to pCA2-439.1 but inserted in the opposite orientation to *lacZ* transcription
pCA1s-1DR-439	pMP220 derivative with the first DR and the first spacer of CRISPR Array 1 of *V. vulnificus* YJ06 cloned as an *Eco*RI DNA fragment in the same orientation as *lacZ* transcription.
pCA1as-1DR-439	Similar to pCA1as-1DR-439.1 but inserted in the opposite orientation to *lacZ* transcription
pCA2s-1DR-439	pMP220 derivative with the first DR and the first spacer of CRISPR01 of *V. vulnificus* YJ06 cloned as an *Eco*RI fragment in the same orientation as *lacZ* transcription
pCA2as-1DR-439	Similar to pCA2s-1DR-439 but inserted in the opposite orientation to *lacZ* transcription
pCRISPR-439	pGEM-T Easy derivative containing the leader region, the first DR and the first spacer of CRISPR Array 1 of *V. vulnificus* YJ06.

Table M8. Oligonucleotides used for the construction of plasmids for *in vivo* spacer acquisition assays.

Primer	Sequence (5´-3´)	Description
439O1	GGGAGATCTATCACACACGTTTTCAGGCC	F adaptive operon 439
439O2	GGGAGATCTGAAACTGGAAATGTTTACC	R adaptive operon 439
439-ΔCAf	GGGAGATCTTGCAACGCAAACCAGCACC	F adapative operon 439 Δ-CRISPR Arrays

439-ΔCAr	GGGAGATCTGTCATATAGGGTGTATTACTC	R adapative operon 439 Δ-CRISPR Arrays
439-YAAAf	TCGATATGCCGCCGCTTTTGTTGTTCTG	F RTCas1-439 YAAA
439-YAAAr	CAGAACAACAAAAGCGGCGGCATATCGA	R RTCas1-439 YAAA
439-E517Af	ACGAGGGACAGCAGGCGCGGCCGCA	F RTCas1-439 E517A
439-E517Ar	TGCGGCCGCGCCTGCTGTCCCTCGT	R RTCas1-439 E517A
439-E597Af	GACTTGATGGCAGGCTATCGAC	F RTCas1-439 E597A
439-E597Ar	GTCGATAGCCTGCCATGCAAGTC	R RTCas1-439 E597A
439-ΔCas2Af	GGGCTCGAGAATCACGTAATGTTCACGGCT	F adaptive operon ΔCas2A
439-ΔCas2Ar	GGGCTCGAGTTTTGATTATATTGATTAGAGGTAC	R adaptive operon ΔCas2A
439-ΔCas2Bf	GGGCTCGAGAAGCTGATCAACCAAAGCATG	F adaptive operon ΔCas2B
439-ΔCas2Br	GGGCTCGAGAGAGTAATACACCCTATATGAC	R adaptive operon ΔCas2B
439-tdIf	GGGCATATGCTTTGGTTGATCAGCTTTG	F analysis splicing *td*I
SP6	ATTTAGGTGACACTATAG	R analysis splicing *td*I
CA1-439f	GGGGGATCCATCACACACGTTTTCAGGCC	F CRISPR Array1-439
CA1-439r	GGGAGATCTCCTTTGTTGATAGTTTAAGAAGT	R CRISPR Array1-439
CA2-439f	GGGGGATCCGATAAAAACGACAATGTAACGT	F CRISPR Array2-439
CA2-439r	GGGAGATCTACAAAACGGCTACGAAAACCT	R CRISPR Array2-439
432O1	GGGGGATCCTACGGTATTGAATTGGGTTTTC	F adaptive operon 432
432O2	GGGGGATCCCTTATTCTTTAAAGTACAGCAC	R adaptive operon 432
BsaBI/BamHI/ SalI-5’	TCATCGGATCCG	F adapative operon 432 Δ-CRISPR Array1
BsaBI/BamHI/ SalI-3’	TCGACGGATCCGATGA	R adapative operon 432 Δ-CRISPR Array1
432pRTf	GGGGGATCCGACCAGTGAGGCATTTTTAATC	F adaptive operon 432 (RT-Cas1-Cas2)
432OpΔRTf	GGGGGATCCCCCAATTTGGGAAGTTCAACT	F adaptive operon 432 ΔRT
432OpΔRTf	GGGGGATCCGATGAACATCTGAAA	R adaptive operon 432 ΔRT

* The restriction recognition sites are underlined.

M.4.4. Plasmids for interference assays in Vibrio *strains*

Plasmids for interference assays in Vibrio strains are listed in table M9. The pVSV-105 backbone was used. All plasmids were verified by sequencing. The oligonucleotides used to make plasmids for protein expression and purification are listed in table M10.

Table M9. Plasmids for interference assays in *Vibrio* strains.

Plasmid	Description
pVSV-105	Shuttle vector used as positive control of the conjugation.
pVSV-105-ks	pVSV-105 derivative containing the target sequence of spacer 1 of CRISPR Array 1 from *V. vulnificus* YJ016 in sense orientation.

pVSV-105-sk	pVSV-105 derivative containing the target sequence of spacer 1 of CRISPR Array 1 from *V. vulnificus* YJ016 in antisense orientation.
pVSV-105-OmpC	pVSV-105 derivative containing the constitutive promoter of *ompC* gen from *V. vulnificus* YJ016 cloned as *Sph*I.
pVSV-105-OmpC-ks	pVSV-105-OmpC derivative containing the target sequence of spacer 1 of CRISPR Array 1 from *V. vulnificus* YJ016 in sense orientation.
pVSV-105-OmpC-sk	pVSV-105-OmpC derivative containing the target sequence of spacer 1 of CRISPR Array 1 from *V. vulnificus* YJ016 in antisense orientation.

Table M10. Oligonucleotides used for the construction of plasmids for interference assays

Primer	Sequence (5´-3´)	Description
ks1	CGTCTTACTAATACACCGCACCACTTCTTAAACTATCAACAAAGAAACCGCATG	5’ protospacer1 Sense
ks2	CGGTTTCTTTGTTGATAGTTTAAGAAGTGGTGCGGTGTATTAGTAAGACGGTAC	3’ protospacer1 Sense
sk1	CGTCTTACTAATACACCGCACCACTTCTTAAACTATCAACAAAGAAACCGGTAC	5’ protospacer1 Antisense
sk2	CGGTTTCTTTGTTGATAGTTTAAGAAGTGGTGCGGTGTATTAGTAAGACGCATG	3’ protospacer1 Antisense
OmpCf	GGGACATGTACAATCGAACAGTGTTCATAAG	F promotor OmpC Vv YJ016
OmpCr	GGGGCATGCGTTTAGCTGTCCATAATCTTTTTG	R promotor OmpC Vv YJ016

* The restriction recognition sites are underlined.

M.5. Nucleic acids extraction

M.5.1. Total DNA extraction from Cyanobacteria

To extract total DNA from the Cyanobacteria *Scytonema hofmanni* PCC7110 the culture was directly harvested from the solid medium and was resuspended in 400 µL TE (pH 7.5). Then, a volume equivalent to 150 µL of sterile glass beads (212-300 nm diameter), 20 µL SDS 10% and 450 µL phenol:chloroform was added. Cell lysate was gently mixed in the vortex and was placed on ice 1 minute. Later, it was centrifuged for 15 minutes at 4ºC and 14,000 rpm. The phenol:chloroform step was repeated four times and every time the supernatant was transferred to a new tube. To precipitate the genomic DNA 2-2,5 volumes of absolute ethanol and 1/10 sodium acetate 3 M (pH 5.2) was added and the tube was put for at least 1 hour at -20ºC and then pelleted by centrifugation for 15 minutes at 14,000 rpm 4ºC. The pellet was washed with 1 ml of 70% EtOH for 3-5 minutes at 14,000 rpm 4ºC. The supernatant was removed and the pellet was dried at room temperature in the laminar flux cabin. Finally, it was resuspended in TE or MiliQ water and stored at 4ºC and quantified using a spectrophotometer (NanoDrop® ND-1000).

M.5.2. Plasmid DNA purification

M.5.2.1. Plasmid DNA extraction by magnesium salts

This procedure was used for rapid plasmid extraction from *E. coli* cultures (Studier, 1991). An overnight-grown culture of *E. coli* was collected in 1.5 ml microcentrifuge tubes and centrifuged at 13,000 rpm for 2 minutes. The supernatant was removed, and the pellet resuspended in 100 µL MiliQ water. 0,1 M NaOH, 10 mM EDTA and 2% SDS was added and mixed immediately by vortexing. Tubes were heated for 2 minutes in a boiling-water bath. Therefore, 50 µL of 1 M $MgCl_2$ was added and mixed by vortexing and the tube was placed on ice for 2 minutes. The precipitate was pelleted by centrifuging at 13,000 rpm for 1 minute. Then, 50 µL of 5 M potassium acetate (pH 4.8) was mixed into the supernatant in the same tube by brief vortexing (inverted tube to avoid resuspension of the pellet that contains basically linear chromosomic DNA) and the tube was centrifuged another 5 minutes at 13,000 rpm. The supernatant was removed to a new tube containing 600 µL of 100% cold EtOH, mixed by vortexing and incubated at room temperature for at least 5 minutes. The tube was centrifuged for 5 minutes at 13,000 rpm, the supernatant removed by aspiration, and 200 µL of 70% cold EtOH was added, centrifuged at 13,000 rpm for 5 minutes. The supernatant was removed by aspiration. The pellet was dried at room temperature or at 37ºC in thermoblock for 10-15 minutes. Finally, it was resuspended in 10-30 µL of a 10 µg/µL RNase solution in MiliQ Water. The plasmid DNA was stored at -20ºC.

M.5.2.2. Plasmid DNA purification by commercial kit

Plasmid DNA was extracted using the commercial kit *Illustra plasmidPrep Mini Spin Kit* (GE HEalthcare). The main advantage of the use of this kit versus the previous methods was the obtention of clean DNA and the absence of RNase in the final steps. The procedure was carried out as follows: 1.5-3 ml of culture was harvested in 1.5 ml microcentrifuge tubes by centrifugation at 13,000 rpm for 30 seconds. For the cell lysis step the pellet was resuspended by vortexing in 175 µL of *Lysis buffer type 7*. 175 µL of *Lysis buffer type 8* were added and mixed by inversion until get a clear and viscous solution. The sample was neutralized adding 350µL of *Lysis buffer type 9* and mixing by inversion. After that, it was centrifuged at 13,000 rpm for 4 minutes and the supernatant was transferred to a new tube in which a kit column had previously been placed. After a new centrifugation at 13,000 rpm for 30 seconds to remove the supernatant, bound DNA was washed with 400 µL of *Wash buffer type 1* and a new centrifugation at 13,000 rpm for 1 minute was performed. To elute the DNA, the kit column was transferred to a new tube and 20-50 µL of *Elution buffer type 4* or MiliQ water was added according to the required DNA concentration. It was incubating for 30 seconds at room temperature and centrifuged at 13,000 rpm for another 30 seconds. The sample was stored at -20ºC.

M.5.2.3. Supercoiled plasmid DNA purification

To allow the purification of ultrapure supercoiled DNA with high yields in order to perform spacer acquisition assays *in vitro* (section M.17) the *QIAGEN® Plasmid Midi Kit* was used. Once the culture (25-100 ml) is growth the bacterial cells were harvested by centrifugation at 6000 g for 15 minutes at 4ºC. Then, the bacterial pellet was resuspended in 4 ml of *Buffer P1* (contains RNase A). 4 ml of *Buffer P2* was mixed thoroughly by vigorously inverting the sealed tube 4-6 times and incubated at room temperature for 5 minutes. The *Buffer P3* was then added and mixed immediately by inverting 4-6 times and incubated on ice 15 minutes for enhancing precipitation. This step was followed by centrifugation at

20,000 xg for 30 minutes at 4ºC. In the meantime, a QIAGEN-tip 100 was equilibrated by applying 4 ml of *Buffer QBT* (the column was allowed to empty by gravity flow). The supernatant was transferred to the QIAGEN-tip and allowed it to enter the resin by gravity flow. Later, 10 ml *Buffer QC* was used twice to wash the QIAGEN-tip. The plasmid DNA was eluted with 5 ml of *Buffer QF*. The eluted DNA was precipitated by adding 0.7 volumes (3.5 ml) of isopropanol at room temperature. Immediately, was mixed and centrifuged at 15000 g for 30 minutes at 4ºC and the supernatant was removed. At this point, the DNA pellet was washed with 2 ml of room-temperature 70% ethanol and centrifuged at 15,000 xg for 10 minutes. Then, the supernatant was carefully decanted without disturbing the pellet. The pelleted plasmid DNA was dried 5-10 minutes and resuspended in a suitable volume of TE Buffer or MiliQ water.

M.5.3. Total RNA extraction

This procedure used for isolation of RNA from *E. coli* cultures is based on the method described in Cabanes *et al.*, 2000. The bacterial pellet from a 3 ml culture from *E. coli* (optical density at 600 nm > 1.5) was resuspended and incubated for 10 minutes at 65ºC in 2 ml of prewarned lysis solution (1.4% SDS, 4 mM EDTA and 0.4 centrifuged mg/ml Proteinase K solution). Proteins were removed by adding 150 µL of 5 M NaCl at 4ºC for 10 minutes. Then, the sample was centrifuged at 13,000 rpm for 15 minutes at 4ºC. The supernatant was transferred to a new tube and the nucleic acids were precipitated by adding 1 ml of absolute ethanol for at least 1 hour at -80ºC. The precipitated nucleic acids were at 13,000 for 30 minutes at 4ºC. After removed the ethanol, the pellet was resuspended in 85 µL of MiliQ water and digested with 50 units of RNase-free DNase I (Roche) for 1 hour at 37ºC. The RNA was first extracted with 1 volume (100 µL) of phenol (pH 4.5): chloroform:isoamyl alcohol (25:24:1) and was centrifuged at 13,000 rpm for 10 minutes at 4ºC. The aqueous phase (at the top) was mixed with 1 volume of chloroform:isoamyl alcohol (24:1) and was centrifuged at 13,000 rpm for 5 minutes at 4ºC. Finally, the phase containing the RNA was extracted and precipitated with 3 volumes (600 µl) of ethanol 100% and 75 mM NaOAc (pH 5.2) for at least 1 hour at -80ºC. Then, the sample was centrifuged at 13,000 rpm for 30 minutes at 4ºC and the RNA pellet was washed with 500 µl of ethanol 70% (stored at -20ºC). The RNA pellet was dried at room temperature and resuspended in 20 µl of nuclease-free water. RNA concentration was determined by the use of a spectrophotometer (NanoDrop® ND-1000).

M.6. Cloning and enzymatic manipulation of DNA

M.6.1. DNA digestion with restriction endonucleases

DNA was digested with commercially available restriction endonucleases according to the manufacturer's recommendations (New England Biolabs and Roche). Typically, DNA was digested in the presence of 5 units of enzyme per microgram and 1x Reaction Buffer, in a volume ranged from 10 µl (0.5 µg of plasmid DNA) for analytical digestions to 50 µl (2-5 µg) for preparative digestions and from 1 to several hours depending on the enzyme. In the case of double digestions, generally, these were simultaneously carried out with both enzymes, choosing the optimal buffer for both. As long as the 2 enzymes were incompatible (buffer or temperature), successive digestions were performed, starting with the enzyme with lower ionic strength requirements or lower temperature requirement. After digestion, the DNA was visualized on agarose gel (section M.8.1.) or was further purified (section M.6.3.).

M.6.2. Dephosphorylation of DNA fragments

The 5' end phosphate groups derived from digested DNA vectors were removed by incubation in the presence of the Antarctic Phosphatase (NEB). Typically, 5-10 µg of linearized DNA was added directly to a reaction containing 1x Antarctic Phosphatase Buffer and 2 units in a total volume of 50 µl. This reaction was incubated at 37ºC for 50 minutes. The reaction was terminated by inactivating the enzyme by heating for 10 minutes at 65ºC. Then the DNA was purified as indicated in the following section.

M.6.3. DNA fragments purification

The cleanup of DNA fragments from enzymatic reactions was carried out by the commercial kit *Illustra GFX PCR DNA and Gel Band Purification Kit* (GE Healthcare). Briefly, following manufacturer's indications, 500 µl of *Capture Buffer* was added for every 100 µl of a sample. After mixing the sample is transferred to a *GFX MicroSpin*^{TM} column placed onto a collection tube and centrifuged at 16000 g for 30 seconds and the eluted liquid is discarded. Later, 500 µl of *Washing Buffer* was added and newly the sample was centrifuged. The column was placed onto a new 1.5 ml microcentrifuge tube. Finally, to elute 10-50 µl MiliQ water was added, dried for 1 minute at room temperature and centrifuged another minute at 16000 g. The sample was stored at -20ºC. Around 10-20% of the sample was analyzed in agarose gel to determine its concentration.

This kit was also used for the purification of sliced bands of DNA separated by electrophoresis in agarose gels. In this case, for slice the band of interest the UV transilluminator was used once DNA in the gel is dyed with GelRed (Biotiym, Inc; minimizing the exposure time of the gel to UV). The sliced band was transferred to a 1.5 ml microcentrifuge tube and weighed. For every 10 mg of agarose 10 µl of *Capture Buffer* and the mixture was heated at 60ºC until the agarose got completely melted and following the protocol as is indicated in the previous paragraph.

M.6.4. Ligation of DNA fragments

The reactions in which several of digested DNA fragments of PCR products were carried out typically using a 3:1 insert:vector molar ratio. To perform self-ligations of a vector, the DNA was diluted 10 times with deionized water to decrease its concentration and avoid dimers formed by 2 vectors. In all the reactions, the ligation was carried out with the enzyme T4 DNA ligase (Roche) in the presence of 1x Ligase Buffer and in a final volume of 10-20 µl at 4ºC or room temperature for up to 24 hours.

M.7 Bacterial transformation

M.7.1. Transformation of chemically competent cells

Bacterial cells were made competent according to the rubidium chloride method of Rodríguez & Tait (1983). A 100 ml culture of one of the different *E. coli* strains used in this work (section M.2.) was grown until OD_{600} reached 0.4 (exponential phase). The growth was stopped placing the culture on ice for 15 minutes. Cells were pelleted by centrifugation at 6,000 rpm for 10 minutes at 4ºC. The pellet was resuspended in 30 ml of pre-chilled RF1 solution (100 mM RbCl, 30 mM potassium acetate, 10 mM $CaCl_2$, 50 mM $MnCl_2$, 15% glycerol. pH adjusted to 5.8 with 1 M acetic acid. Filter sterilized and stored at 4ºC). The resuspended cells were incubated on ice for 15 minutes and centrifuged again at 6,000 rpm for 10 minutes at 4ºC. The supernatant was discarded and the cells were resuspended in 4 ml of ice-cold RF2 solution (10 mM RbCl, 5 mM $CaCl_2$, 100 mM MOPS (pH 6.5), 15% glycerol. pH was adjusted to 6.5 with 1 M KOH. Filter and sterilized at 4ºC). Cells were dispensed in aliquots of 100 µl in pre-chilled 1.5 ml microcentrifuge tubes, quick-frozen in liquid nitrogen and stored at -80ºC. The efficiency range of these competent cells is approximately 10^6 cells/µg DNA.

For transforming reaction, competent cells aliquots were placed on ice for 20 minutes. Then, 50-500 ng of plasmid DNA or ligated DNA fragments were added to an aliquot of competent cells and incubated for 20 minutes. Tubes were transferred from ice to 42ºC (water bath or thermoblock), heat-shock for 2 minutes and placed on ice immediately to cool for 5 minutes. 900 µl of LB medium was added and incubated for 1 hour at 37ºC. Finally, all or part of the transformation mix was plated onto LB plates with an appropriate antibiotic and incubated overnight at 37ºC.

M.7.2. Electroporation of electrocompetent **E. coli** *cells*

A 500 ml culture of *E. coli* was grown until the OD_{600} reached 0.4-0.5 and, immediately, the growth was stopped locating culture on ice for 20-30 minutes. Cells were then harvested by centrifugation at 6,000 rpm for 15 minutes and 4ºC. The supernatant was decanted, and the pellet was resuspended in 100 ml of pre-chilled

10% glycerol (stored at 4ºC). Cells were harvested by centrifuging at 6,000 rpm for 15 minutes at 4ºC. The supernatant was removed and the pellet was resuspended in 20 ml of pre-chilled 10% glycerol (stored at 4ºC). Cells were harvested by centrifugation at 6,000 rpm for 15 minutes at 4ºC. The supernatant was carefully aspirated with a sterile Pasteur pipette. Finally, the pellet was resuspended in 2 ml of pre-chilled 10% glycerol (stored at 4ºC). The suspension was split into 50 µl aliquots in sterile 1.5 ml microcentrifuge tubes and snap froze with liquid nitrogen. Cells were stored frozen in the -80ºC freezer. The efficiency range of these competent cells is approximately 10^7 cells/µg DNA.

For electroporation, *Eppendorf 2510* electroporator was used. The electrocompetent cells were thawed on ice for 15 minutes. In sterile conditions, 50-500 ng plasmid DNA or ligation mix (this latter previously dialyzed to remove salts) was added to each aliquot and was incubated on ice for 15 minutes. The dialysis was carried out with bidistilled water using nitrocellulose filters VSWP 0.25 µm (MILLIPORE). The aliquot with the DNA was transferred along the wall of a 0.2 cm electroporation cuvette. The electroporator launched with an electric pulse of 1800 V for 3-5 milliseconds. Immediately, cells were transferred from the cuvette into a 1.5 ml microcentrifuge tube containing 1 ml LB medium without antibiotics. Transformed cells were outgrown by incubating the tubes at 37ºC for 1 hour. Later, all or part of the transformation mix was plated onto LB plates with an appropriate antibiotic and incubated overnight at 37ºC.

M.7.3. Plasmid conjugation between **E. coli** *and* **V. vulnificus** *strains*

Plasmids were mobilized from *E. coli* to *V. vulnificus* YJ016 or R99 strains by filter-mating conjugation (Gulig *et al.*, 2009). To select for *V. vulnificus* and against donor *E. coli* during conjugations, St or TCBS agar containing 10^5 U/ml colistin and appropriate antibiotic was used.

M.8. DNA amplification

M.8.1. Polymerase Chain Reaction (PCR).

Routine PCR reactions (analytical PCR) used to check the presence of certain inserts were carried out with a homemade *Taq* DNA polymerase (Engelke *et al*., 1990). These were performed in a final volume of 25 µl reaction including 10-100 ng of template DNA, 125 µM dNTPs, 200 µM of each specific primer, 2.5 µl of 10x reaction buffer (10 mM Tris-HCl pH 8.3, 50 mM KCl and 2.5 mM $MgCl_2$), 2 U *Taq* DNA polymerase and MiliQ water until complete the 25 µl of the reaction.

PCR reactions which require a low error rate in the final product were carried out either with Accuprime *Taq* DNA Polymerase High-Fidelity (Invitrogen) or Phusion High-Fidelity DNA polymerase (Finnzymes). In a final volume of 50 µl reaction including 10-100 ng of template DNA, 125 µM dNTPs, 500 µM of each specific primer, 1x of the specific reaction buffer of each polymerase, 4 U of the corresponding polymerase and MiliQ water until complete the 50 µl of the reaction.

PCR amplification was performed with an *Eppendorf Mastercycler* thermal cycler. PCR conditions varied depending on the specific denaturing temperature of each polymerase (94ºC for all with the exception of the Phusion which has a denaturing temperature of 98ºC), the annealing primer temperature (typically 60-62ºC), the extension time (from 15 seconds to 2 minutes) and the number of cycles (generally 25, 30 or 35). All the reactions start with an initial step of denaturing for 3-5 minutes at 94 or 98ºC and a final extension step for 4-10 minutes at 72ºC.

Each set of reactions included a negative control (minus DNA template). Amplified products were resolved on agarose gel (section M.9.1.) or purified for later use (section M.6.3.).

M.8.2. Adenilation of PCR products

Occasionally, the amplified PCR product was cloned into pGEM-T Easy vector which presents prominent Thymidine-ends. Then, for the correct cloning it is necessary a PCR product with prominent Adenine-ends. However, some polymerases present 3'-5' exonuclease activity and leave blunt ends after amplification. For this reason, the addition of adenines on both ends of the amplified products when Phusion or Accuprime polymerases are used, is required. With this aim, from 3 to 7 µl of PCR product, previously clean-up or purified, and a mix composing by the following reagents was added: 5 U *Taq* DNA polymerase, 1x specific Taq reaction buffer, 200 µM dATP and MiliQ water (until a final volume of 10 µl). The reaction mix was incubated for 30 minutes at 70ºC and the product is directly used for the ligation reaction (section M.6.4.) with pGEM-T Easy vector.

M.8.3. Overlap-extension PCR (OE-PCR)

This method is a variant of the conventional PCR used to perform point mutations in a sequence. In this thesis, the OE-PCR has been used to generate the point mutations in the RTCas1 protein from *V. vulnificus* YJ016 at the RT (YADD to YAAA) and the Cas1 domain (E517A and E597A point mutations). Firstly, 2 independent PCRs were carried out as described in section M.8.1 with the Phusion High-Fidelity DNA polymerase: one of them with 439a as the forward primer and 439-YAAAr or E517Ar or E597Ar as the reverse primer; and the other PCR with 439-YAAAf or E517Af or E597Af as the forward primer and 439b as the reverse primer (Tables M6 and M8). In both cases, a 1:500 dilution of pMal-Flag-439 was used as DNA template (Table M5). Both PCR products were clean-up (section M.6.3.) and mixed at a 1:1 ratio. Later, a 1:10 dilution of the mix was done and 1 µl was used as DNA template to perform the second round of PCR with the external primers: 439a and 439b. The final products of PCR contained their corresponding point mutation, namely: RTCas1-YAAA, RTCas1-E517A and RTCas1-E597A. Later, these fragments were cloned in the corresponding plasmids to obtain the pMal-Flag-439 derivatives (Table M5) and pAGDt-439 derivatives (Table M7).

This approach was also used to generate a plasmid containing the whole adaptive operon from *V. vulnificus* YJ016 with a 6x His-tagged Cas2B. Thus, two PCRs were carried out: one PCR with 439O1 and HisCas2Br as the forward and reverse primers (primers sequence in tables M6 and M8), respectively, amplified the RTCas1 and Cas2A including the intergenic region between Cas2A and Cas2B. The former primer overlaps with the forward primer of the second PCR, in which primers HisCas2Bf and Cas2B-439-HindIIIr were used to amplify the His-Tagged Cas2B. A 1:500 dilution of pAGDt-439 was used as DNA template for the first PCR and pET16b-Cas2b for the second PCR (tables M5 and M7). The rest of the protocol continues as is described above to obtain a final PCR product used to generate pMal-OpHisCas2B-439 plasmid (table M5).

M.8.4. Reverse Transcription PCR (RT-PCR)

Upon total RNA extraction (section M.5.3), a synthesis of cDNA (complementary DNA) was carried out, consisting of an initial step of annealing of the RNA with random primers, following by the cDNA synthesis through the extension of the oligonucleotides with a commercial reverse transcriptase. With this purpose, the following components were added to a nuclease-free microcentrifuge tube: 1-5 ng of total RNA annealing with 50 ng of random hexamers, 10 mM dNTPs and MiliQ water until 15 µl. The mixture was heated for 5 minutes at 65ºC and quickly chilled on ice. The contents of the tube were collected by brief centrifugation and at this point, 1x First-Strand Buffer, 0.1 M DTT and 40 units of *RNaseOUT*^{TM} were added, mixed gently and incubated for 2 minutes at 42ºC. Finally, 100 units of *SuperScript*^{TM} II RT (Invitrogen) were added, mixed by pipetting and the reaction was carried out for 50 minutes at 42ºC. The reverse transcription was inactivated by heating for 15 minutes at 70ºC.

The last step was the amplification of the freshly synthesized cDNA by a PCR amplification with the *Taq* DNA polymerase purified in the laboratory (section M.8.3.). Only about 10% of the cDNA synthesis reaction was used for the PCR because of higher volumes may not increase amplification and may result in decreased amounts of PCR products. As a negative control, samples without the cDNA synthesis step were included to ensure that it was not remaining DNA contaminating the RNA samples. The result was analyzed in 0.8% agarose gels as described below (section M.9.1.).

M.9. Nucleic acids electrophoresis

M.9.1. Non-denaturing agarose gel electrophoresis

Plasmid DNA, PCR products o restriction DNA fragments were resolved by 0.8-2% agarose gel electrophoresis (*Seakem LE*, Cambrex/Iberlabo) in TAE buffer (Tris-HCl 40 mM, glacial acetic acid 0.1142% (v/v) and EDTA 2 mM). A 6x solution was used as loading buffer (composition: 0,50% Orange G (w/v), EDTA Na_2 0.01 M and and 50% glycerol) was used. Gels were stained with 0.002% GelRed (Biotium, Inc.). The stained gel was visualized under UV light in a *Bio-Rad Gel Doc 2000* transilluminator. Images were captured, cropped, and printed with Quantity One software 4.3.1. (BioRad).

M.9.2. Denaturing agarose gel electrophoresis

Total RNA analysis was resolved by 1.4% agarose gel electrophoresis in MOPS 1x buffer (3-(N-morpholino)-propanesulfonic acid; Buffer 4x composition: 80 mM MOPS, 20 mM sodium acetate and 4 mM EDTA, adjusted to pH 7 with NaOH) and 0.05% (v/v) formaldehyde. 1 µl of total RNA was mixed with 4 µl of MiliQ water and 2 µl solution of 1% GelRed (Biotium, Inc.) in loading dye (1.8% sucrose, 1x TBE (89 mM Tris-HCl pH 8, 89 mM boric acid and 2mM EDTA). The stained gel was visualized under UV light in a *Bio-Rad Gel Doc 2000* transilluminator. Images were captured, cropped, and printed with Quantity One software 4.3.1. (BioRad).

M.9.3. Molecular-weight size markers

The molecular-weight size markers used in this work are the following:

- Marker II: DNA from λ phage digested with *Hind*III (Roche). Composed of 8 fragments: 125, 564, 2027, 2322, 4361, 6557, 9416, 23130 bp.
- Marker III: DNA from λ phage digested with *Hind*III and *Eco*RI (Roche). Composed of 13 fragments: 125, 564, 831, 947, 1375, 1584, 1904, 2027, 3530, 4268, 4973, 5148, 21226 bp.

- Marker Φ29: DNA from Φ29 phage digested with *Hind*III (Universidad Autonoma de Madrid – Servicio de Fermentación). Composed of 14 fragments: 72, 156, 273, 453, 579, 611, 759, 1150, 1331, 1933, 2201, 2498, 2899, 4370 bp.
- Marker pGEMT: DNA from pGEM-T plasmid digested with *Hinf*I and *Eco*RI (Promega). Composed of 15 fragments: 36, 51, 65, 75, 126, 179, 222, 350, 396, 460, 517, 676, 1198, 1605, 2645 bp.

M.10 Sequencing and analysis of plasmid DNA and PCR products

For classic Sanger sequencing (Sanger *et al.*, 1977) the services of the sequencing of DNA/Genomics units from the Instituto de Parasitología y Biomedicina López-Neyra (IPBLN-CSIC) and from the Estación Experimental del Zaidín (EEZ-CSIC) were currently used. For the preparation of the different samples, the specification of each service was followed. The visualization of the chromatograms of the sequences was carried out using *Chromas Lite* v2.0.1 software. The routine analysis of the sequences (searching for restriction targets, comparing several sequences, *in silico* cloning…) was performed with *Clone Manager Professional Suite* v6.00.

M.11. Protein purification and manipulation

M.11.1 Protein purification by amylose beads

pMal-Flag derivatives with the different RTs or RTCas1cloned (table M5) were used to transform *E. coli* strain Rosetta2 (DE3). Single transformed colonies were then grown overnight in LB medium supplemented with ampicillin, chloramphenicol and 0.2% glucose, at 37ºC, with shaking. A flask containing 50 ml LB was inoculated with 1% of the overnight culture, and the bacteria were grown until mid-exponential growth phase at 37ºC (OD_{600} 0.4-0.6), with shaking. Then, IPTG was added to a final concentration of 0.3 mM and the cultures were incubated overnight at 20ºC. Cells were harvested by centrifugation, and the pellet was resuspended in column buffer (CB: 20 mM Tris–HCl (pH 7.5), 200 mM NaCl, 1 mM

EDTA, 1 mM DTT and 1x EDTA- free protease inhibitor (Roche)) at 4ºC. Cells were lysed by three freeze-thaw cycles and subjected to sonication (*Sonifier*® Cell Disrupters, Branson Ultrasonics). The lysate was cleared by centrifugation (16000 g, 15 minutes, 4ºC).

Proteins were purified with a liquid chromatography system using empty Econo-Pac columns (30 ml; BioRad), loaded with 1 ml of amylose beads (NEB Amylose High Flow Resin). The crude protein was loaded into the columns with the amylose beads, incubated for 2 hours at 4ºC (gentle shaking). The removal of the unbound proteins was performed by washing the column five times with 2 ml CB. Then, bound proteins were eluted in CB supplemented with 10 mM maltose. The proteins were concentrated with an Amicon ultracentrifugation filter (Ultracel 30-K) and dialyzed against storage buffer (SB: 10 mM Tris–HCl (pH 7.5), 1 mM DTT, 50% glycerol). The diverse proteins were stable in SB for several months at -20ºC.

M.11.2. Protein purification by Fast Protein Liquid Chromatography (FPLC)

For further purification of MBP-tagged and His-Tagged proteins used in the current thesis (Plasmids described in Table M5) a FPLC was used. Firstly, the plasmid of interest was used to transform *E. coli* strain Rosetta2 (DE3) (chloramphenicol resistant) or BL21 AI, and single transformed colonies were then grown overnight in 10 ml LB medium supplemented with the appropriate antibiotic/s and 0.2% glucose, at 37ºC, with shaking. 2 flasks, each one containing 500 ml LB, were inoculated with 5 ml of the overnight culture, and the bacteria were grown in the shaker at 37º until DO_{600} 0.4 was reached. At this point, 0.3 mM IPTG was added when Rosetta2 strain was used and 1 mM IPTG together with 0.2% L-arabinose for BL21 AI strain and the cultures were incubated overnight at 20ºC. Cells were harvested by centrifugation, and the pellet was resuspended in Buffer A (typically: 20 mM HEPES–KOH (pH 7.4), 500 mM KCl, 0.1% Triton X-100, 2 mM DTT, 0.5 mM phenylmethylsulfonyl fluoride (PMSF), 1x EDTA- free protease inhibitor (Roche) and 10% glycerol) at 4ºC. The solution was incubated for 30 minutes with lysozyme. After lysis by pressure with FRENCH® Press (Thermo Electron), the lysates were cleared by centrifugation (16,000 xg, 30-60 min, 4ºC). In the indicated cases, an additional step was used by adding polyethyleneimine (PEI) to the supernatant on ice with stirring to a final concentration of 0.4%. After 10 min, precipitated nucleic acids were removed by centrifugation (16,000 xg, 30 min, 4ºC).

The next steps were carried out using FPLC with the *ÄKTApurifier* system (GE healthcare). The purification of MBP-tagged proteins was performed with 5 ml MBPTrap^{TM} HP column (GE Healthcare). Upon the proteins was loaded into the column, unbound proteins were removed by washing with at least 5 column volumes (CV) with Buffer B (typically, 20 mM HEPES–KOH (pH 7.4), 500 mM KCl, 2 mM DTT and 10% glycerol). Then, bound protein was eluted with a linear gradient 0-10 mM Maltose (Gradient: 10 minutes at 2 ml/min flow) and 1 ml fractions were collected. Fractions containing the recombinant proteins were identified by SDS-PAGE gels, pooled, and stored at 4ºC or dialyzed against storage buffer (typically, 20 mM HEPES–KOH (pH 7.4), 1 mM DTT and 40% glycerol).

His-tagged proteins were purified using a 5 ml HisTrap^{TM} HP column (GE Healthcare). The protein was then into the column, which was charged with nickel. Unbound protein was washed with at least 5 CV with Buffer B. The next step was the clean-up of unspecific bound protein washing with Buffer B supplemented with 50 mM imidazole. Then, bound protein was eluted with a linear gradient of 50-1000 mM imidazole (Gradient: 10 minutes at 2 ml/min flow). Fractions that contain the protein of interest were analyzed by SDS-PAGE gels, pooled and stored at 4ºC or dialyzed against storage buffer (typically, 20 mM HEPES–KOH (pH 7.4), 1 mM DTT and 40% glycerol).

The purification of co-expressed MBP-tagged and His-tagged proteins, such as MBP-RTCas1-Cas2A-His-Cas2B or MBP-Cas2A-His-Cas2B, was carried out using MBPTrap^{TM} HP column and HisTrap^{TM} HP column in tandem.

For further purification purposes, the purified proteins using one or both the affinity columns described in this section, were dialyzed against Buffer B with low salts (200 mM KCL) and was loaded onto a 1 ml HiTrap^{TM} Heparin HP column (GE healthcare). Unbound protein was washed with at least 5 CV with Buffer B with low salts. Bound protein was eluted using a linear gradient of 200-1500 mM KCl

(Gradient: 8 minutes at 0.5 ml/min flow) and 0.5 ml fractions were collected. Fractions that could contain the protein of interest were analyzed by SDS-PAGE gels and pooled and stored at 4ºC or dialyzed against storage buffer (typically, 20 mM HEPES–KOH (pH 7.4), 1 mM DTT and 40% glycerol).

M.11.3. Size exclusion chromatography (SEC)

Proteins purified as indicated in section M.11.1 and M.11.2 were loaded in gel filtration columns for analytical and further purification purposes. *Superdex^{TM} 75 Increase 10/300 GL* (for proteins with molecular weight (MW)from 3 to 70 kDa), or *Superdex^{TM} 200 Increase 10/300 GL* (MW 10-600 kDa), were used on an *ÄKTApurifier^{TM}* system (GE healthcare). The column was equilibrated with 2 Column Volumes (CV) of Buffer B (20 mM HEPES–KOH (pH 7.4), 500 mM KCl, 2 mM DTT and 10% glycerol) placed on ice. Equilibration was not necessary between runs with the same buffer. Running conditions were set with a flow rate of 0.25-0.35 ml/min and a pressure limit of 3 MPa. 50-500 µl. The protein sample was loaded into the loop with the corresponding volume and the column run was initiated. After 0.2 CV, the protein sample was automatically injected into the column; 0.25 ml fractions were collected in 1.5 ml microcentrifuge tubes. Fractions corresponding with the peaks were analyzed in SDS-PAGE gels and those containing the protein of interest were pooled and stored at 4ºC or dialyzed against storage buffer (typically, 20 mM HEPES–KOH (pH 7.4), 1 mM DTT and 40% glycerol).

For high-resolution gel filtration at a preparative scale, *HiLoad^{TM} Superdex^{TM} 200 prep grade* column (GE Healthcare) was used on an *ÄKTApurifier^{TM}* system. The column was equilibrated as described above. The run conditions were set with a flow rate of 0.75-1 ml/min and a pressure limit of 0.3 MPa. 2-4.5 ml of sample protein was loaded into the loop and the run was initiated. 0.2 CV after the protein sample was automatically injected into the column; 1-1.2 ml fractions were collected in 1.5 ml microcentrifuge tubes. Fractions corresponding with the peaks were analyzed in SDS-PAGE gels and those containing the protein of intereset were pooled and stored at 4ºC or dialyzed against storage buffer (typically, 20 mM HEPES–KOH (pH 7.4), 1 mM DTT and 40% glycerol).

Gel Filtration Standard Kit (Bio-Rad) containing a mixture of molecular weight markers ranging from 1.35 to 670 KDa was used for making the calibration curve of every size exclusion chromatography column used in this work. Once the straight-line pattern was made the molecular weight of the proteins located in each peak could be determined.

M.11.4. Determination of protein concentration

Protein concentration was determined by the Bradford method (Bradford, 1976) using the dye solution of *BioRad*, according to the kit manufacturer's protocol that allows measuring a protein concentration range from 0.125 to 2.5 mg/ml. Firstly, a straight-line pattern was made with bovine serum albumin (BSA), performing 3-5 dilutions from a stock solution of 10 mg/ml, assuming that this protein has a linear range between 0.2 and 0.9 mg/ml. Simultaneously, the test samples were prepared. 800 µl of each dilution was pipetting in a 1.5 ml microcentrifuge tube and 200 µl of Bradford reactive were added in every tube and mixed gently. The samples were incubated at room temperature for at least 5 minutes and then DO_{595} was measured for both, patterns and test samples, in a *Ultrospec II (Pharmacia LKB)* spectrophotometer. Once the straight-line pattern was made the concentration of the test samples could be extrapolated.

M.11.5. Protein Electrophoresis in denaturing SDS-PAGE gels

SDS polyacrylamide gel electrophoresis (SDS-PAGE) was used to protein direct visualization by staining of proteins between 10 and 150 kDa. The gels were made following the protocol described by (Laemmli, 1970), and had a 10-15% acrylamide/bis-acrylamide content. Typically, 0.75 mm gels were prepared in *REAL® Sub-mini 10x10 Dual System (REAL)* and the volumes (in ml) used to have denaturing conditions are described in tables M11 and M12.

Tables M11 and M12. SDS-PAGE gels formulation (amounts in ml)

Resolution Gel (bottom)	Acrylamide %		
	10%	12%	15%
MiliQ water	5,8	5,2	4,14
Acrylamide/bis-acrylamide (40%)	3	3,6	4,5
Tris-HCl 1,5M pH 8,8	3	3	3
SDS 10%	0,12	0,12	0,12
APS 10%	0,12	0,12	0,12
TEMED	0,012	0,012	0,012

Packaging gel (top)	Acrylamyde %
	5 %
MiliQ water	2,225
Acrylamide/bis-acrylamide (40%)	0,375
Tris-HCl 1,5M pH 6,8	0,38
SDS 10%	0,03
APS 10%	0,03
TEMED	0,005

Firstly, the bottom gel is prepared to mix the components from top to bottom order of table M11. Immediately, isopropanol is added, allowing a totally horizontal polymerization in the gel surface. The polymerization takes about 30 minutes. Later, the top gel is prepared in the same way placing the comb where the wells were formed, and it is left polymerizing for 20 minutes.

The protein samples were prepared by adding 5x loading buffer (0.05% bromophenol blue, 0.313 M Tris-HCl pH 6.8, 10% SDS, 0.05 M DTT and 50% glycerol) supplemented with 600 mM β-mercaptoethanol and were immediately denaturing by boiling 3-5 minutes before loading into the gel wells. Gels were run in running buffer 1x (25 mM Tris-HCl, 192 mM glycine and adjusted to pH 8.3) run at 120-150 V until the bromophenol blue reaches the front of the gel. To estimate the molecular weight of the resolved proteins the wide-range marker *Kaleidoscope Molecular Marker* (Bio-Rad; ranging from 10 to 250 kDa) was used.

Proteins visualization were carried out by the next staining methods:

- Coomassie Blue Staining: after electrophoresis run, the gels were immersed in methanol:glacial acetic acid:H_2O (5:1:4 ratio) solution altogether with 0.25% *Coomassie R-250 Brilliant Blue* (Bio-Rad) for 15-30 minutes at room temperature and softly shaking. The fading of the gels were carried out submerging in methanol:glacial acetic acid:H_2O (0.5:1.5:8 ratio) solution until precise visualization of the protein bands.
- Silver Nitrate Staining: the staining was performed following a modification of the protocol described by (Blum *et al.*, 1987). The gels were immersed in 10% EtOH and 0.5% acetic acid for 3 minutes twice for the fixation. Then, fixation solution was removed, and the gels were newly submerged for the staining in 250 ml of 0.2% $AgNO_3$ solution for 30 minutes in darkness and softly shaking. Immediately, 4 quickly washing steps were carried out. Finally, the development solution (composing: 25 mg $NaBH_4$, 1 ml 37% formaldehyde and 250 ml 1.5% NaOH) was added controlling the apparition of the protein bands until a correct visualization. At this moment, the development solution was removed, the gels were clean-up with MiliQ water and, finally, the stop solution (0.75% Na_2CO_3) was added for 3-10 minutes.

In both types of staining, MiliQ water was used for short-time gels conservation at 4ºC. For long-time conservation periods (more than 3 months) 0.2% sodium azide was added to the MiliQ water.

M.11.6. Protein detection by Western Blotting.

10-100 ng of protein sample was loaded into an SDS-PAGE gel. After the run, the protein is transferred from the gel to a PVDF membrane. With this aim, the membrane was activated by immersion in methanol for 10 sec and, subsequently, was equilibrated in transfer buffer (25 mM Tris-HCl pH 8.3, 190 mM glycine and 20% methanol). 14 slices of 3MM Whatman paper (8.5 x 5.5 cm) also were equilibrated in transfer buffer, together with the SDS-PAGE gel. In a semi-dry electrophoretic pre-wetted with water, 7 slices of 3MM Whatman were stacking, then the PVDF membrane was placed, followed by the SDS-PAGE gel and, finally, other 7 3MM Whatman papers complete the stacking. The semi-dry electrophoretic unit was set at 50 mA for 50-70 min.

The PVDF membrane was clean-up with TBST buffer (25 mM Tris-HCl pH 7.5, 150 mM NaCl, 0,1% Tween 20) and blocked for 1 hour in softly shaking with 10 ml of TBST with 0.2 g of AmershamTM ECLTM Prime Blocking Reagent (GE Healthcare). The membrane was quickly rinsed with 10 ml TBST and incubated with the appropriate dilution of primary antibody in 10 ml TBST for 1 hour. Generally, a 1:200,000 dilution of the Anti-His6-Peroxidase (Sigma), a monoclonal antibody directly conjugated to horseradish peroxidase which allows specific and sensitive detection of histidine-tagged proteins, was used. The antibody excess is removed rinsing the membrane with TBST twice, followed by 6 five-minute washing steps with TBST. For signal development, 1.5 ml of AmershamTM ECLTM Prime Western Blotting Detection Reagent (GE Healthcare) was prepared by adding solution A (luminol) and solution B (peroxide) in 1:1 ratio. The membrane was placed in a Gel Documentation System (Bio-Rad) in which the detection reagent was added in darkness and the image was acquired by using the Quantity One v4.6.2 software (Bio-Rad).

M.11.7. Proteolytic cleavage of purified proteins

Upon protein purification, if necessary, MBP-tag or His-tag were removed by proteolytic cleavage. The proteases used in this work together with their recognition sites are listed in table M13.

Table M13. Proteases used in this work.

Proteases	**Cleavage Site**	**Reference**
Xa Factor Protease	Ile-Glu/Asp-Gly-Arg |	New England Biolabs
Pierce^{TM} HRV 3C Protease	Leu-Glu | Val-Leu-Phe-Gln-Gly-Pro	Thermo Scientific

*The precise cleavage position is indicated with “|”

The corresponding enzyme:substrate ratio was tested for all target proteins on a small scale (25 µl reaction) before scale-up. In all the cases, the optimal enzyme:substrate ratio was between 1:25 and 1:100, meaning 1 unit of protease required for cleavage of every 25-100 µg of the target protein. Although each protease uses its proper reaction buffer, both proteases work with high-efficiency in a wide range of buffers, including those with high glycerol content (>20%), in which the efficiency is slightly reduced. Generally, HRV 3C protease was used due to their higher specificity in comparison with Xa Factor protease. The cleavage reaction was performed in the required volume and was incubated overnight at 4ºC for complete cleavage (>90% efficiency). After cleavage, the protease and the different tags (MBP or His-tag) were removed by affinity chromatography (section M.11.2.) or size exclusion chromatography (section M.11.3) columns.

M.12 *In vitro* assays

M.12.1 Exogenous RT activity assay

The assays performed to study exogenous RT activity were carried out as described (Moran *et al.*, 1995; Matsuura *et al.*, 1997). RT activities of the purified proteins were assessed with poly(rA)/oligo(dT)$_{18}$ to obtain a cDNA of high molecular weight. In the reaction, a negative control was included using identical substrate without oligo(dT)$_{18}$. As positive control commercial RTs such as AMV RT (Roche) and SuperScript II (Life Technologies) were used. The poly(rA)/oligo(dT)$_{18}$ substrate was prepared mixing poly(rA) 1 mg/ml (Sigma) and oligo(dT)$_{18}$ 1 mg/ml (GE Healthcare) in a 9:1 ratio and boiling in RT buffer (10 mM KCl, 25 mM MgCl2, 50 mM Tris–HCl (pH 7.5), 5 mM DTT) for 2 min and then placed on ice. In negative controls oligo(dT)$_{18}$ is substituted by miliQ water.

The reaction was carried out by incubating the substrate with 1mM unlabeled deoxythymidine triphosphate (dTTP) and 5 µCi [α-32P]dTTP (800 Ci/mmol; GE healthcare) in 1x RT Buffer. Then, the reaction was initiated by adding the RT or RTCas1 protein (final concentration 0.1–0.5 µM) in a final volume of 10 µl and incubating for 10 min at 37ºC. The reaction was stopped by spotting 8 µl of the reaction mixture onto Whatman DE81 paper. The paper was dried and washed in 250 ml of 2× SSC to eliminate unincorporated labeled dTTP. Radioactivity was quantified with a scintillation counter (Beckman Coulter). All reactions were performed in triplicate and the mean values were obtained.

M.12.2 Protein:protein interaction assays

M.12.2.1 Pull-down assays

pMal-Flag and pET16 derivatives containing the different genes of the *S. hofmanni* PCC 7110 adaptive operon (table M5) were used to transform *E. coli* strain Rosetta2 (DE3). Single transformed colonies were then grown as indicated in section M.11.1. A flask containing 50 ml LB was inoculated with 1 ml of the overnight culture, and the bacteria were grown to exponential growth phase at 37ºC, with shaking. When the culture reached OD_{600} ~ 0.6, 0.3 mM IPTG was added and the cultures were incubated overnight at 20ºC. 2 ml aliquots were harvested by centrifugation, and the pellet was frozen at store at -80ºC.

The required MBP-RT aliquots were unfrozen and resuspended in 1 ml column buffer (section M.11.1) at 4ºC. Cells were lysed by sonication (Sonifier® Cell Disrupters, Branson Ultrasonics). Then, the lysate was cleared by centrifugation (16,000 xg, 15 minutes, 4ºC) and the supernatant is collected into a 1.5 ml microcentrifuge tube. 80 µl of amylose beads (NEB Amylose High Flow Resin) are added to the tube and incubated with the MBP-tagged protein for at least 1 hour at 4ºC. During this step, the His-tagged proteins (Cas1, Cas2, WYL or Ph) are unfrozen and resuspended in 1 ml of Binding Buffer (50 mM Tris-HCl pH 6,8; 100 mM NaCl and 10 mM $CaCl_2$). The sample was sonicated and cleared as indicated above. At this moment, the MBP-tagged protein was three-times washed with 1.5 ml of CB to remove unbound proteins. Then, the supernatant of the just prepared His-tagged proteins was added to tube containing the MBP-RT and the mix was incubated for 2-4 hours at 4ºC. Unbound proteins are removed by washing five times with 1.5 ml of BB. Bound proteins were resuspended with 40 µl of 5x loading buffer (section M.11.5).

Protein interaction was detected by 15% SDS-PAGE gel (section M.11.5) followed by western blotting to detect the presence of the His-tagged proteins (section M.11.6).

M.12.2.2 Flag-tagged protein co-immunoprecipitation

pVSV-105-FlagRTCas1 and pVSV-105-RTCas1 plasmids (Table M5) were used to transformed *E. coli* CC118 λpyr. These plasmids were transferred to *V. vulnificus* YJ016 strain by filter-mating conjugation (section M.7.3). Single transformed colonies were then grown overnight in St medium supplemented with chloramphenicol at 28ºC, with shaking. A flask containing 200 ml St was inoculated with 1ml of the overnight culture, and the bacteria were grown to exponential growth phase at 28ºC, with shaking. When the culture reached an optical density of ~ 0.6, cells were harvested by centrifugation (4,000 rpm 10 min at 4ºC) and the pellet was frozen and stored at -80ºC. The pellet was resuspended in CB (section M.11.1) and incubated with lysozyme (30 min). Cell lysis was carried out with FRENCH® Press (Thermo Electron), and the lysate were cleared by centrifugation (12,000 xg, 10 minutes, 4ºC).

The 8 ml supernatant was added to a 15 ml tube and incubated (gentle shaking) overnight at 4ºC with 40 µl of resin containing anti-Flag antibody (FLAG® Immunoprecipitation Kit; Merck). The resin was prepared following kit manufacturer's protocol. The 8 ml were centrifuged (8,200 xg for 1 min at 4ºC). The unbound protein was removed, and the precipitated resin is transferred to pre-cooled *SigmaPrep^{TM}* spin columns (Merck). The resin is centrifuged (8,200 xg for 1 min at 4ºC) and then, the resin was three-times washed with 500 µl of 1x Wash Buffer (FLAG® Immunoprecipitation Kit; Merck). The bound protein is eluted by the addition of 3 µL of 3XFLAG peptide solution (prepared following kit's manufacturer protocol) together with 100 µL of Wash Buffer. The 103 µL were transferred to the capped *SigmaPrep^{TM}* spin columns with the resin. The resin is resuspended and incubated 30 min at 4ºC. The columns were gently inverted every several minutes. The uncapped column was placed on a pre-cooled 1.5 ml microcentrifuge tube and were centrifuged to eluted Flag-tagged proteins. Proteins that co-purified with Flag-RTcas1 will be identified using SDS-PAGE fractionation followed by peptide mass fingerprinting. RTCas1 without Flag will serve as negative control.

M.12.3 In vitro spacer acquisition assays

M.12.3.1 In vitro spacer integration assays

Spacer integration reactions were performed based on protocol described in Nuñez et al., 2014. Generally, the reaction buffer consists of 20 mM HEPES-KOH, pH 7.4, 100mM KCl, 10 mM MgCl2, 1 mM DTT and 5% glycerol. For reactions with the RTCas1–Cas2A-Cas2B complex, separately or co-purified proteins (sections M.11.1, 2 and 3) were pre-incubated for 30 min at 4 °C. A dsDNA protospacer (AAAAGAAAACCCGGCAGCTTGCCATCCCCACCGT and their complementary; HPLC purified) was incubated with the protein(s) for 10– 15 min at 4 °C, followed by the addition of the target plasmids (pCRISPR-439; table M7) containing the *V. vulnificu*s CRISPR Array 2 (purified as described in section M5.2.3.). The reactions were conducted at 37 °C for 1 h and quenched with 1x loading buffer containing a final concentration of 50 mM EDTA. The products were analyzed on 1% agarose gels in 1x TAE and then staining with GelRed (Biotium, Inc.). Unless stated otherwise, all of the reactions were conducted with 150 nM protein, 100 nM protospacers and 7.5 nM pCRISPR-439 to clearly visualize supercoiled, relaxed or linear plasmid products.

M.12.3.2 Radiolabeled protospacer integration assays.

dsDNA substrate (sequence in the previous section) was radiolabeled using [γ-32P]-ATP (PerkinElmer) by phosphorylation in 5’OH-ends. The reaction was carried out in 10 µl final volume radiolabelling 10-25 pmol probes using 1 µl of

T4 polynucleotide kinase (New England Biolabs), 1-2 µl [γ-^{32}P]-ATP (6,000 mCi/mmol) (PerkinElmer), 1 µl 10x phosphorylation buffer (Tris-HCl 50 mM pH 7.5, 10 mM $MgCl_2$, 5 mM DTT, 0.1 mM EDTA and 0.1 mM spermidine) and filled up to 10 µl with MiliQ water. The reaction was performed 2-3 hours at 37ºC. Then, the volume was raised up to 25 µL and wash clean-up with a *Sephadex Microspin G25* column (New England Biolabs) to eliminate the excess of radiolabeled ATP. Counts per minute (cpm) were quantified with a scintillation counter (Beckman Coulter).

The reactions were carried out in 1x reaction buffer (20 mM HEPES-KOH, pH 7.4, 100mM KCl, 10 mM MgCl2, 1 mM DTT and 5% glycerol). Unless otherwise noted, 150 nM of RTCas1–Cas2A-Cas2B was first incubated with 100 nM protospacers at 4 °C for 10–15 min, followed by the addition of 7.5 nM of pCRISPR-439 plasmid. The reactions were conducted as described in the previous section. After electrophoresis in 1% non-denaturing agarose gels, the DNA was transferred from the gel onto a positively charged nylon transfer membrane (Pall Corporation) using an alkaline transfer system. The nylon membrane was visualized by overnight expositions on *BAS-IP MS2040* screens (Fujifilm) using *Phosphor Imager Personal FX system* (Bio-Rad). The bands were analyzed using the *Quantity One* v4.6.2 software (Bio-Rad). Radiolabeled [γ-32P]-ATP Marker III (section M.9.3) was used as a molecular size marker.

M.12.3.3 CRISPR DNA Cleavage/ligation assays

CRISPR Array 2 from *V. vulnificus* YJ016 (248 bp dsDNA containing the leader sequence, 2 Directs Repeats (DRs) and the first 2 spacers) was internally radiolabeled with [γ-^{32}P]-dTTP (6,000 mCi/mmol) (PerkinElmer). With this aim, a PCR with primers 3222f-439 (sequence) and 3469r-439 (sequence) using the Phusion High-Fidelity was carried out. Every sample was prepared in a final volume of 50 µl: 10 µl 5x Phusion Buffer, 200 µM d(A, C, G)Ps, 30 µM dTTP, 50 µCi [α-^{32}P]-dTTP, 0.5 µM each primer, 4 units Phusion High-Fidelity polymerase, 2 µl 1:50000 dilution 248 dsDNA substrate (previously amplified and purified) and filled up with MiliQ water. The conditions of the PCR were an initial 1-minute denaturing step at 98ºC and then 25 cycles of denaturing at 98ºC for 20 seconds and annealing at 64ºC for 20 seconds, and a final extension step of 4 minutes at 72ºC. Then, the radiolabeled substrate was washed in an S-300 column (GE Healthcare) to remove the radiollabelled nucleotide not incorporated.

The labelled PCR product was loaded into a 6% acrylamide in 1x TBE buffer (Composing: 100 mM Tris-HCl (pH 8.3), 100 mM boric acid and 1 mM EDTA) gel with 0.75 mm thickness to further purify it. The gel was pre-run at 15 V for 10-15 minutes. 1x Loading buffer was added to the substrate which was loaded into the gel. The run was performed for 2 hours at 10 V. After the run, the gel was covered with cling film and was exposed for 10 minutes on *BAS-IP MS2040* screens (Fujifilm) using *Phosphor Imager Personal FX system* (Bio-Rad). The bands were analyzed using *Quantity One* v4.6.2 software (Bio-Rad) and the image was printed to visualize where was the band of interest (248 bp) and sliced it. The sliced band was transferred to a 1.5 ml microcentrifuge tube and radioactivity was quantified with a scintillation counter (Beckman Coulter).

Later, the slice of gel with the radiolabeled substrate was crushed, the tube was frozen on dry ice and 500 µl of extraction buffer (10 mM Tris-HCl (pH 7.5), 1 mM EDTA, 500 mM ammonium acetate) was added to the tube and frozen again. The tube was incubated overnight on a thermoblock at 37ºC. Then, the acrylamide was removed using a system based on glass wool that retains the acrylamide and the rest of the solution was transferred to another tube. The radiolabeled substrate was washed adding 1 volume of ØCIA (chloroform:isoamyl alcohol (24:1); pH 8), vortexed and centrifuged at 13,000 rpm for 10 minutes at 4ºC. The aqueous phase was collected and mixed with 3 volumes of 100% EtOH, 50 mM sodium acetate and 10 µl of lineal acrylamide. The substrate was precipitated (for 1 hour on dry ice or for 2 hours on -80ºC freezer) and was centrifuged at 13,000 rpm for 30 minutes at 4ºC. 100% EtOH was carefully discarded with the pipette and 180 µl 70% EtOH was added to the tube and then was centrifuged at 13,000 rpm for 5-10 minutes at 4ºC. 70% EtOH was carefully discarded with the pipette, and the tube was left at room temperature until complete evaporation of the ethanol. Radioactivity of the dry radiolabeled substrate was quantified with a scintillation counter and the substrate was resuspended in the volume required to have 150,000 cpm/µl.

Separately purified proteins were mixed and incubated for 16 hours at 4 °C to allow complex formation. Reactions were carried out in a 10 µl final volume: 2 µl of each protein individually or the complex (150 nM final), 2 µl radiolabeled substrate (1-10 nM), 2.5 µl protospacer (2.5 µM ssDNA/dsDNA: sequence), 1 µl 10x reaction buffer (20 mM HEPES KOH (pH 7.4), 100 mM KCl, 10 mM $MgCl_2$, 1 mm DTT, 5% glycerol) and MiliQ water. The reactions were incubated at 37ºC for 1 hour and stopped by adding stop solution (1x TE buffer, linear acrylamide and sodium acetate (8:1:1)). The mix was washed with 1 volume of ØCIA (100 µl), vortexed and centrifuged at 13,000 rpm for at least 10 minutes at 4ºC. The aqueous phase was collected and transferred to a tube containing 250 µl 100% EtOH and the tube was frozen at least 2 hours at -80ºC. Then, the samples were centrifuged at 13,000 for 30 minutes at 4ºC. 100% EtOH was carefully discarded with the pipette and 180 µl 70% EtOH was added to the tube and centrifuged at 13,000 rpm for 5-10 minutes at 4ºC. 70% EtOH was carefully discarded with the pipette, and the tube was left at room temperature until the ethanol was completely evapored. Samples were resuspended by vortexing in 10 µl of Loading Buffer (1x TE Buffer and 1x Loading Buffer (97.5% formamide, 10 mM EDTA, 0.3% colorants: xylene cyanol and bromophenol blue). Radiolabeled pGEM marker was used as molecular size marker.

DNA was analyzed in a 6% polyacrylamide 7 M urea gel. The gel was pre-heat at 50 V for 45-60 minutes until getting 50-55ºC. At this point, 5 µl of the samples were loaded in the gel that was run for 2 hours at 50 W. Then, the gel was transferred to a 3MM Whatman paper that were covered with plastic film and were vacuum dried at 80ºC for 60-120 minutes with *Gel Dryer 583* system (Bio-Rad). Radiolabeled bands were visualized by overnight or several days exposition on *BAS-IP MS2040* screens (Fujifilm) using *Phosphor Imager*

Personal FX system (Bio-Rad). Bands were analyzed using the *Quantity One* v4.6.2 software (Bio-Rad).

M.13. *In vivo* assays

M.13.1. β-Galactosidase assay

This assay was used for the determination of promoter activity of the different CRISPR Arrays based on the expression level of the reporter β-galactosidase gene *lacZ* in pCA constructs. β-galactosidase assays were performed as described by Miller (Miller, 1972). Briefly, *E. coli* DH5α strain was transformed with the different pCA constructions. Then, individual transformed colonies were grown at 37°C overnight. Cultures were diluted 1:50 in fresh LB medium and incubated until the cultures reached the log phase (~0.6). Then, cultures were cooled for 20 minutes on ice and bacterial density was recorded by measuring optical density at 600 nm. Then, 100 µl of the cultures were mixed with 900 µl Z buffer (60 mM Na_2HPO_4, 40 mM NaH_2PO_4, 10 mM KCl, 1 mM $MgSO_4$, 50 mM β-Mercaptoethanol and adjusted the pH to 7). Later, 50 µl chloroform and 25 µl 0.1% SDS were added to the sample-buffer mixture and after vortexed for 30 seconds. The samples were incubated at 28°C for 5 minutes and the reaction was initiated by adding 0.2 ml o-nitrophenyl-β-D-galactopyranoside (ONPG) (4mg/ml; Fluka). The reaction proceeds for 10 min at 28°C and was stopped by adding 500 µl of a 1 M Na_2CO_3 solution. Samples were centrifuged 2 min to eliminate cell debris and absorbance was measured at 420 nm. The results are expressed in Miller Units, following formula:

$$\beta - galactosidase\ activity\ (U)\ = \frac{1000\ \times\ DO_{420}}{t \times\ V \times DO_{600}}$$

Where *t* is the reaction time expressed in minutes and *V* is the volume of reaction expressed in ml.

M.13.2 Spacer acquisition assay in vivo

M.13.2.1 Spacer acquisition assay.

E. coli strain HMS174 (DE3) was co-transformed by electroporation with pAGDt plasmids harboring *V. vulnificus* YJ016 RTCas1–Cas2A–Cas2B adaptive operon and derivatives, and pCA plasmids containing the CRISPR Array with only the first DR and the first spacer (Table M7). Individual colonies were cultured overnight at 37ºC in LB medium supplemented with ampicillin and tetracycline. The culture was diluted 1:500 in LB medium, and split into triplicates, which were cultured with the same antibiotics and 0.1 mM IPTG for 14–18 h. The bacterial cells were harvested by centrifugation and plasmids containing CRISPR Arrays were isolated by standard plasmid miniprep procedures to serve as a template for PCR amplification and the preparation of NGS samples.

M.13.2.2 Amplification of CRISPR Arrays and preparation of NGS sample

Leader proximal spacers were amplified by PCR from 3-4 ng of plasmid DNA per µl of PCR mix, with a forward primer binding to the leader sequence of the corresponding CRISPR Array and a reverse primer binding to the first native spacer (Table M14). For each biological replicate, a 25 µl PCR mixture was subjected to the following cycling sequence: 94ºC for 4 min; 30 cycles of 94ºC for 30 s and 62ºC for 30 s. The dominant amplicon contained the first native spacer from the unexpanded CRISPR Array. Electrophoresis was performed in a 2% agarose gel for the excision of gel slices corresponding to a molecular weight of ~ 300 bp (70 bp above the 233-bp band, consistent with the expected size of an amplicon from the expanded CRISPR Array). The slices were purified with the Illustra GFX PCR DNA and Gel Band Purification Kit (GE Healthcare) and eluted in 30 µl of the buffer. We then used 2 µl of the eluted product for the second round of a semi-nested PCR in a 50 µl reaction mixture, with barcoded Illumina sequencing adaptors annealing to the leader region (closer to the first repeat) and to the first native spacer (Table M14), as follows: 94ºC for 4 min; 35

cycles of 94ºC for 30 s and 62ºC for 30 s. Expanded CRISPR Array amplicons were separated from unexpanded arrays by an additional round of purification by electrophoresis in a 2% agarose gel, and the final product was eluted in 10 µl of the buffer. The resulting samples were quantified with *Qubit* (Life Technologies) and analyzed on a *2100 Bioanalyzer* (Agilent Technologies). Libraries were sequenced on an Illumina Miseq at the Genome Sequencing Unit of the IPBLN-CSIC (Granada, Spain).

Table M14. Oligonucleotides used for spacer acquisition assays and Illumina Miseq sequencing.

Primer	Sequence (5´-3´)	Description
sCA2f-439	GAGAGATTTTGAAGCACGCC	F CRISPR02-439 leader
sCA2r-439	ACGGCTACGAAAACCTTGTG	R CRISPR02-439 spacer1
sCA1f-439	TTCCACTGGTTATGGCGTGA	F CRISPR01-439 leader
sCA1-439r	TTGTTGATAGTTTAAGAAGTGGT	R CRISPR01-439 spacer1
1461f	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GacaaccAAGCACAGCACGGTTACAG	Anchor forward CRISPR01 and CRISPR02
1462f	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GttaaccAAGCACAGCACGGTTACAG	Anchor F CRISPR01 and CRISPR02
1463f	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GcgcgtcAAGCACAGCACGGTTACAG	Anchor F CRISPR01 and CRISPR02
1464f	TCGTCGGCAGCGTCAGATGTGTATAAGAGACA GtggcatAAGCACAGCACGGTTACAG	Anchor F CRISPR01 and CRISPR02
1470r	GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC AGctcaggACGGCTACGAAAACCTTGTG	Anchor rreverse CRISPR02
1471r	GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC AGcaacacACGGCTACGAAACCCTTGTG	Anchor R CRISPR02
1472r	GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC AGagaactTTTA AGAAGTGGTGCGGTGT	Anchor R CRISPR01
1473r	GTCTCGTGGGCTCGGAGATGTGTATAAGAGAC AGaggagcTTTAAGAGGTGGTGCGGTGT	Anchor R CRISPR01
sCA1f-432	TTCTTTGAGGGTTACTTTCAGA	F CRISPR02-432 leader
sCA1r-432	CATGATGAGGGCGAAAGCC	R CRISPR02-432 spacer1

*For primers 1461f to 1473r, the 6-nucleotide barcodes are indicated by letters in lower case.

M.13.2.3 Data processing pipeline

FASTQ files were mate-paired with fastq-join (https://github.com/brwnj/fastq-join), with a minimum overlap of 40 nt. They were then converted to FASTA format using the fastq-to-fasta tool with FASTX-Toolkit v0.0.14 (htpp:// http://hannonlab.cshl.edu/fastx_toolkit/) and trimmed with Cutadapt software (Martin *et al.*, 2011). In all samples, ~ 90% of total read pairs were successfully merged, and ~ 80% of the merged read pairs had the correct primer-encoded barcodes located exactly at the ends of the amplicon. Using a custom script written in Python v2.7, spacers were identified, grouped on the basis of unique start and end coordinates (unique spacers), and mapped on the plasmid and genome with Bowtie2.0, with two mismatches allowed. This approach preserves strand information.

M.13.2.4 Construction and splicing efficiency of td intron constructs

The 393-bp intron sequence and its native exons (CTTGGGT/CTACCGT) were inserted in the *Sal*I restriction site of the pAGDt-439 plasmid, just downstream from the adaptation operon, as a *Sal*I/*Xho*I insertion. The intron was introduced at this location due to this region was frequently incorporated as a new spacer in previous experiments (Figure R310). In this context, the splicing product was easy to distinguish from non-spliced transcripts and DNA. *In vivo* splicing efficiency was tested by extracting total RNA (section M.5.3) and further reverse transcription 1.5 µg (SuperScriptII; Life Technologies) with random hexamers in a 20 µl reaction mixture (section M.8.4). Then, 1 µl of cDNA was subjected to PCR amplification in a 25 µl PCR mixture with the Accuprime Polymerase and the Cas2.A-439f and SP6 primers (table M6 and M8, respectively) as described in section M.8.4. The PCR products were analysed by electrophoresis in a 0.8% agarose gel, to check that splicing rates were close to 100%.

M.13.2.5 tdI intron spacer acquisition assay

Spacer acquisition assay described in section 13.2.1 was optimized in pursuit of the detection of the maximum number of spacers after Illumina-Miseq sequencing. After co-transformation with pAGDt-439-*td*I and pCA2s- 1DR, two different sets of experiments corresponding with 80 and 200 individual colonies were selected for the standard spacer acquisition assay. Upon plasmid extraction, individual PCRs were performed, as previously described (section M.13.2.1). The first purification step was carried out by mixing the PCR products in groups of 10 different colonies and then performing the second PCR step. The PCR mixtures were then combined and two additional band purification steps were performed to increase the proportion of expanded arrays. With this method, 50–70% of the reads after Illumina-MiSeq sequencing corresponded to the expanded array, corresponding to >10,000 newly acquired spacers per assay performed with the *td*I construct.

M.13.3 Type III CRISPR-Cas systems interference assay

These experiments were performed in order to demonstrate whether the type III-D system present in *V. vulnificus* YJ016 was functional. *V. vulnificus* strain R99 lacking the III-D CRISPR-Cas system was used as negative control. pVSV-105 derivatives (Table M9) carrying the protospacer matching with the first spacer of CRISPR Array 1 from the YJ016 strain were used to conjugate both *V. vulnificus* strains. The oligonucleotides used to construct pVSV-105 derivatives are listed in Table M10. Once the conjugation was performed, bacteria were plated on St or TCBS agar and grown overnight at 28ºC. Relative conjugation efficiencies were calculated as CFU μg^{-1} DNA of the construct divided by the CFU μg^{-1} DNA of the positive control plasmids. The average value and standard deviation of three times conjugation were shown.

Results

Chapter 1: Phylogenetic relationships of RTs associated with CRISPR-Cas systems

R.1.1 Background

Prokaryotic genomes harbor a plethora of uncharacterized reverse transcriptases. Most prokaryotic RTs are thought to be group-II intron-encoded proteins (IEPs), Retron/retron-like sequences and diversity-generating retroelements (DGRs). However, large-scale genomic surveys and phylogenetic analyses have revealed many other predicted RTs that remain uncharacterized (Kojima and Kanehisa, 2008; Simon and Zimmerly, 2008; Toro and Nisa-Martínez, 2014). In this chapter, an overview of the up-to-date distribution of bacterial and archaeal RTs is presented.

Then, the analysis is focused on a particular lineage of RTs phylogenetically related to those encoded by group II introns, which have been found associated with type III CRISPR-Cas systems, adjacent or fused at the C-terminus to Cas1(Kojima and Kanehisa, 2008; Toro and Nisa-Martínez, 2014). Type III systems target both RNA and transcriptionally active DNA (see section I.4.1.2). Thus, the presence of an RT domain in type III CRISPR-Cas systems could expand immunity against RNA phages or highly transcribed regions. Moreover, it is also shown that type VI CRISPR-Cas systems, which target RNA (see section I.4.2.3), have also recruit an adaptive unit with an RT-Cas1 fusion present, suggesting acquisition of spacers from RNA molecules by type VI systems. Nevertheless, the current knowledge of RT-containing CRISPR-Cas systems remain limited.

To explain how RT first arrived at a CRISPR-Cas context a parsimonious evolutionary scenario known as the "single point of origin" model has been proposed, possibly through a random group II intron retrotransposition event (Silas *et al.*, 2017a). However, the aim of this chapter is to provide novel insights on the origin and evolutionary relationships of RTs functionally linked to CRISPR-Cas systems. Here, a "multiple origins" model is suggested based on strong evidence that RTs have been recruited several times during evolution by these adaptive immune systems.

R.1.2 Distribution of RTs between bacteria and archaea

In order to perform the most exhaustive phylogenetic analysis of bacterial reverse transcriptases, completed and draft bacterial genome were analyzed in search of sequences annotated either as “RNA-directed DNA polymerases” (145,379 sequences) or “Reverse Transcriptases” (52,684 sequences). The dataset was enlarged as explained in Figure M.2. The final dataset included 198,760 predicted RT proteins to perform phylogenetic analysis and address the distribution of main RT groups in bacterial phyla. This large dataset was processed by selecting the RT domain (RT0-7) of at least 200 amino acids and unique sequences by multistep clustering at 85% sequence identity to remove closer relatives (see section M.1.1.1). This procedure yielded a final dataset of 9,141 predicted RT-sequences, expanding from hundreds to thousands the number of non-redundant RTs sequences (Toro and Nisa-Martinez 2014).

In Archaea, the record of RTs is reduced compared with bacteria because of the lower number of complete genomes available in databases (i.e. more than 3 x 10^5 bacterial genomes versus almost 5 x 10^3 archaea in the PATRIC database). Here, the largest dataset of archaeal RTs until the date was also compiled using Uniprot and Patric databases. A total of 1,491 putative archaeal RTs sequences were processed as explained in section M.1.1.2., yielding a final dataset of 411 representative archaeal RT sequences.

The bacterial phylogenetic tree was constructed from a multiple sequence alignment (MSA) of the final dataset of 9,141 non-redundant RTs (Figure R1.1A). The phylogenetic tree supports that the majority of RTs belong to the 3 main groups: group-II introns, the largest group with 47% of total RTs, Retron/retron-like sequences (25%), and DGRs (12%) (Figure R1.2A). The remaining 16% clustered into distinct groups including RTs previously reported being linked to CRISPR-Cas systems, Group 2-like (G2L), Abi-like or UG (Unknown) groups. Overall, these data are in concordance with those reported previously (Kojima and Kanehisa 2008; Simon and Zimmerly 2008; Toro and Nisa-Martínez 2014). In contrast, the archaeal RT phylogenetic tree shows a different pattern (Figure R1.1B).

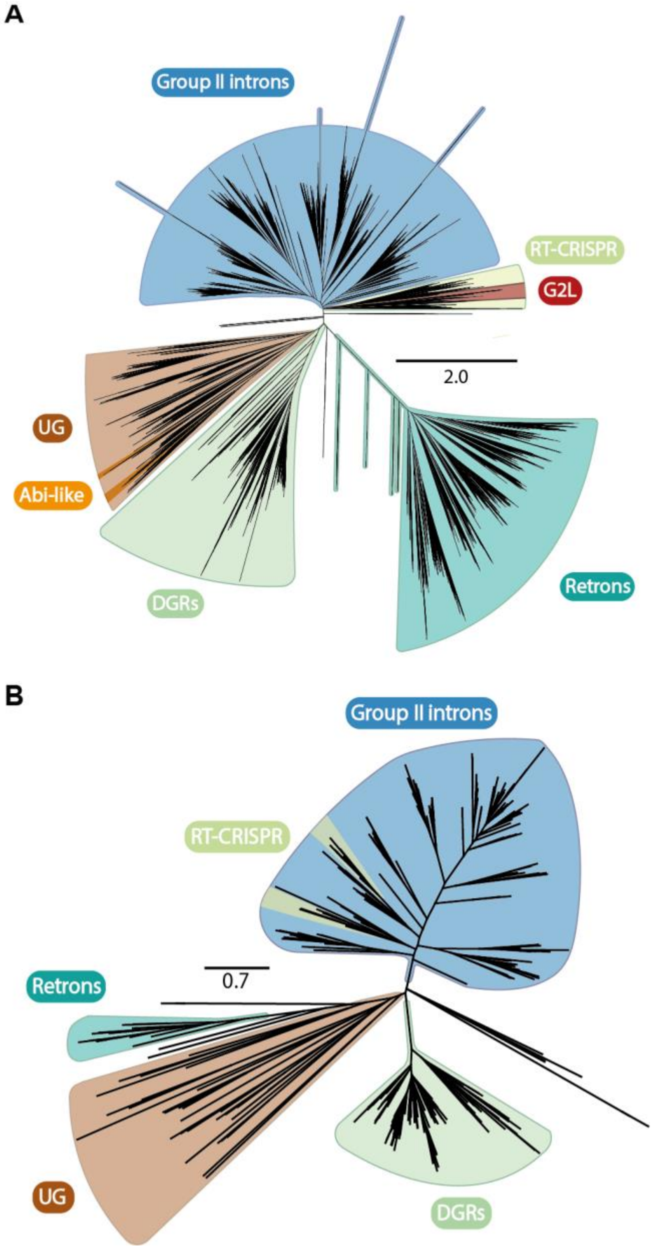


Figure R1.1 Phylogeny of prokaryotic RTs. Phylogeny of bacterial **(A)** and archaeal **(B)** RTs. The unrooted trees were constructed from an alignment of 9,141 and 411 unique predicted protein sequences, respectively, using FastTree program. The branches corresponding to Group II introns, RT associated with CRISPR-Cas systems (RT-CRISPR), Group II-like (G2L), Retrons, Diversity-generating retroelements (DGRs), Abi-like RTs and RTs from unknown groups (UG) are indicated and shaded with different colors.

The archaeal RT phylogeny reveals that the highest proportion of DGRs stands out (30% of total archaeal RTs) in contrast to what occurs in bacteria (with only 12%) (Figure R1.2A). The opposite occurs with retrons which only represent 4% of total Archaea RTs and since they belong to different phylogenetic clades might be a consequence of a recent event of horizontal gene transfer (HGT) from bacterial retrons. Besides, no entries of the G2L and Abi-like groups have been found between archaeal RTs. The remaining RT groups maintain a similar percentage between archaea and bacteria, being group II intron the most abundant (52%), which likely have proliferated even though originally were also acquired from bacteria (Toro *et al.*, 2003).

RTs are unevenly spread among the different prokaryotic phyla. Looking deeper into each bacterial phylum, predictably group II introns are the most prevalent RT group in most bacterial phyla. Nevertheless, retrons are the predominant type in phylum Proteobacteria, whereas DGRs constitute about 80% of total RTs in CPR phylum; the greatest bias observed towards a specific RT group in bacteria (Figure R1.2B). With regards to the distribution of RT groups within the different archaeal phyla/groups, group II introns are highly dominant in Euryarchaeota, TACK group and Asgard group (55-75% of RTs). However, a different scenario appears in DPANN group (Diapherotrites, Parvarchaeota, Aenigmarchaeota, Nanoarchaeota and Nanohaloarchaeota) and unculture/unclassified archaea where DGRs represent about 80% of total RTs (Figure R1.2C).

In Bacteria, Flavobacteria-Chlorobi-Bacteroidetes group (FCB), Candidate Phyla Radiation (CPR) and Cyanobacteria show high rates of non-redundant RTs relative to the number of sequenced genomes of these phyla. By contrast, phylum Actinobacteria present a low RT diversity despite being the third phylum with more sequenced genomes (Figure R1.3A). As long as the RT distribution is considered based on the RT groups, Proteobacteria is the predominant phylum in most RT groups, whereas Firmicutes account for a large proportion of group II introns and Abi-like RTs. Similarly, CPR group contains about 40% of all singular DGRs (Figure R1.3B). These data support the idea that certain bacterial phyla have recruited specific groups of RTs to improve their responses to the ecological conditions of their natural environments.

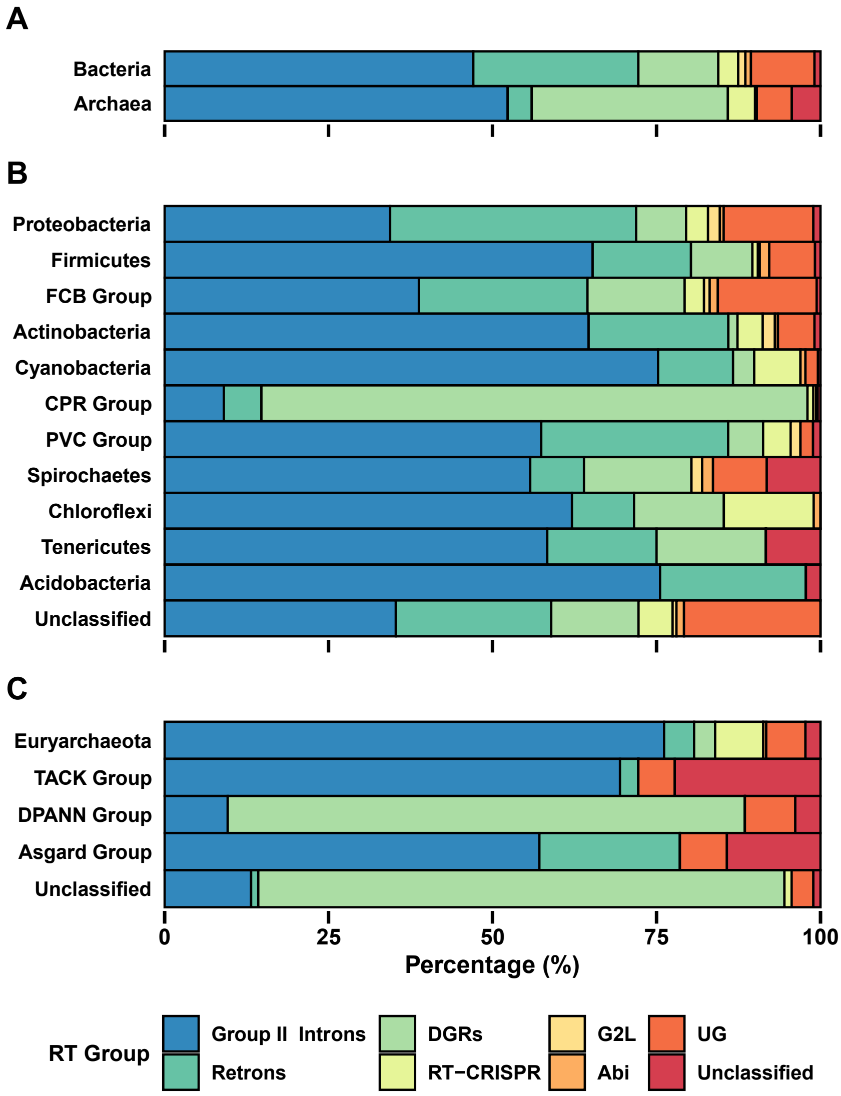


Figure R1.2 Distribution of prokaryotic RTs per phylum. (A) RT groups in bacterial and archaeal genomes. **(B)** RT groups in main bacterial phyla. **(C)** RT groups in main archaeal phyla. Charts showing the proportions of RTs proteins that correspond with Group II Introns, Retrons, DGRs, RTs associated with CRISPR-Cas systems (RT-CRISPR), Group II-like RTs (G2L), RT-based Abi systems (Abi), Unknown Groups (UG) and Unclassified RTs.

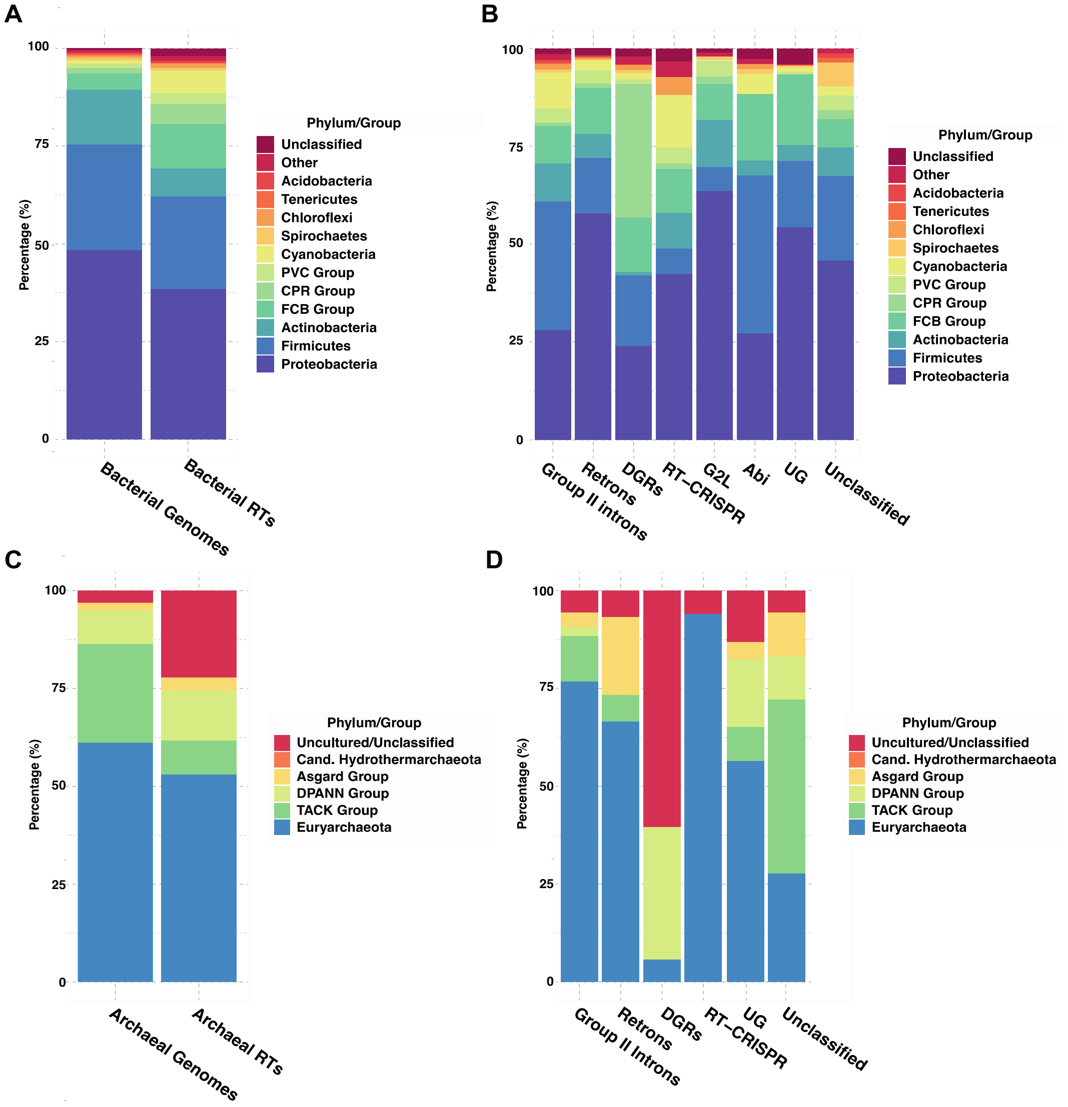


Figure R1.3. Distribution of every RT group in different prokaryotic phyla. (A) Non-redundant RTs per bacterial phylum relative to the number of sequenced bacterial genomes per phylum. Charts showing the proportions of main bacterial phyla/groups corresponding to total bacterial genomes sequenced (left bar) and the proportion of non-redundant RTs per bacterial phylum/group (right bar). **(B)** Distribution per RT lineage in the main bacterial phyla. Proportion of the RT groups in the main bacterial phyla/groups. **(C)** Non-redundant RTs relative to the number of sequenced archaeal genomes. Charts showing the proportions of main archaeal phyla/groups corresponding to total archaeal genomes sequenced (left bar) and the proportion of non-redundant RTs per archaeal phylum/group (right bar). **(D)** Distribution per RT lineage in the main archaeal phyla. Proportion of the RT groups in the main archaeal phyla/groups

As in Bacteria, in most archaea phyla/groups there is a correlation between the total genomes sequenced from a phyla/group and the total number of RTs found in the group concerned, except for the TACK group (Thaumarchaeota, Aigararchaeota, Crenarchaeota and Korarchaeota phyla), which accounts for 25% of total archaeal genomes but only harbor 7% of non-redundant RT sequences (Figure R1.3C). Regarding the dominance of phyla for every RT group, Euryarchaota constituted the most extended phylum, although it is worth to highlight that 90% of all DGRs were found in DPANN group and unclassified archaea (Figure R1.3D).

R.1.3 Reverse Transcriptases associated with CRISPR-Cas systems

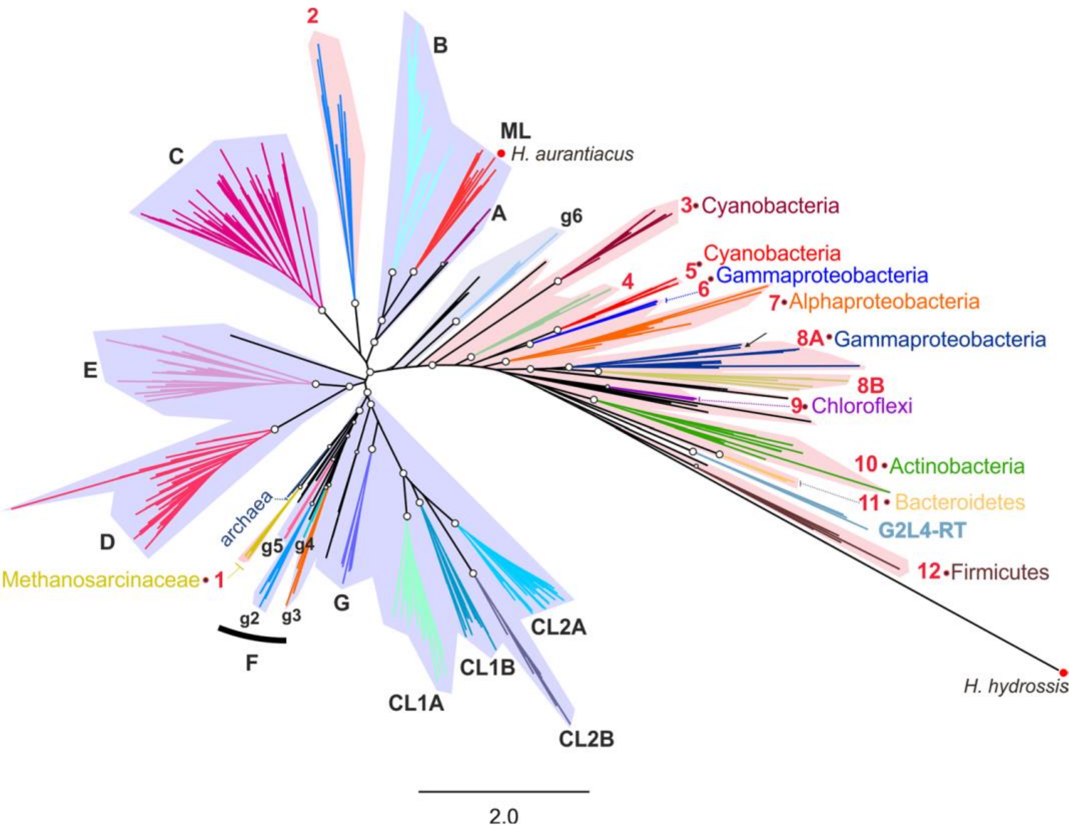
R.1.3.1. Phylogeny of Reverse Transcriptases associated with CRISPR-Cas systems

To generate the phylogeny of RTs associated with CRISPR-Cas systems a phylogenetic tree was built using a dataset of 537 sequences as described in section M1.1.1. This dataset encompass sequence from different origins: RTs associated with CRISPR-Cas systems described in previous studies, protein sequences carrying both Cas1 and RT domains present in the NCBI database, RTs associated with CRISPR-*cas* loci found in complete archaeal genome, together with group II intron RTs and more closely related RT-like sequences. The phylogenetic clustering of the protein dataset identified 12 major clades of reverse transcriptases associated with CRISPR-Cas systems. In contrast to the extensive horizontal transfer observed for CRISPR-cas systems, most of the clades identified of RTs associated with these adaptive systems were limited to particular phyla suggesting host-dependent functioning (Figure R1.4; Appendix A1).

Thus, RTs associated to CRISPR-*cas* loci were found in a broad range of bacterial phyla including Cyanobacteria (clades 3 and 5), (alpha and gamma) Proteobacteria (clades 6 to 8), Chloroflexi (clade 9), Actinobacteria (clade 10), Bacteroidetes (clade 11) and Firmicutes (clade 11). Representatives of various bacterial phyla were found in clade 2 (Planctomycetia, Bacteroidetes, Delta and

Epsilon Proteobacteria) and clade 4 (Planctomycetia, Chlorobi, Gamma and Delta Proteobacteria, no-rank phyla).

In archaea, all RT sequences related to CRISPR-Cas systems are clustered in clade 1. It is worth to note that these archaeal RTs are present in only two genera from the Methanosarcineacea family *(Methanosarcina and Methanomethylovorans)* and they share a common ancestor with two RT sequences from unlcultured archaea that do not appear to be associated with CRISPR-*cas* loci. Furthermore, clade 1 seem to have branch off from a common node to class F group II intron RTs (Figure R1.4). Moreover, these RT remained adjacent to *cas1* genes but not fused. Taken together, these findings suggest that RT-CRISPR module may have been acquired by *Methanosarcina* spp, by a horizontal gene transfer (HGT) event from bacteria, in which this association is more widespread.



(Figure Legend in the next page)

Figure R1.4. Unrooted phylogenetic tree encompassing the diversity of RTs associated with CRISPR-Cas systems. The tree includes 118 RT sequences associated with CRISPR-Cas systems and 419 closely related RT sequences (Methods). Note that the RTs associated with CRISPR-*cas* locus (highlighted with red dots) from *Herpetosiphon aurantiacus* (GI: 159898445) and *Haliscomenobacter hydrossis* (GI: 332661943) correspond to a group II intron and a retron/retron-like RT, respectively. The arrow indicates the position of the *M. mediterranea* (MMB-1) RT. Group II intron classes and varieties are highlighted in color and their names are indicated in black. All group II introns RTs are shadowed in light purple. The RT clades associated with CRISPR-*cas* loci are highlighted in color and their names are indicated in red. All RTs associated with CRISPR-Cas systems are shadowed in light pink. Open circles at the nodes indicate that the node concerned has a FastTree support value ≥0.92. The phyla restricted to particular RT-CRISPR clades are indicated.

On the other hand, although some RTs are associated with partial CRISPR-Cas systems (37%), formed only by the adaptive unit and the CRISPR Array, most RTs are presented in a complete CRISPR-*cas* locus (63%) and they were always found to be associated with all subtypes of type III systems (III-A to D), with predominance of Csm complexes, which are found in subtypes III-A and III-D.

R.1.3.2 Co-evolution of RT and Cas1 domains

To understand the phylogenetic relationships between RTs and Cas1 in CRISPR-Cas modules a phylogenetic tree using 148 unique Cas1 sequences was constructed (Figure R1.5; section M1.2.2). The Cas1 phylogeny essentially matched that of the associated nearby or fused RT for the majority of the clades. The Cas1 phylogenetic tree reveals two main lineages of Cas1 proteins associated with RTs (fused or separate), one of which contained most of the sequences including the archaeal Cas1, providing further support for its acquisition from bacteria. The other lineage contained Cas1 proteins from the CRISPR-Cas systems identified in the clade 3, which is restricted to cyanobacteria. In addition to the RTCas1 fusion, this particular clade is characterized by a more distant *cas1* gene present within the CRISPR-Cas. However, the phylogeny of the two *cas1* genes of these systems suggests that they have a distinct evolutionary origin (Appendix A2).

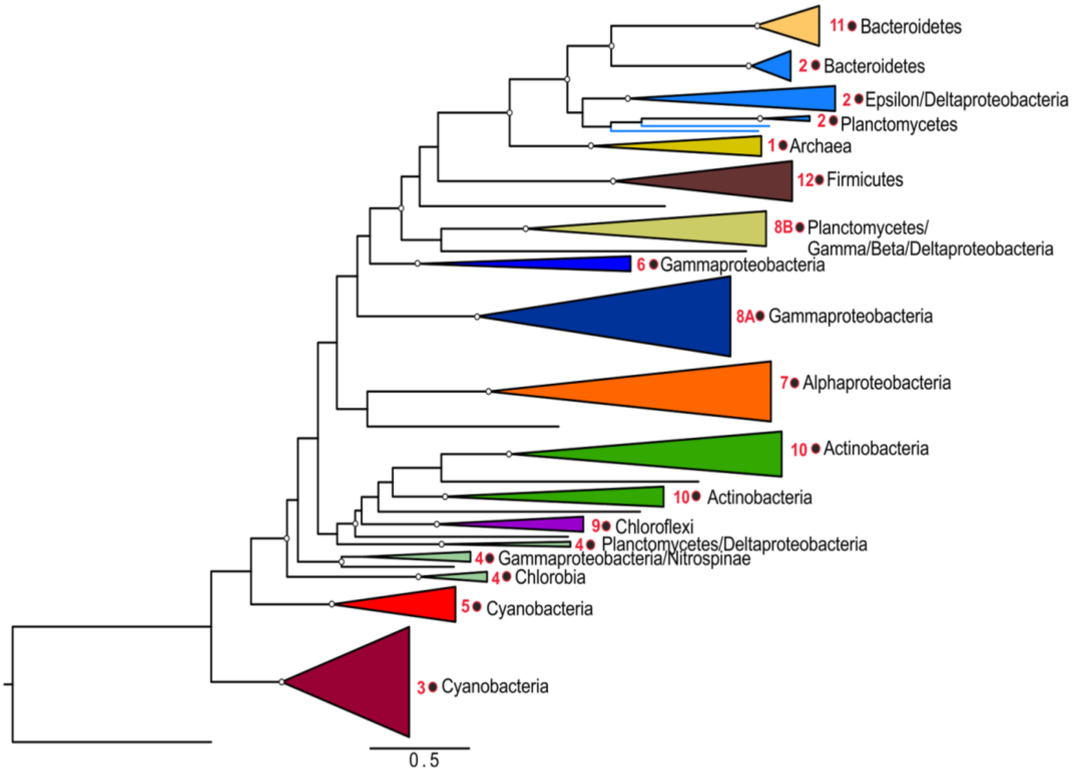


Figure R1.5 Phylogenetic tree of Cas1 associated with RTs. The phylogenetic reconstruction was performed with a total of 148 Cas1 proteins. The identified clades were named and colored in agreement with the RT-associated clade shown in Figure R1.4. FastTree support values ≥ 0.92 are indicated at the nodes. The Cas1 protein (unknown subtype) from *Arthrospira platensis* (GI:479129297; Makarova *et al.*, 2015) was used as an outgroup.

The comparison between the RT and Cas1 tree suggests extensive co-evolution. In addition, RT and Cas1 domains appear to have a common evolutionary history that may have led to co-adaptation through a direct relationship between these two proteins (e.g., a physical interaction) within particular protein complexes, which need to be further investigated. Nevertheless, the Cas1 sequences corresponding to RT-CRISPR clades 2, 4, 8 and 10 are polyphyletic, and some of these clades (clades 2 and 4) are subdivided according to particular phyla. This implies that either the association of the RT and Cas1 is a more recent evolutionary event or that it has occurred multiple times in these CRISPR-Cas modules.

R.1.3.3 “Single point” versus “various origins” of RTs associated with CRISPR-Cas systems

Overall, the findings explained above suggest that recruitment of RT proteins by type III CRISPR-Cas systems occurs several times during evolution. However, a parallel study based on a phylogenetic tree using 134 RTs associated with CRISPR-Cas systems suggested a parsimonious evolutionary scenario known as “single point origin” model (Silas et al., 2017a). This model proposes that upon a random group II intron retrotransposition even, RT domain arrives in a genomic context close to a CRISPR-Cas system. Then, the C-termini of the RT ended up fusing with the Cas1 domain. Finally, in some cases a Cas6 domains was acquired at the N-termini of the RT (Silas *et al.*, 2017a). To shed light on this contradictory issue about the different hypothesis used to explain origin and evolutionary relationships of the RTs associated with CRISPR-Cas systems, an integrated dataset merging RT sequences from both studies was used to perform a new phylogenetic analysis as described in the following section.

Firstly, a comparison of our dataset with the 134 RTs used to build the phylogenetic tree in Silas *et al.*, (2017a) was carried out. The different compilation methods used to retrieve RT- like sequences associated with CRISPR-cas systems may be the explanation for the several differences observed between the two datasets. In Silas *et al.*, (2017a), the RT alone sequences clustered on branches 7 and 10 were not included in our previous analysis. Conversely, the RTs grouped into clade 11 and fused to Cas6 domain at their N-termini and to Cas1 domain at the C-termini present in our study, were not included in Silas et al., (2017a). However, their analysis only contains a few members of clades 6 and 10, which not form consistent branches and appear as unclassified entries.

On the other hand, our dataset lacks all the RT sequences that were clustered in the branch 10 of Silas *et al.*, (2017a) analysis, which were placed at the base of the phylogenetic tree. All RTs from this branch were harbor in *Streptomyces* and *Streptococcus* genomes and, in contrast with the rest of RTs, they were associated with type I-E CRISPR-Cas systems. To test if this particular group of RTs represent a functional association between RT and CRISPR-Cas systems, a phylogenetic analysis with a dataset of 742 RT sequences with representatives from all RT groups and these putative RTs were carried out. This analysis reveals that the RT sequence from *Streptococcus oralis* SK10 was found embedded between uncharacterized UG2 group of RTs, whereas those from *Streptomyces clavuligerus* ATCC27064, *Streptomyces lydicus* A02 and *Streptomyces* MUSC164 formed a phylogenetically distant clade to group II intron classes. Moreover, these last three RTs had a large number of substitutions per site (FastTree: 3.2), the canonical YADD sequence in domain 5 was replaced by WGDD sequence, and they also lacked the conserved domains RT0 and 7. Therefore, the *Streptococcus* and *Streptomyces* RT-like sequences described above were not considered for the following analysis, as they appear to be distantly related to group II intron-encoded RTs.

Finally, to further investigate the origin and evolutionary relationships of RTs associated with type III CRISPR-Cas systems, 21 RT sequences identified in branches 7, 8 and 9 from Silas *et al.*, (2017a) analysis which were absent in our previous study were added to 537 sequences of our dataset (section M1.1.1; Appendix A1). Upon the alignment of at least 250 positions of the integrated 558 RTs sequences the phylogenetic tree was built using FastTree as described in section M1.2. This analysis provides additional support to our previous phylogeny, as a total of 13 clades of RTs associated with CRISPR-Cas systems were identified: the 12 clades of our previous study and confirmed new one clade, hereafter referred to as clade 13, which corresponds to the branch 7 identified in Silas *et al.*, (2017a). On the other hand, most of the RT sequences clustered together on branch 9 in Silas et al., (2017a), were grouped together with sequences of clades 2A (*Caminibacter mediatlanticus* TB-2), and 2B (*Bacteroides fragilis* str. 3988T B14 and *Bacteroides barnesiae* DSM 18169). However, our analysis was unable to confirm the minor branch 8 identified by Silas *et al.*, (2017a), even though the RT sequence from *Roseburia inulivorans* DSM 16841 RT clustered within our clade 12 (phylum Firmicutes). The rest of sequences from this branch remained unclassified and might have a different origin from the others. The clades identified in this study (clades 1–13) together with the equivalent branches identified by Silas et al., (2017a) are shown in Figure R1.6.

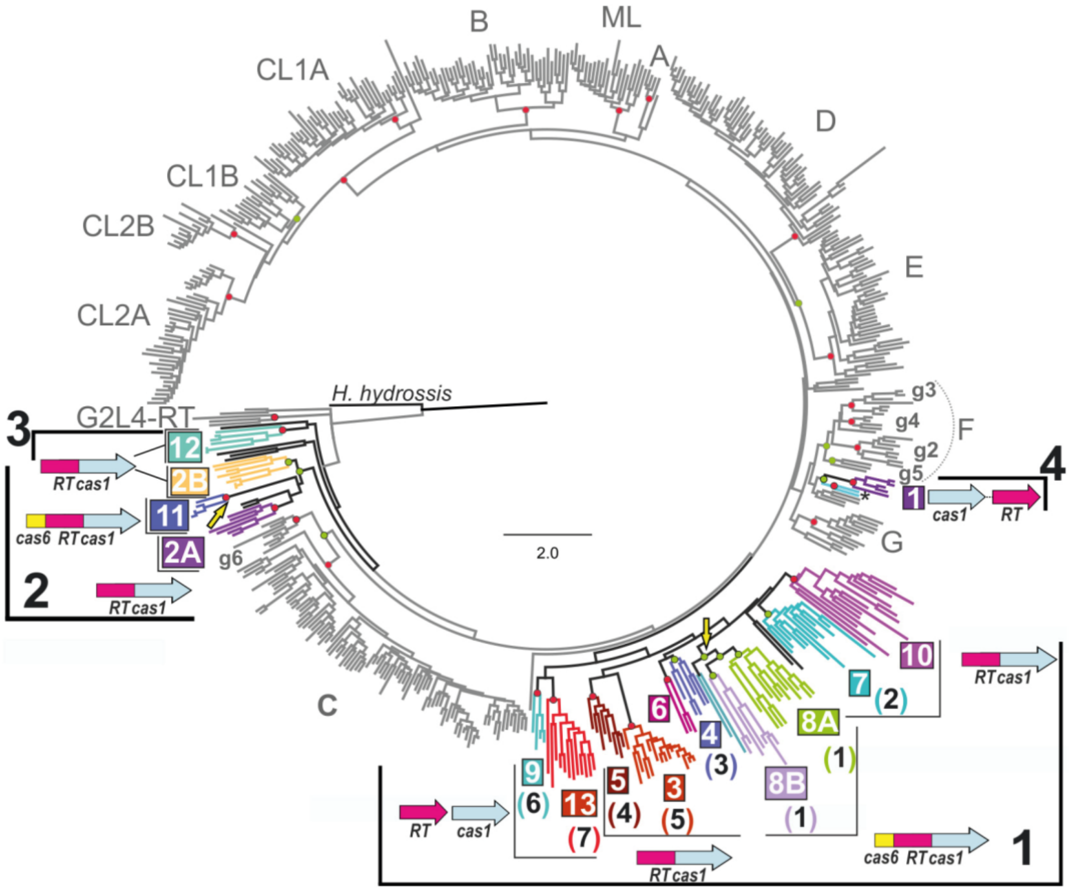


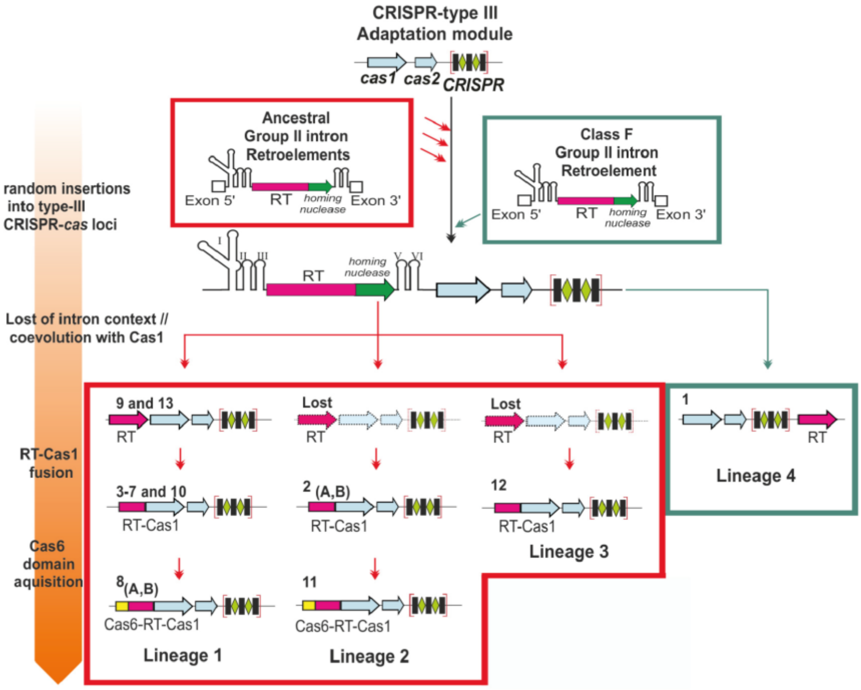
Figure R1.6 Phylogeny of RTs associated with CRISPR-Cas systems. The tree was inferred with FastTree, from an alignment containing 558 RT sequences (537 from the previous analysis and 21 from Silas *et al.*, (2017a), including 138 RT sequences closely related to group II intron-encoded RTs associated with CRISPR-Cas systems (see section M1.1). Group II intron classes and varieties together with G2L4 group are indicated in gray. Branches corresponding to RTs associated with CRISPR-*cas* loci are in shown in black and the clades and host bacterial phyla or archaeal family are indicated by highlighting in color. The possible bacterial lineages 1, 2 and 3, and archaeal lineage 4 are indicated. Two singular RT sequences related to archaeal RTs of clade 1 are indicatated in blue below clade. The most common domain or gene organization for each clade is indicated. Independent genes are shown with distinct arrows, while fused genes are displayed as single arrows with multiple colors. The name of the clade (colored and boxed) and its correspondence with the branches (number in brackets, respectively) identified by Silas *et al.*, (2017a) are shown. The two independent Cas6 domain acquisitions are indicated in yellow. Circles at the nodes indicate that the node concerned has a FastTree local support value ≥ 0.9 and either a standard non-parametric bootstrap value ≥ 70% (Red) or SH-aLRT and Ufboot values ≥ 80% (Green) in the corresponding phylogenetic analyses. Outgroup: retron/retron-like RT from *Haliscomenobacter hydrossis* (GI: 332661943).

As a consequence of the incorporation of additional RT sequences into our dataset, the sequences clustered in clade 2 (RTCas1 fusion) branching off from a well-supported node (FastTree local value 0.98, and SH-aLRT and Ufboot values of 94%) common to sequences formerly clustered in clade 11 (Cas6RTCas1 fusions), hereafter termed as clade 2/11. Within clade 2, RT sequences may be split into two subclades, 2A and 2B, Subclade 2A contain representatives belonging to the new proposed phylum Epsilonbacteraeota, comprising the Epsiolonproteobacteria and Desulfurellales (Waite *et al.*, 2017), whereas all sequences of the subclade 2B belong to the phylum Bacteroidetes.

Interestingly, clade 11, closely related with clade 2 in the current analysis, contain RTs that belong to Bacteroidetes phylum. The relationship between both clades was not unexpected, because Cas1 proteins associated with RTs from clade 2 and clade 11 were found to might have a common origin (Figure R1.5). Therefore, this analysis supports the idea that in an ancestor of some *Porphyromonas* species (clade 11) a Cas6 domain was acquired in a preexisting RTCas1 fusion within clade 2B (phylum Bacteroidetes). This acquisition of the Cas6 domain in the clade 11 occur independently of the acquisition event of the same domain in clade 8 (Figure R1.6). This finding provides support to the “various origin” model of RTs associated with CRISPR-Cas systems. On the other hand, clade 12, which appear at the base of the phylogenetic tree in this analysis, would form an independent lineage from the rest of the clades (Figure R1.6; Appendix A3).

The phylogenetic tree built with the FastTree program provided support for an inner node (local value of 0.96) that comprises from clade 3 to clade 10, including the clade 13. All these clades would be in agreement with the “single point of origin” model: first would occur the acquisition of an RT (clades 9 and 13), followed by a fusion of the RT to Cas1 (clades 3–7 and 10) and, finally, the acquisition of a Cas6 domain by a RTCas1 fusion (clade 8). Nevertheless, the same phylogenetic analysis carried out with other ML methods (SH-aLRT and Ufboot values of 53.5 and 67%, respectively) did not support this parsimonious evolutionary hypothesis, due to splitting of the clades containing RTs alone (clades 9 and 13) from the rest of the clades in the non-parametric bootstrapping analysis (Appendix A3). Thus, to solve this issue further phylogenetic studies with more RT sequences associated with CRISPR-Cas modules are required to confirm or refuse these evolutionary relationships.

Due to the long diversification time and sequence saturation with mutations of RTs sequences associated with CRISPR-Cas systems, most of the internal nodes lacked reliable support, hindering inferences about the evolutionary relationships between group II intron RTs and the different RT-CRISPR clades. However, the topology of the several trees built during these analysis (Figure R1.6 and Appendix A3) reflected a closer relationship of most RT-CRISPR clades with the exception of clade 1 with class C group II introns. Furthermore, RTs related to CRISPR-Cas modules could be subdivided into three bacterial lineages: a major lineage comprising clades 3–10 and 13 (lineage 1), the clade 2/11 (lineage 2), and clade 12 (lineage 3). Additionally, the archaeal RTs from clade 1 would constitute the lineage 4. All these lineages could indicate different acquisition events of an RT by CRISPR-Cas systems, possibly from ancestral group II introns (Figure R1.7).



(Figure Legend in the next page)

Figure R1.7 Contribution of group II introns to the origin of the reverse transcriptases associated with type III CRISPR-Cas systems. The scheme represents the several steps (left side) in the association between a group II intron-encoded RT to type III CRISPR-Cas systems. Different CRISPR-*cas* loci were independently invaded by ancestral group II introns or by a more recent Class F group II intron (red and green arrows, respectively). The intron RNA was lost, and the remaining RT co-evolved with the adjacent Cas1 protein, a process that may have occurred independently four times during evolution (Lineages 1–4). Subsequently the RT and Cas1 were fused (Lineages 1–3), and later a Cas6 domain was acquired independently twice (Lineages 1 and 2). The clades are indicated together with their gene organization. Dashed gene loci indicate RT arrangements lost or not yet identified.

However, to overcome all the problems raised in this analysis and to provide more insight into the evolutionary origins of this particular group of RTs associated with CRISPR-Cas systems an extensive study was carried out as described in the following section.

R.1.3.4 Multiple origins of RTs linked to CRISPR-Cas systems

A large survey of CRISPR-encoded RTs was performed using the dataset of 9,141 non-redundant RT sequences analyzed in the section R1.2 of this thesis. This dataset represents the up-to-date landscape of prokaryotic reverse transcriptases. Thus, it results very suitable to dig into the evolutionary history of RTs linked to CRISPR-Cas systems. To avoid manual searching and improve the results, a computational pipeline was designed for identifying RTs sequences with a CRISPR-*cas* loci in their genomic neighborhood (Toro *et al.*, 2019a). Briefly, all genes located 30 kb upstream and downstream from the RT gene were analyzed seeking the presence of an adaptation, an effector or a whole CRISPR-Cas module.

With this computational pipeline a total of 280 non-redundant RTs associated with CRISPR-Cas systems were detected (Appendix A1). The analysis confirms the 13 clades previously described expanding the number of RTs and RTCas1 sequences clustered in these clades. Interestingly, the analysis also revealed the existence of other two additional clades (clade 14 and 15) with a distinct origin from group II introns. Thus, RT sequences from clade 14 have evolved from Retron/retron-like

RTs and those from clade 15 have their origin in the Abi-P2 group. Furthermore, although sequences from clade 14 are also associated with type III CRISPR-Cas systems as the rest of the clades, RTs from clade 15 were found to be associated with type I-C systems (Figures R1.8). A summary of the distribution of the RTs linked to CRISPR-*cas* loci is shown in Table R1. An example of the genomic architecture of each clade is provided in Figure R1.9.

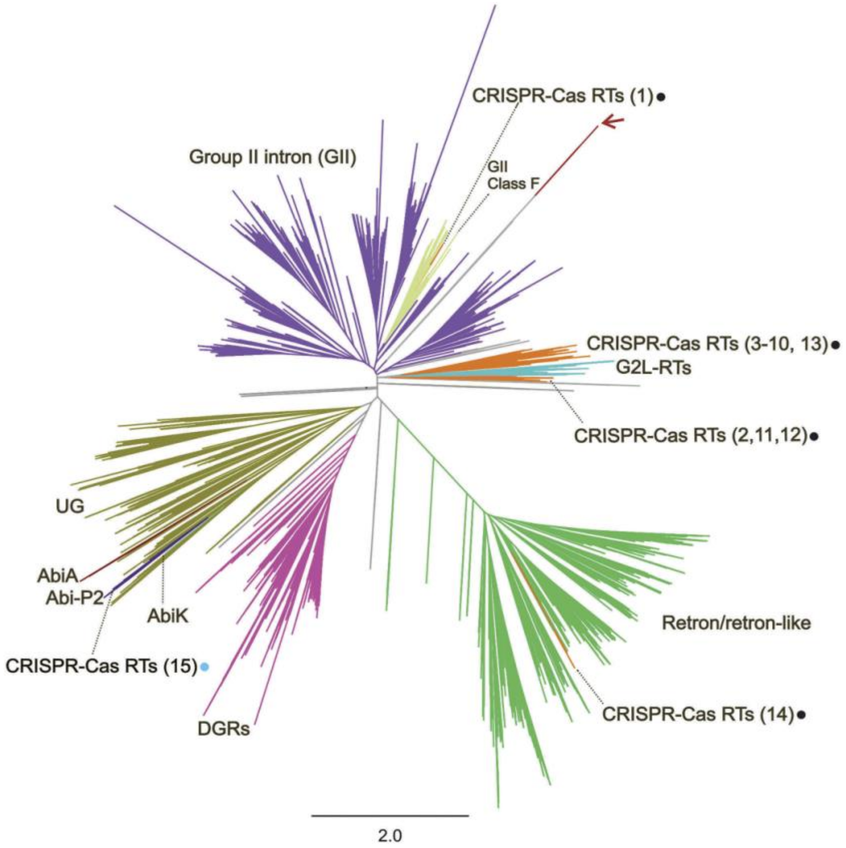


Figure R1.8 Multiples origins of RTs associated with CRISPR-Cas systems. The unrooted tree was constructed from an alignment of 9,141 unique predicted RT protein sequences obtained with the FastTree program as described in section M.1.1. The branches corresponding to group-II introns (GII), GII class F, Retron/retron-like, DGRs, CRISPR-Cas, G2L, Abi and UG RTs are indicated and highlighted with distinct colors. The 15 clades of RTs linked to CRISPR-Cas systems are shown with their specific number indicated in brackets. The CRISPR-Cas system type associated with these RTs are indicated by dots: type III (black) or type I-C (blue). The red arrow indicates the branches corresponding to the putative RTs linked to type I-E CRISPR-Cas systems described by Silas *et al*. (2017a). A relevant subtree is shown in Figure R1.10.

Table R1 Distribution of RTs associated with CRISPR-Cas systems. (a)Number of representative RTs described in this study (corresponding to Appendix A1. (b)Number of records with partial or unknown effector complex associated. (c)Three of the records Cas6RTCas1 fusion gene without a recognizable Cas1 domain.

	Clade	Taxonomic adscription	*Records*[a]	CRISPR-Cas System Type						T
				Type III			Type I		RT	RT Ca
				Cmr (B-C)	Csm (A-D)	(−)[b]	I-C	I-E		
Prokaryotes	**All**	**Taxonomic adscription**	**280**	**73**	**105**	**94**	**2**	**6**	**97**	**13**
Archaea	**1**	Euryarcheota (Methanosarcinaceae)	4	1	3	-			4	0
Bacteria	**2**	Planctomycetia, Bacteroidetes, (Delta, Epsilon) Proteobacteria	33	9	14	10	-	-	3	30
	3	Cyanobacteria	18	9	5	4	-	-	-	18
	4	Planctomycetia, Chlorobi, Iginavibacteria,(Beta, Delta, Gamma) Proteobacteria	22	6	7	9	-	-	8	13
	5	Cyanobacteria	20	10	2	8	-	-	13	7
	6	Gammaproteobacteria	6	-	4	2	-	-	-	6
	7	(Alpha, Delta) Proteobacteria, Deinococci	30	5	16	9	-	-	5	25
	8	Planctomycetia, (Beta, Gamma, Delta, Zeta) Proteobacteria, Nitrospira	40	13	20	7	-	-	-	7
	9	Chloroflexi, Anaerolineae	8	-	-	8	-	-	7	1
	10	Actinobacteria	23	4	14	5	-	-	2	21
	11	Bacteroidetes, Saprospiria, Planctomycetia, (Delta, Epsilon) Proteobacteria	4	2	-	2	-	-	-	-
	12	Firmicutes	21	2	11	8	-	-	7	8
	13	(Gamma, Delta) Proteobacteria, Nitorspirae, Anaerolineae	26	5	6	15	-	-	25	-
	14	Flavobacteria, Cytophagia, Saprospiiria	10	6	2	2	-	-	10	-
	15	Gammaproteobacteria	2	-	-	-	2	-	2	-
	Silas-Branch10	Actinobacteria	6	-	-	-	-	6	6	-
	Unkonwn	Proteobacteria, Fusobacteria, Anaerolineae; Synergistia, Planctomycetia	7	1	1	5	-	-	4	2

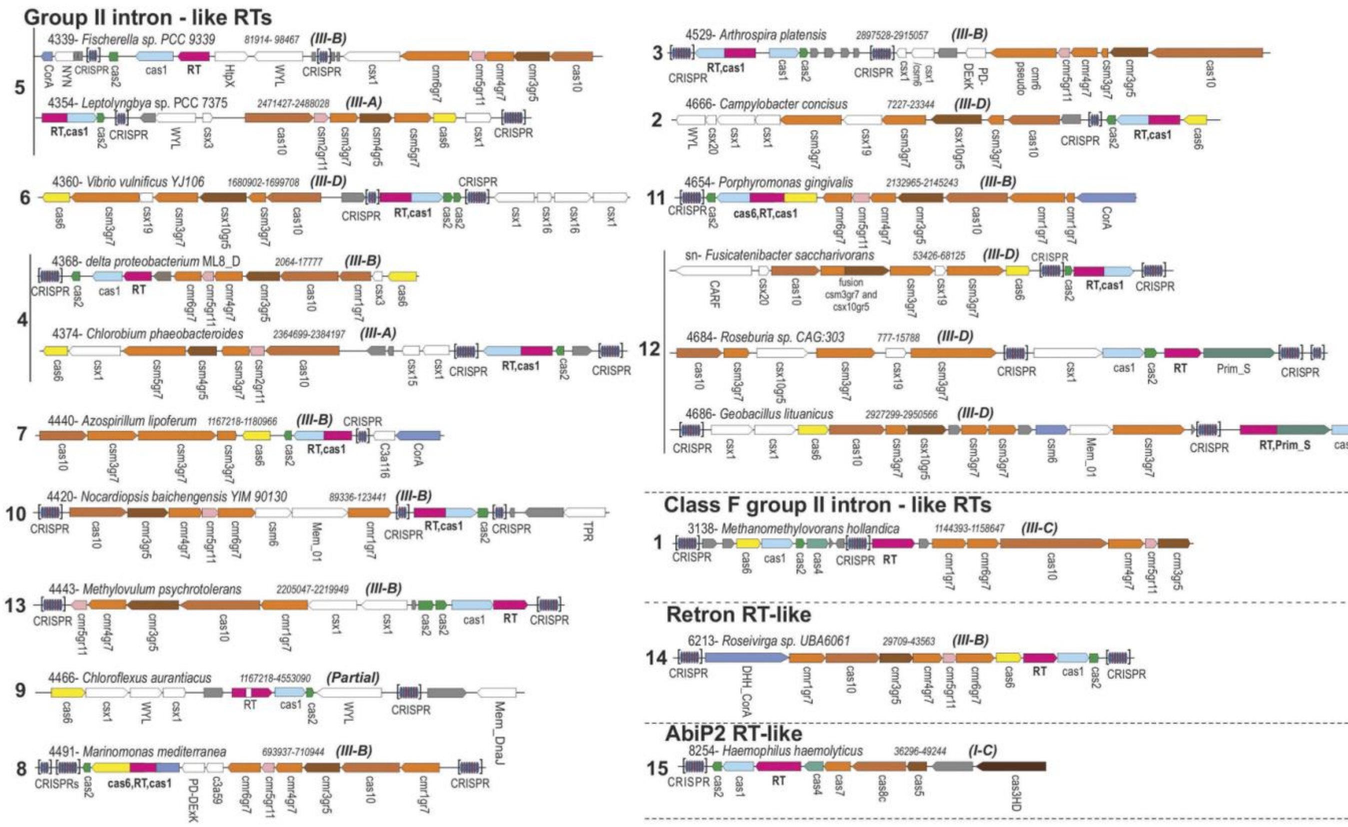
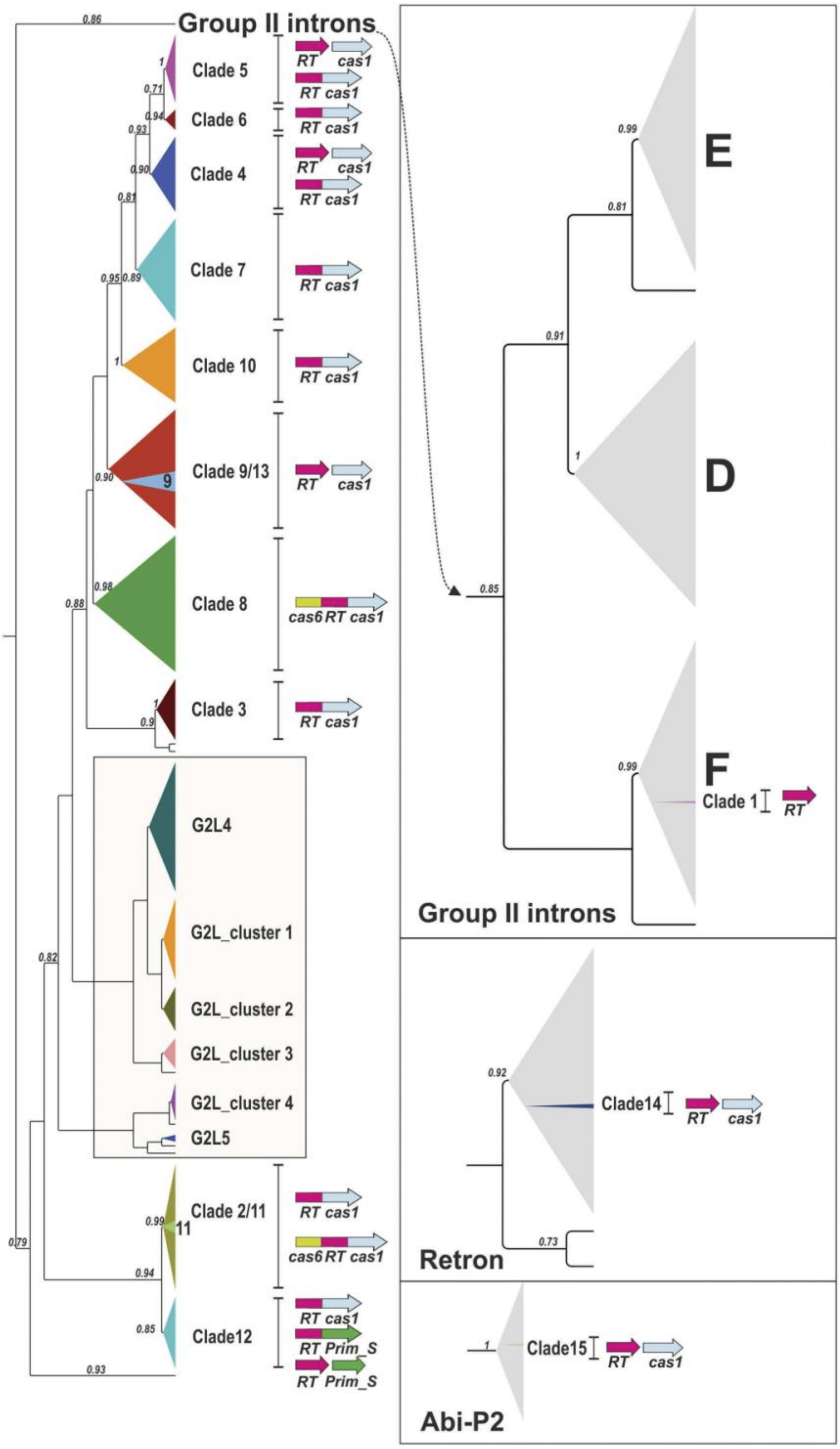

Group II intron - like RTs
5
4339- Fischerella sp. PCC 9339 81914- 98467 (III-B)
CorA NYN CRISPR cas2 cas1 RT HtpX WYL CRISPR csx1 cmr6gr7 cmr5gr11 cmr4gr7 cmr3gr5 cas10
4354- Leptolyngbya sp. PCC 7375 2471427-2488028 (III-A)
RT,cas1 cas2 CRISPR WYL csx3 cas10 csm2gr11 csm3gr7 csm4gr5 csm5gr7 cas6 csx1 CRISPR
6
4360- Vibrio vulnificus YJ106 1680902-1699708 (III-D)
cas6 csm3gr7 csx19 csm3gr7 csx10gr5 csm3gr7 cas10 CRISPR RT,cas1 cas2 cas2 CRISPR csx1 csx16 csx16 csx1
4
4368- delta proteobacterium ML8_D 2064-17777 (III-B)
CRISPR cas2 cas1 RT cmr6gr7 cmr5gr11 cmr4gr7 cmr3gr5 cas10 cmr1gr7 csx3 cas6
4374- Chlorobium phaeobacteroides 2364699-2384197 (III-A)
cas6 csx1 csm5gr7 csm4gr5 csm3gr7 csm2gr11 cas10 csx15 csx1 CRISPR RT,cas1 cas2 CRISPR
7
4440- Azospirillum lipoferum 1167218-1180966 (III-B)
cas10 csm3gr7 csm3gr7 csm3gr7 cas6 cas2 RT,cas1 CRISPR C3a116 CorA
10
4420- Nocardiopsis baichengensis YIM 90130 89336-123441 (III-B)
CRISPR cas10 cmr3gr5 cmr4gr7 cmr5gr11 cmr6gr7 csm6 Mem_01 cmr1gr7 CRISPR RT,cas1 cas2 CRISPR TPR
13
4443- Methylovulum psychrotolerans 2205047-2219949 (III-B)
CRISPR cmr5gr11 cmr4gr7 cmr3gr5 cas10 cmr1gr7 csx1 csx1 cas2 cas2 cas1 RT CRISPR
9
4466- Chloroflexus aurantiacus 1167218-4553090 (Partial)
cas6 csx1 WYL csx1 RT cas1 cas2 WYL CRISPR Mem_DnaJ
8
4491- Marinomonas mediterranea 693937-710944 (III-B)
CRISPRs cas2 cas6,RT,cas1 PD-DExK c3a59 cmr6gr7 cmr5gr11 cmr4gr7 cmr3gr5 cas10 cmr1gr7 CRISPR
3
4529- Arthrospira platensis 2897528-2915057 (III-B)
CRISPR RT,cas1 cas1 cas2 CRISPR csx1 csx1 /csm6 PD-DExK cmr6 pseudo cmr5gr11 cmr4gr7 csm3gr7 cmr3gr5 cas10
2
4666- Campylobacter concisus 7227-23344 (III-D)
WYL csx20 csx1 csx1 csm3gr7 csx19 csm3gr7 csx10gr5 csm3gr7 cas10 CRISPR cas2 RT,cas1 cas6
11
4654- Porphyromonas gingivalis 2132965-2145243 (III-B)
CRISPR cas2 cas6,RT,cas1 cmr6gr7 cmr5gr11 cmr4gr7 cmr3gr5 cas10 cmr1gr7 cmr1gr7 CorA
12
sn- Fusicatenibacter saccharivorans 53426-68125 (III-D)
CARF csx20 cas10 fusion csm3gr7 and csx10gr5 csm3gr7 csx19 csm3gr7 cas6 CRISPR cas2 RT,cas1 CRISPR
4684- Roseburia sp. CAG:303 777-15788 (III-D)
cas10 csm3gr7 csx10gr5 csm3gr7 csx19 csm3gr7 CRISPR csx1 cas1 cas2 RT Prim_S CRISPR
4686- Geobacillus lituanicus 2927299-2950566 (III-D)
CRISPR csx1 csx1 cas6 cas10 csm3gr7 csx10gr5 csm3gr7 csm3gr7 csm6 Mem_01 csm3gr7 CRISPR RT,Prim_S cas
Class F group II intron - like RTs
1
3138- Methanomethylovorans hollandica 1144393-1158647 (III-C)
CRISPR cas6 cas1 cas2 cas4 CRISPR RT cmr1gr7 cmr6gr7 cas10 cmr4gr7 cmr5gr11 crm3gr5
Retron RT-like
14
6213- Roseivirga sp. UBA6061 29709-43563 (III-B)
CRISPR DHH_CorA cmr1gr7 cas10 cmr3gr5 cmr4gr7 cmr5gr11 cmr6gr7 cas6 RT cas1 cas2 CRISPR
AbiP2 RT-like
15
8254- Haemophilus haemolyticus 36296-49244 (I-C)
CRISPR cas2 cas1 RT cas4 cas7 cas8c cas5 cas3HD

Figure R1.9 Architectures of genomic loci for the representative subtypes of CRISPR-Cas systems associated with RTs. Grou RTs (ancient, clades 2–13; and recent, clade 1), Retron RT-like (clade 14) and Abi-P2 RT-like (clade 15). For each locus, the node n respective nucleotide coordinates and CRISPR-Cas system subtype are indicated. Genes are shown roughly to scale; CRISPR Arra in brackets and are not to scale. Homologous genes are color-coded, with the exception of most of the ancillary genes, which are s unknown proteins are shown in grey.

Independently of the major effort carried out with this computational analysis, the predicted RTs from *Streptomyces* species reported to be associated with type I-E CRISPR-Cas systems (Silas *et al.*, 2017a) present an uncertain origin as described in the preceding section. These protein sequences form a distinct long branch in the phylogenetic tree of RTs (Figure R1.8). The current phylogeny show that this reduced group of RTs share a common node with group II intron RT sequences. However, this fact is not consistently supported by the phylogenetic analyses performed in the previous section. Thus, their position in the tree could be consequence of a long branch attraction (LBA) phenomenon due to the large number of substitutions per site (2.4) presented in the sequence of these RTs.

The most recent phylogenetic analysis reveals that the 13 previously reported clades of RTs associated with CRISPR-Cas systems that may have evolved from a retrotransposition event from an ancestral group-II intron are polyphyletic and they can be subdivided into three major lineages. The main difference respect to analysis performed in the previous section, in which four different lineages were detected (Figure R1.7), lies in the current study clades 2, 11 and 12 form a single lineage instead of the two reported above. The archaeal RTs branching within class F introns and the group comprising clades 3–10 and 13 keep on forming independent lineages. Curiously, the RT sequences present in the bacterial clades (2 to 13) were found to be closely related to other uncharacterized group of RT sequences, known as G2L (Group II-like). As RTs linked to CRISPR-Cas systems, G2L-RTs also lack the characteristic ribozyme RNA structure which represent one of the main features of group II introns. Moreover, G2L-RTs also lacks CRISPR-Cas modules in their vicinity. The novel G2L-RT sequences detected in the current analysis cluster with other RTs members of the previously described G2L4 and G2L5 groups (Kojima and Kanehisa, 2008; Simon and Zimmerly, 2008; Toro and Nisa-Mártinez, 2014; Zimmerly and Wu, 2015) and within four new additional G2L clusters (G2L_cluster 1 to 4). Thus, the fact that G2L-RTs together with RT- CRISPR from clades 2 to 13 branch off from a common node, raise the possibility of a common origin of both lineages of RTs from an ancestral group II intron (Figure R1.10).



(Figure Legend in the next page)

Figure R1.10 Identified lineages of RTs associated with CRISPR-Cas systems. The subtree shows the three lineages evolving from group-II introns, one from Retron/retron-like and one from Abi-P2 RTs. The CRISPR-Cas RTs and neighboring group-II intron classes (F, D and E); G2L; Retron and Abi-P2 clades are depicted schematically, with collapsed branches (FastTree support ≥0.85). For the CRISPR-Cas RT clades, the most common RT domains or gene organizations are indicated. Prim_S indicates an archaeo-eukaryotic primase AE_Prim_S-like domain.

The large analysis also reveals that clades 9 and 13 cluster into a single clade, hereafter referred as clade 9/13. However, RT sequences from clade 9, most of them belonging to phylum Chloroflexi (except from a possible lateral transfer event found in *Poribacteria bacterium* WGA-4CII), contain an insertion of ~105 to 145 amino acid residues just upstream from the RT4 domain (this insertion was removed to avoid artefacts in the alignment used to build the phylogenetic tree shown in Figure R1.8). Together with systems linked to RTs from clade 6, one of the features that define CRISPR-Cas systems associated with RTs from clade 13 is the frequent presence of two different *cas2* genes within the CRISPR-*cas* loci (Figure R1.9). The present phylogenetic analysis also reveals the homogeneity of some of the clades: clades 9 and 13 only contain RTs alone, clades 3, 6, 7 and 10, only RTCas1 fusions, finally, clade 8 only present Cas6RTCas1 fusions (Figure R1.10). By contrast, clades 4 and 5 harbor both RTs and RT-Cas1 fusions (Figure R1.10). Furthermore, the RT and RTCas1 fusions belonging to these clades branch off from different single nodes, suggesting single RTCas1 fusion events occurring within the clades, rather than fission events.

In this large dataset, clades 2 and 11 continue branching together as described in the previous section (Figure R1.10; Appendix A4). Additionally, the current phylogeny also reveals that clade 12 branch off together with clade 2 and 11 from a well-supported common node, suggesting that these clades descended from a common ancestor (Figure 1.10; Appendix A3). The previously reported members of these clades only contain only RTCas1 (clades 2 and 12) and Cas6RTCas1 fusions (clade 11). However, the current phylogenetic analysis reveals several RTs sequences adjacent to or fused at the C-terminus to an archaeo-eukaryotic primase (AEP) domain (AE_Prim_S_like) similar to the small catalytic subunit PriS which are found at the base of clade 12 (Appendix A4). Furthermore, despite the evolutionary proximity of RT domains from this clade, Cas1 domain from clade 12 RTCas1 fusions have a distinct origin that those Cas1 proteins adjacent to the RTPrim_S fusions in the other members of the clade, suggesting two different events of acquisition of RT domains from clade 12 by CRISPR-*cas* loci. Indeed, the generation of the RTCas1 fusion protein could trigger the loss of the AE_Prim_S_like domain of these particular group.

Nevertheless, only 16 protein sequences present an RT-Prim_S domain architecture in the NCBI database, including some of RTs from clade 12. The phylogeny of the AEPs reveals a total of 13 different families, 12 of which can be grouped into three major clades: the AEP proper clade, the NCLDV-herpesvirus primase clade, and the Prim-Pol family. All these families share three conserved motifs (I, II and III) essential for catalysis (Iyer *et al.*, 2005; Kazlauskas *et al.*, 2018). The presence of the three conserved motifs in the AE_Prim_S_like domain adjacent or fused to the RTs from clade 12 suggest that this domain is certainly a member of the AEP family. A phylogenetic reconstruction with 62 known AEPs sequences indicated that the AEP domain of the RT-CRISPR sequences formed a new lineage of primases within the AEP proper clade phylogenetically close to archaeal and eukaryotic PriS proteins, NHEJ primases, and Lef-1-like primases of baculoviruses (Figure R1.11).

Until this larger computational survey, all RTs associated with CRISPR-Cas sytems were thought to have evolved from group II introns. Nevertheless, the pipeline carried out in the current study reveals the existence of two novel clades (14 and 15) that branched off from Retron and Abi-P2 RT sequences, respectively (Figure R1.8 and R1.10). The novel clade 14 is restrain to the classes Flavobacteria, Cytophagia and Sphingobacteria within the phylum Bacteroidetes. The association between retron/retron-like RTs and CRISPR-cas loci constituted a singular event, since the most related RT sequences to those from clade 14 lack CRISPR-Cas systems in their neighborhood, suggesting the existence of a common ancestor within this phylum (Appendix A5). Referring to clade15, two RT sequences closely related to Abi-P2 group were found to be related to type I-C CRISPR-Cas systems

(Figure R1.8 and R1.10). These are harbored by *Basfia succiniciproducens* and *Haemophilus haemolyticus* strain HK386 both from order Pasteurellales.

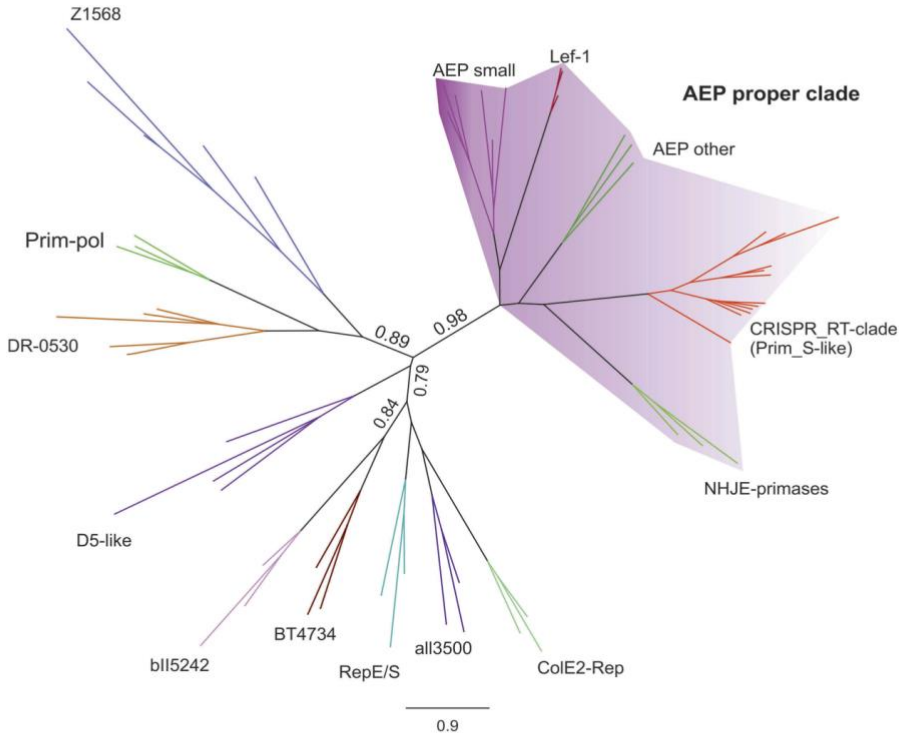


Figure R1.11 Phylogeny of CRISPR-Cas encoded RT AE_Prim_S_like domains. The tree was built using FastTree program, from an alignment of 62 protein sequences, including Prim-Pol clade (Z1568-like family, DR0530-like family, all3500-like family, bll5242-like family, ColE2 Rep-like family, RepE/RepS family), members of the AEP proper clade (AEP small_PriS proteins, NHEJ primases, Lef-1-like primases of baculoviruses and other related sequences), BT4734-like family, and the AE_Prim_S_like domain of 14 unique RT proteins with this architecture (NCBI database). All the clades except the all3500-like family (FastTree support 0.65) have a FastTree support ≥0.85.

Similar to what happen in clade 14, the closest Abi-P2 sequences are not in the vicinity of any CRISPR-Cas system. A search for close relatives to this two RTs led to the identification of additional Abi-P2 RT sequences linked to type I-C systems. Interestingly, the phylogeny of these RTs reveals that they split into two subgroups which correspond to distinct ecological niches: *H. haemolyticus* group appears to be restrain to the human microbiome, whereas members from the *B. succiniciproducens* group are present in livestock- and animal-associated habitats (Appendix A6).

The analysis of the data of this computational survey, indicates that most of the RT are associated with type III CRISPR-Cas systems. However, it was noticed that two RT sequences were actually associated with type VI-A CRISPR-Cas system. Furthermore, each sequence belonged to a different clade: the RT protein from *Rhodovulum* sp. MB263 is member of clade 7, whereas the RT from *Eubacteriaceae bacterium* CHKCI004 is part of clade 12 (Appendix A1). The presence of RTs in the neighborhood of type VI systems in two distinct clades could indicate a broader relationship between both elements. Thus, in order to find more examples of type VI CRISPR-Cas systems associated with RT sequences a search in different databases was carried out as it is explained in the following section.

R.1.3.6. Reverse Transcriptases associated with type VI CRISPR-Cas systems

Type VI systems are Class 2 CRISPR-Cas systems in which the interference machinery is formed by a single effector nuclease, known as Cas13, that exclusively targets single-stranded RNA (see section I.4.2.3). Although the adaptation stage remains largely uncharacterized in these systems, a recent study has shown that an acquisition-deficient type VI-B system from *Flavobacterium columnare* is able to acquire spacers in *trans* using the adaptation module from a type II-C CRISPR-*cas* locus (Hoikkala *et al.*, 2020). As type VI are RNA-targeting systems, it has not been unreasonable to think that these particular systems could contain adaptation modules able to acquire spacers from RNA phages. Aiming to answer this issue a search for RTs associated with Cas13 effectors was carried out. Briefly, non-redundant Cas13 proteins were detected using the different profiles for the four Cas13 subtypes (A to D). Furthermore, the diversity of Cas13 homologs was increased using metagenomic data (Toro *et al.*, 2019b). Then, the computational pipeline used in the previous section (Toro *et al.*, 2019a) was used to look for RT sequences on the vicinity (± 30 kb) of Cas13.

Interestingly, some type VI systems were found associated with adaptation modules, several including RTCas1 fusions. All these fusions were only related to subtype VI-A, in fact, approximately a 15% of all subtype VI-A systems contain an adaptive module with an RTCas1 fusion, indicating that these systems may be able to acquire spacers from RNA. A dataset of 49 unique Cas13a proteins was used to analyze the relationships of the Cas13a proteins with an associated RT domain in the type VI-A CRISPR-*cas* loci (Appendix A7). The phylogenetic tree constructed using these sequences reveals two major groups of Cas13a sequences linked to RTCas1 fusions, hereafter referred as type VI-A/RT1 and VI-A/RT2 systems (Figure R1.12A). Furthermore, these sequences split into two distinct clades clustering with other related Cas13a proteins lacking the RTCas1 fusion. The type VI-A/RT1 systems are formed by the Cas13a from *Rhodovulum* sp. MB263, and other associated Cas13a sequences (*Rhodovulum kholense* and *Rhizobium* sp. SPY-1), whereas type VI-A/RT2 systems comprised the Cas13a protein from Eubacteriaceae bacterium CHKCI004 and related sequences (*Drancourtella* sp. An57 and several *Eubacterium rectale* strains. All these species contain an RTCas1 fusion associated with the Cas13a (Figure R1.12B).

Cas13a proteins of types VI-A/RT1 and VI-A/RT2 are grouped within two separate clades 1 and 2, respectively (Figure R1.12A). Clade 1 comprises Cas13a sequences mainly from class Alphaproteobacteria, including families Rhodobacteraceae, Rhodospirillaceae, and Rhizobiaceae. Furthermore, one example of a sequence from class Spirochaetia is contained within clade 1, possibly corresponding to a lateral transfer. Contrary, the Cas13a sequences from clade 2 belong to three families of class Clostridia (Lachnopiraceae, Ruminococcaceae, and Eubacteriaceae). Due to these differential distributions, it could be hypothesized that the two type VI-A/RT systems could have emerged independently during evolution. In clade 1, the Cas13a proteins either present an adaptive module with an RTCas1 fusion or lack this module. The RTCas1 found in clade 1 are part of the clade 7 of the RT phylogeny, most of them associated with type III-D systems (Table R1 and Appendix A1).

Thus, it is plausible that CRISPR-Cas adaptation genes from one of these type III CRISPR-*cas* loci were recruited by a type VI-A CRISPR-Cas system (Figure R1.13). By contrast, in clade 12 the Cas13a proteins that lack an RTCas1 fusion present a different adaptation module, which may indicate distinct events of domestication of these modules by Cas13a proteins of this clade.

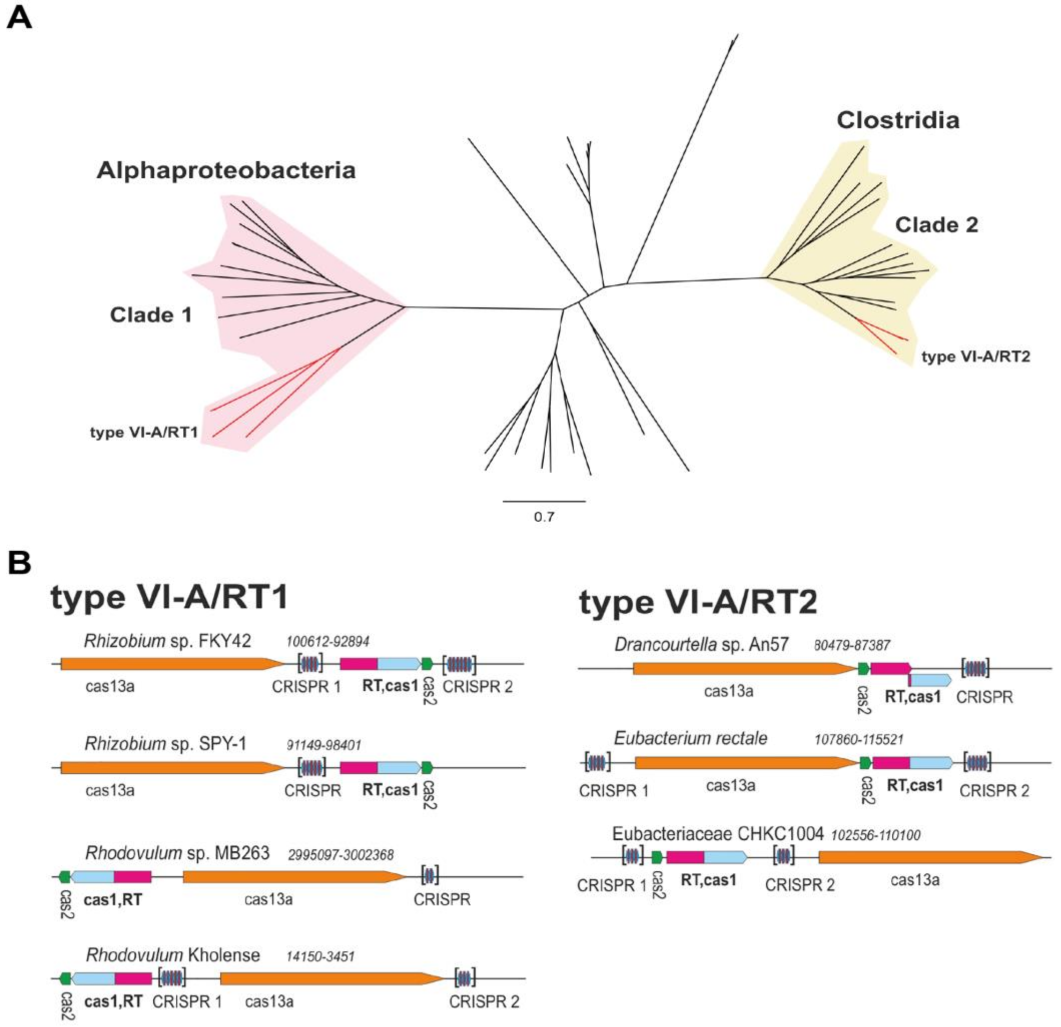


Figure R1.12 Cas13a proteins associated with RTCas1 fusions. (A) Phylogeny of Cas13a proteins. The unrooted tree was constructed with the FastTree program from an alignment of unique predicted Cas13a proteins identified in genomics and metagenomics databases. The corresponding sequence, accession number, species name is provided in Appendix A7. The branches corresponding to the Cas13a with an associated RTCas1 fusion denoted type VI-A/RT1 and type VI-A/RT2 are indicated in red. Clade 1 is mostly restricted to Alphaproteobacteria and clade 2 to Clostridia. **(B)** Architectures of genomic loci for the representative variants of type VI-A/RT1 and type VI-A/RT2 systems. For each locus, the species, nucleotide coordinates, and loci are indicated. Genes are shown roughly to scale; CRISPR Arrays are indicated in brackets and are not shown to scale.

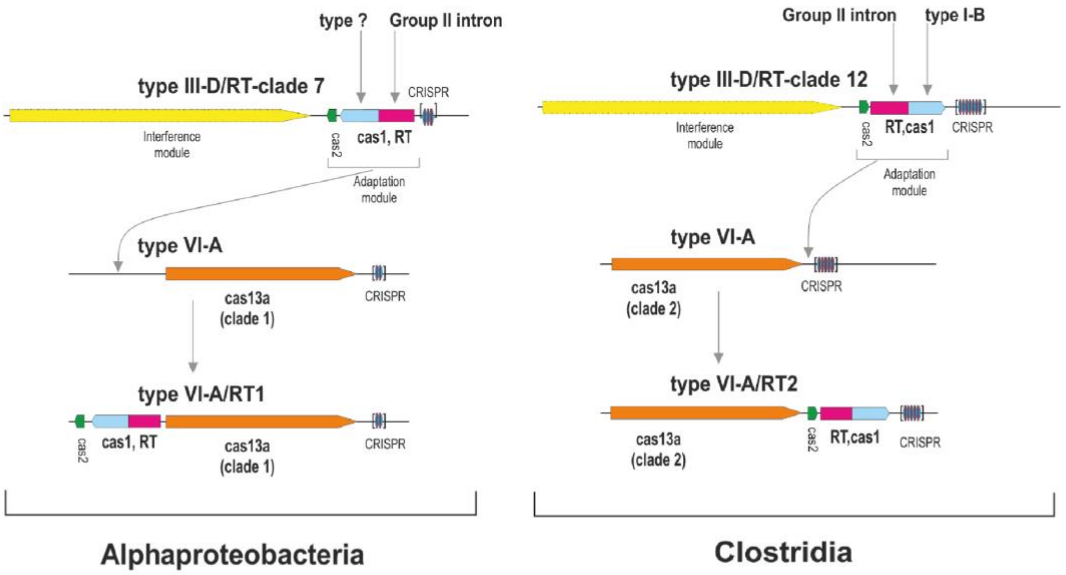


Figure R1.13 Origin of the type VI-A/RT1 and type VI-A/RT2 subtypes. The figure depicts a hypothetical scenario for the origin of adaptation modules for RT-containing type VI-A CRISPR-Cas systems. Different adaptation modules containing RTCas1 and Cas2 proteins were captured independently by distinct Cas13a proteins, probably from type III-D systems. Note that the interference module of type III systems encodes a multisubunit effector complex.

To trace the origin of adaptation modules linked to type VI-A/RT2 systems, a search for novel sequences displaying similarity to those of the Cas1 and RTCas1 proteins was performed. A dataset of non-redundant Cas1 proteins was used to construct a phylogenetic tree (Appendix A8.A), which reveals that Cas1 and RTCas1 sequences clustered into two independent groups. The Cas1 alone proteins clustered with a group of Cas1 sequences mostly associated to type III-A CRISPR-Cas systems. Nevertheless, the RTCas1 sequences of clade 2 clustered with a group of RTCas1 proteins linked to type III-D systems. These findings suggest that type VI systems within clade 2 have independently domesticated Cas1 and RTCas1 sequences from type III-A and type III-D adaptation modules, respectively.

On the other hand, Cas2 proteins associated with Cas13a proteins within clade 2 were used as a query to search for Cas2 homologs in order to infer the evolutionary origin of Cas2 of clade 2. Through an approach similar to that described above, a final dataset of 537 Cas2 proteins was used to build a phylogenetic tree (Appendix A8.B). As a result of this analysis, the Cas2 proteins of type VI-A/RT2 systems were clustered with other Cas2 sequences mostly linked to type III-D systems. However, the Cas2 sequences of type VI-A systems, which lack the RT domain, were grouped together in a separate clade with other sequences mostly associated with type III-A systems. Thus, the Cas2 phylogeny is consistent with the Cas1 phylogeny, providing further evidence to hypothesize that a series of evolutionary events in Clostridia enhanced that a whole type III-D adaptation unit, comprised by RTCas1 and Cas2 proteins, was captured by type VI-A/RT2 systems (Figure R1.13).

The search of novel examples of type VI-A CRISPR-Cas systems associated with RTCas1 fusion protein in metagenomes increased the number of type VI-A/RT1 and RTs systems until 78 homologs (Toro *et al.*, 2019b). Surprisingly, a sequence from sediment metagenome samples (MGYP000128950304) harbored an associated Cas6RTCas1 fusion. As described in the previous sections, according to the RT phylogeny most Cas6RTCas1 proteins are clustered in clades 8 and 11 (Figures R1.6 and R1.10). A phylogenetic tree comprising this protein as well as those from clade 8 and 11 indicated that this fusion protein is closely related to those from clade 8 (data not shown). All these data show a highly dynamic association between different RTCas1 fusions and type VI-A CRISPR-Cas systems and predict that the mining of metagenome data could lead to the discovery of novel associations between RTs and type VI systems.

Chapter 2: Characterization of RT-containing CRISPR-Cas systems

R.2.1 Background

The adaptation stage of CRISPR-Cas mediated immunity consists in the capture of a foreign nucleic acid from an invading agent, known as a prespacer, and their integration into the CRISPR Array as a new spacer between two repeat sequences, the direct repeats (DRs) (see section I.5). It has been demonstrated that Cas1 and Cas2, which form an integrase complex, are the proteins responsible for performing this step (Yosef *et al.*, 2012; Nuñez *et al.*, 2014). Furthermore, Cas1 has been shown to interacts with other several proteins involved in the adaption process (Koonin *et al.*, 2017). As shown in the previous chapter, one of these ancillary proteins is the RT. Among RTs associated with CRISPR-Cas system predominate those closely related to group II introns, which are found either separately or naturally fused at the C-terminus with Cas1. Furthermore, these RTs are usually linked to type III systems, a class 1 CRISPR-Cas system with a multi-subunit crRNA-effector complex able to target both RNA and DNA¡ when it is transcriptionally active (section I.4.1.2). This particular association could suggest that RT-containing CRISPR-Cas systems would be able to integrate novel spacers from RNA sources.

This hypothesis has been recently validated by two independent studies using different systems: the adaptive operon from *Marinomas Mediterranea* MMB-1 containing a Cas6RTCas1 fusion protein (Silas *et al.*, 2016), and the 1 system from *Fusicanibacter saccharivorans*, with an RTCas1 (Schmidt *et al.*, 2018). Both studies show that RT-CRISPR systems can acquire new spacers directly from RNA *in vivo*, in an RT-dependent manner. However, little is known about the biochemical mechanism of how RT-containing integration complex capture and integrate novel spacers within the CRISPR Array. It has been suggested that the acquisition of spacers from an RNA origin occurs by direct ligation of the RNA prespacer into the Direct Repeats (DRs), and then, the 3′ end generated by cleavage of the opposite DNA strand is then poised for use as a primer for target-primed reverse transcription (TPRT) (Zimmerly *et al.*, 1995). Nevertheless, a recent preprint that obtain the cryo-EM structure of the Cas6RTCas1-Cas2 integrase complex from *Thiomicrospira* (Wang *et al.*, 2020), do not show evidence of analogous TPRT reaction in the biochemical assays performed. Thus, the *in vitro* characterization of novel examples of RT-containing CRISPR-Cas systems could provide additional insights into complex formation and the molecular mechanism of spacer acquisition in type III CRISPR-Cas systems.

R.2.2 Biochemical RT activity of diverse RT homologs associated with CRISPR-Cas systems.

In the study carried out in this thesis, the detection of an exogenous RT activity has been established as the first criteria to select CRISPR-Cas systems to be further investigated. The analysis of the phylogenetic relationships of RTs linked to CRISPR-Cas system carried out in the previous chapter has revealed that most of the identified RTs are part of a complete adaptive modules together with *cas1* and *cas2* genes as well as the CRISPR Arrays. Thus, in order to study different RT-containing CRISPR-Cas systems several representatives from different clades were selected to analyze whether the RT is active, including examples of RT alone and RTCas1 fusion proteins. Two criteria were taken into account for the selection of these representatives:

i. To estimate the active state of each adaptive complex it was assumed that the most active systems would be those that present all the components (*RT,* cas1 and *cas2* genes) and at least one CRISPR Array with their leader sequence. Moreover, the number of spacers within a CRISPR Array was also used as an indicator of an active CRISPR-Cas adaptive unit. Additionally, the presence of the rest of the other CRISPR genes such as those involved in the expression or the interference stages would suggest a completely functional system.

ii. Availability of the microorganism or the genomic DNA sources that harbor the whole system.

Based on the above criteria, representatives from several clades of the main lineage of bacterial RTs linked to CRISPR-Cas systems, which includes clades 3 to 10 and clade 13, were selected for further biochemical characterization (Table

M2). Thus, among the selected systems are included examples from clade 4 (Delta-proteobacterium *Desulfobacca acetoxidans* DSM 11109, and *Chlorobium limicola* DSM 245), clade 5 (the Cyanobacterium *Scytonema hofmanni* PCC 7110), clade 6 (Gamma-proteobacterium *Vibrio vulnificus* YJ016) and clade 9 (Chloroflexi *Roseiflexus castenholzii* DSM 13941). Genomic DNA of most bacterial strains was obtained from the German Microorganism and Cell Culture Collection (DSMZ, https://www.dsmz.de/), with the exception of the DNA of *S. hofmanni* PCC 7110, which the living organism was available in solid medium courtesy of Dr. Agustín Vioque (IBVF-CSIC-Seville) (Figure R2.1A) and the DNA of *V. vulnificus* YJ016, provided by courtesy of Dr. Carmen Amaro (University of Valencia).

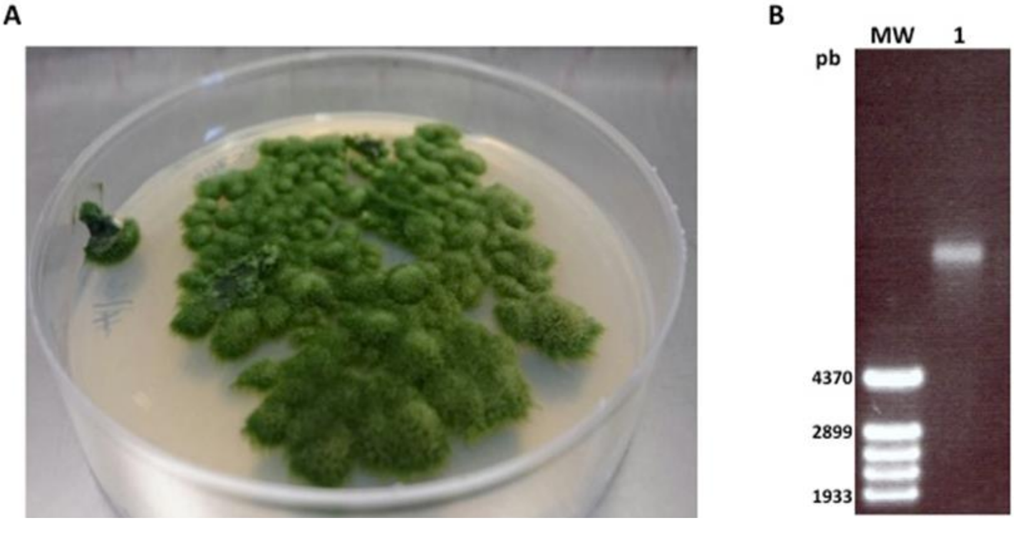
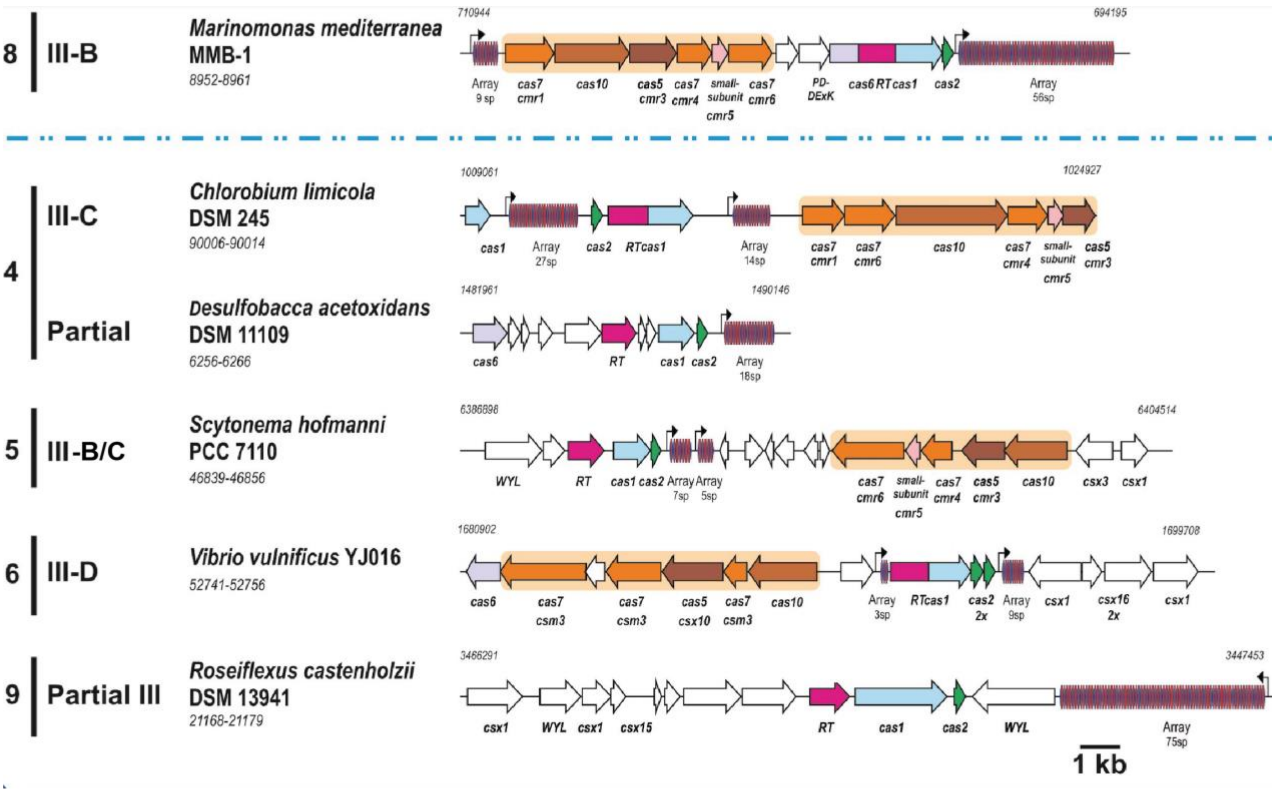


Figure R2.1 The Cyanobacterium *Scytonema hofmanni* PCC 7110. (A) Colonies of *S. hofmanni* PCC 7110 growth in BG11o solid medium. **(B)** Analysis in agarose gel of the total genomic DNA of *S. hofmanni* PCC 7110 (1).

The genomic DNA of *S. hofmanni* PCC 7110 was extracted as is detailed in section M.5.3. (Figure R2.1B). Once obtained the genomic DNA of all the selected strains (Table M2), the DNA fragment encoding the corresponding *RT* or *RTCas1* genes was obtained by PCR amplification using primers including the *BamH*I and/or *Bgl*II restriction sites at the ends of the amplified product for subsequent cloning in the pMal-Flag vector (Tables M5 and M6), which will be used for expression and purification of the selected proteins. The CRISPR-*cas* loci of the selected RT or RT*cas1* genes are shown in Figure R2.2.

The cloning of the diverse amplified products in the vector pMal-Flag with the indicated primers leads to a construction in which the opening reading frame (ORF) of the RT or RT*cas1* genes remains in phase with the Maltose Binding Protein (MBP) encoded by this vector (Tables M5 and M6). Then, these plasmids allow the expression of the RT or RTCas1 as a fusion protein with the MBP, which is placed at the N-termini of the fusion. MBP is a highly soluble protein that increased the solubility of the C-termini fused protein. Furthermore, the presence of the epitope Flag is used for protein detection with anti-Flag antibody.

Upon verification of the cloned fragments in pMal-Flag by Sanger sequencing, the MBP-RT or MBP-RTCas1 fusion proteins were expressed and purified as described in section M.11.1. Briefly, the expression of the recombinant protein was induced with IPTG in *E. coli* Rosetta cells. Then, the culture was harvested, and subsequently the cells lysed. After removing cell debris, the supernatant was mixed with amylose beads which specifically bind the protein of interest. Finally, the recombinant protein was eluted with maltose, pooled and store at -20ºC in a buffer containing 50% glycerol. In all cases, the recombinant protein was highly expressed as observed after induction with IPTG (Appendix B1). However, the purification yield of this protocol was different for every protein (Figure R3.A). The intron-encoded protein (IEP) of the *RmInt1* group II intron was also purified and used as positive control of the entire process (Garcia-Rodriguez *et al.*, 2019). To demonstrate whether the selected RT or RTCas1 are functional, exogenous RT activity was measured. An *in vitro* assay was used to determine the cDNA synthesis carried out by the RT domains through the incorporation of radiolabel [α-P^{32}] dTTP on a poly(rA) substrate, which depends on the presence of an oligo-dT (section M.12.1). This experiment revealed that only two of the five selected RT or RTCas1 proteins presented a significant exogenous RT activity: the RT from *S. hofmanni* PCC 7110 and the RTCas1 fusion from *V. vulnificus* YJ016 (Figure R2.3B).



R.2.2 Architecture of RT-containing CRISPR-Cas systems selected to test exogenous RT activity. The CRISPR-*cas* locus o *mediterranea* MMB-1 characterized in Silas et al., 2016 is also shown. The effector modules are indicated in beige background. The Cas6 is colored in purple. RT and Cas1 domains are indicated in fuchsia and blue, respectively, whereas *cas2* gene are colored in gree Arrays and number of spacers (sp) are indicated. The black arrows indicate the putative promoter sequence. Ancillary and unknow color-coded. For each locus, the clade of the RT-CRISPR phylogeny, the type of CRISPR-Cas system, the name of the organism, th gene locus tag (final digits) and the genomic coordinates are indicated.

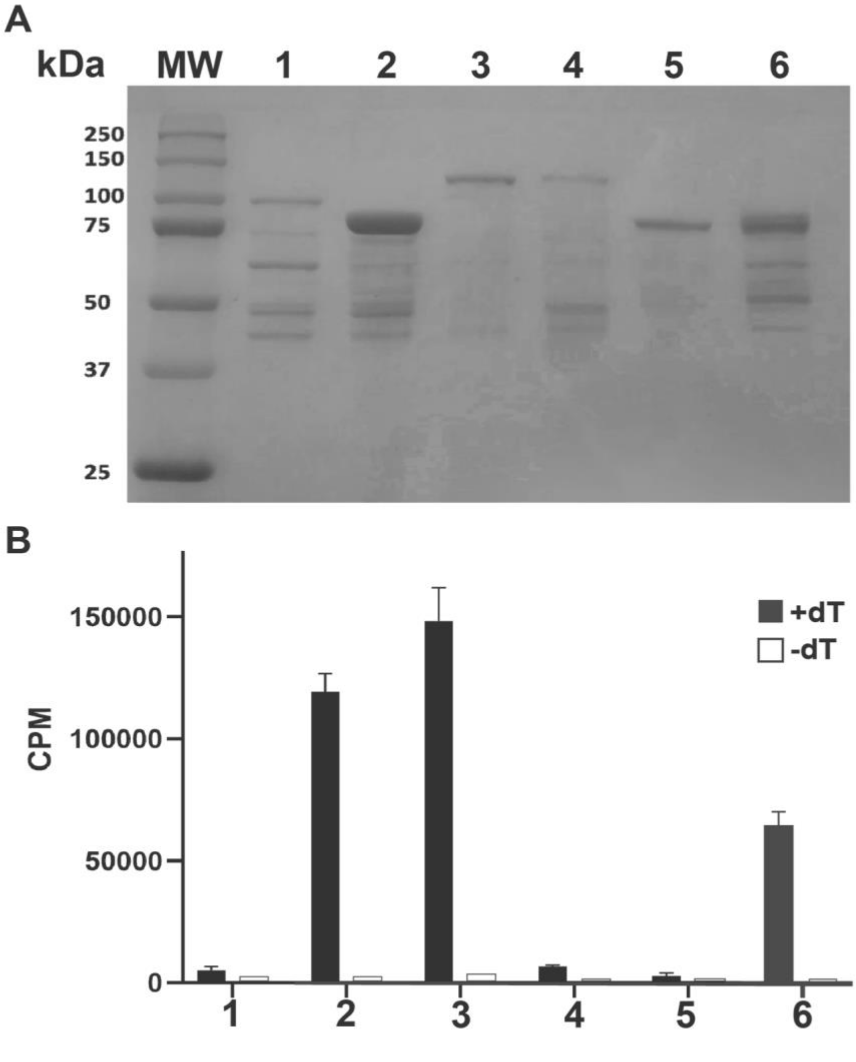


Figure R2.3 Purification and *in vitro* RT activity of selected RT or RTCas1 proteins. (A) SDS-PAGE electrophoresis (10%) showing the recombinant proteins (1-10 µg) after the purification process (section M.11.1). A marker to show the molecular weight (MW) in kilodaltons (kDa) of the selected proteins is displayed on the left. **(B)** Exogenous RT activity of selected RT or RTCas1 proteins. 200 ng of the proteins were incubated with and oligo(dT) and the poly(rA) exogenous substrate (+dT) or only with the substrate (-dT) as described in section M.12.1. The exogenous RT activity is measured in counts per million (CPM). The error bars in the +dT points are based on three replicates. Lane-numbers correspond: (1) MBP-RT from *Roxeiflexus castelhonzii* DSM 13941 (95,7 kDa), (2) MBP-RT from *Scytonema hofmanni* PCC 7110 (83,6 kDa), (3) MBP-RTcas1 from *Vibrio vulnificus* YJ016 (125,4 kDa), (4) MBP-RTCas1 from *Chlorobium limicola* DSM 245 (130,8 kDa), (5) *Desulfobacca acetoxidans* DSM 11109 (82,5 kDa) and (6) MBP-IEP from *RmInt1* group II intron (78 kDa).

Interestingly, the RT from *S. hofmanni* PCC 7110 and the RTCas1 from *V. vulnificus* YJ06 present approximately the double of the exogenous RT activity of the simultaneously purified IEP from *RmInt1* (Figure R3.B). The CRISPR-Cas adaptive modules containing these proteins will be studied in detail in the following sections.

R.2.3 The RT-Cas1-Cas2 adaptive operon from *Scytonema hofmanni* PCC 7110

R.2.3.1 The CRISPR-Cas systems of Scytonema hofmanni *PCC 7110.*

The genome of the cyanobacteria *Scytonema hofmanni* strain PCC 7110 contains a great number and variety of CRISPR-Cas systems. Ten different CRISPR-Cas systems are found in this strain, all of them belong to Class 1 CRISPR-Cas systems (Appendix B2). Thus, this strain contains five type I CRISPR-*cas loci* (two subtypes I-B, two I-D, and one I-U), four type III (three subtype III-B and one III-D) and another mixed CRISPR-*cas* loci, which contains effectors genes of both type I-D and III-D CRISPR-Cas systems. Moreover, all the system of this strain contains at least one CRISPR Array, reaching in some cases up to four (Appendix B2).

The four type III CRISPR-Cas system present in this strain harbor an RT domain closely associated with the adaptive module of these systems. Indeed, examples of different evolutionary stages of the association between RTs and CRISPR-Cas systems co-exist in *S. hofmanni* PCC 7110 since two of them clustered in clade 5 of the phylogeny of RTs linked to CRISPR-*cas* loci, and two RTCas1 fusion proteins which are part of the clade 3 of this phylogeny (Appendix A1 and B2). However, in one of the CRISPR-*cas* locus containing a RT alone lacks a *cas2* gene, while one of the RTCas1 fusions is frameshifted (red-marked in Appendix B2). In any case, the abundance of RT-containing CRISPR-Cas systems could suggest that RT activity has an important role to acquire functional spacers in the cyanobacterial environment.

Among this great diversity, the analysis in this work will be focus on the CRISPR-Cas system which contains the functional RT alone which is described in the previous section (Figure R2.3B). This RT is associated with a complete type III-B/C CRISPR-Cas system, which only lacks a *cas6* gene (Figure R2.4). Nevertheless, the presence of multiple CRISPR-*cas* loci in this strain suggest that the Cas6 protein of another system could act in *trans* to process the two CRISPR Arrays present in this locus. Moreover, in this particular locus, the arrays differ in the leader sequence and the direct repeats (DRs). Importantly, the adaptive module of this system is constituted by an RT-Cas1-Cas2 operon (Figure R2.4).

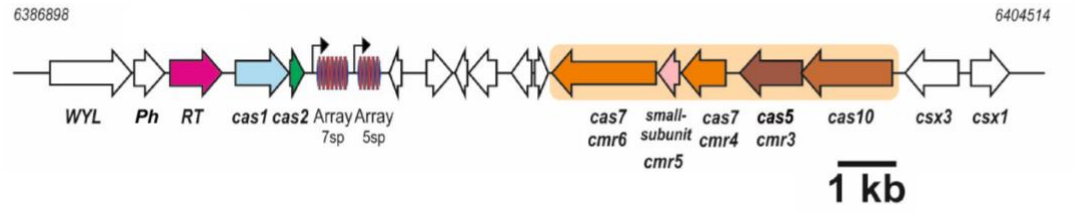


Figure R2.4. Architecture of the type III-B/C CRISPR-*cas* loci of *Scytonema hofmanni* PCC 7110. The CRISPR-*cas* locus consists of a five-gene cassette putatively encoding the type III-B/C effector complex (indicated by a beige background). When available, both a systematic (above) and a 'legacy' (below) names of effector genes are indicated. In the other orientation is encoded the adaptation module, which consists of three genes encoding a RT (fuchsia), *cas1* (blue) and *cas2* (green) genes. Two CRISPR Arrays with 7 and 5 spacers, respectively are located just downstream of the adaptation module. Four ancillary genes (a WYL-domain-containing protein and putative phosphohydrolase encoding genes, *csx1* and *csx3*) and several genes of unknown function are non-color coded. Black arrows indicate the identified leader sequence promoters. The genomic coordinates of the CRISPR-*cas* locus *in* the chromosome of *S. hofmanni* PCC 7110 are indicated.

Interestingly, another seven representatives of the phylum Cyanobacteria present a CRISPR-Cas adaptive module closely related to that associated with the RT alone of *S. hofmanni* PCC 7110 (Appendix B3). The analysis of the genomic context of these systems reveals the presence in half of the loci of two putative ancillary genes usually placed just upstream of the *RT* gene and always in the same order: a WYL-domain-containing protein and a putative phosphohydrolase (Ph) (Appendix B3). As consequence of this conserved genomic organization, it seems that both genes could somehow be part of this particular adaptive module as well. In fact, a WYL-domain-containing protein have been shown to play a regulatory role in type VI CRISPR-Cas systems (Yan *et al.*, 2018). Thus, these two genes will be included in the protein:protein interaction assays that will be shown in the next section in order to test whether these two proteins could have a function related to the RT-containing CRISPR-Cas adaptation module of *S. hofmanni* PCC 7110.

R.2.3.2 The RT alone from S. hofmanni PCC7110 interact with both Cas1 and Cas2.

To study the interaction of the RT alone from *S. hofmanni* PCC 7110 with the other elements of the integrase complex the first step was optimized the purification of all the core proteins involved in the acquisition process (RT, Cas1 and Cas2). With this aim, apart from the ORF containing the *RT* gene already cloned in a pMal-Flag vector (Table M4), *cas1* and *cas2* were cloned independently in a pET16b vector (Table M4). With this aim, Cas1 and Cas2 encoding genes were amplified adding the restriction sites *Nde*I and *BamH*I at each end of the PCR products. This amplification allows the cloning of *cas1* and *cas2* in phase with the N-termini 10xHis-tag contained by the pET16b vector (Table M5). The His-tag will serve for the purification of both proteins by affinity chromatography using a His-Trap column (section M.11.2).

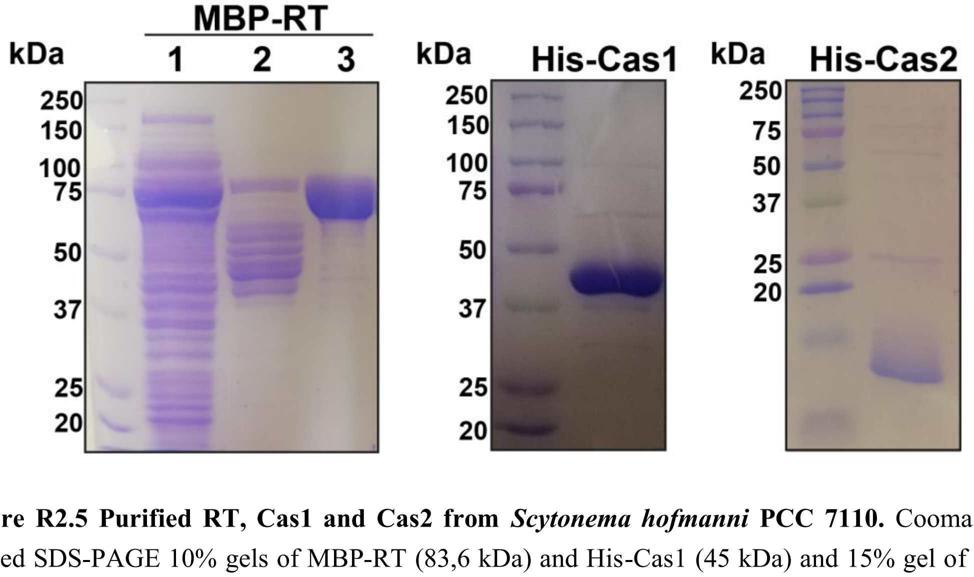


Figure R2.5 Purified RT, Cas1 and Cas2 from *Scytonema hofmanni* PCC 7110. Coomassie-stained SDS-PAGE 10% gels of MBP-RT (83,6 kDa) and His-Cas1 (45 kDa) and 15% gel of His-Cas2 (13kDa) (gel 15%). For MBP-RT the protein is showed after induction with 0,3 mM IPTG (1), after the amylose column step (2) and after the final heparin column step (3).

A single-step protocol is not enough to obtain a highly-purity MBP-RT as other contaminant proteins eluted together with the recombinant protein (Figure R2.3A). To further purify the MBP-RT protein a two-step protocol consisting of two affinity columns in tandem was used. Firstly, an amylose column, and then, a heparin column, which specifically binds nucleic acid-binding proteins, enables a higher purification level of the MBP-tagged protein (section M.11.2). This purification process made it possible to obtain proteins with more than 95% purity (Figure R2.5). In the case of Cas1 and Cas2 proteins, both proteins were purified using a single-step protocol based in a nickel column, that specifically bind the proteins with the His-Tag (section M.11.2). The elution of the bound proteins with an imidazole gradient results in proteins with more than 90% purity (Figure R2.5).

The oligomeric state of the purified proteins was studied using size exclusion chromatography (SEC) as indicated in Section M.11.3. Running 300 ng of each one of the purified proteins (Figure R2.5) on a Superdex 200 increase 10/300 GL, it was revealed that MBP-RT forms a dimer in solution, although there is a percentage of the protein which is found as a monomer (Figure R2.6A). Preparation of His-Cas1 and His-Cas2 proteins indicates that meanwhile Cas1 forms exclusively a monomer, Cas2 is present mainly as a dimer (Figure R2.6B and C). Therefore, it was tested whether RT form a stable complex with Cas1, Cas2 or both. With this purpose, RT and Cas1, RT and Cas2 or the three proteins were mixed (150 ng of each protein) and were incubated for 30 minutes at 4ºC. In all analyzed cases no changes in the position of the peaks were observed, suggesting a null or weak protein interaction in the tested conditions (Figure R.2.6D, E and F). These results suggest that either other requirements are needed to form a stable complex, such as the addition of nucleic acids, or the used tags interfere in complex formation.

An alternative method to investigate the interaction between the components of the CRISPR-Cas adaptive module of *S. hofmanni* PCC 7110 was the use of a pull-down approach (section M.12.2.1).

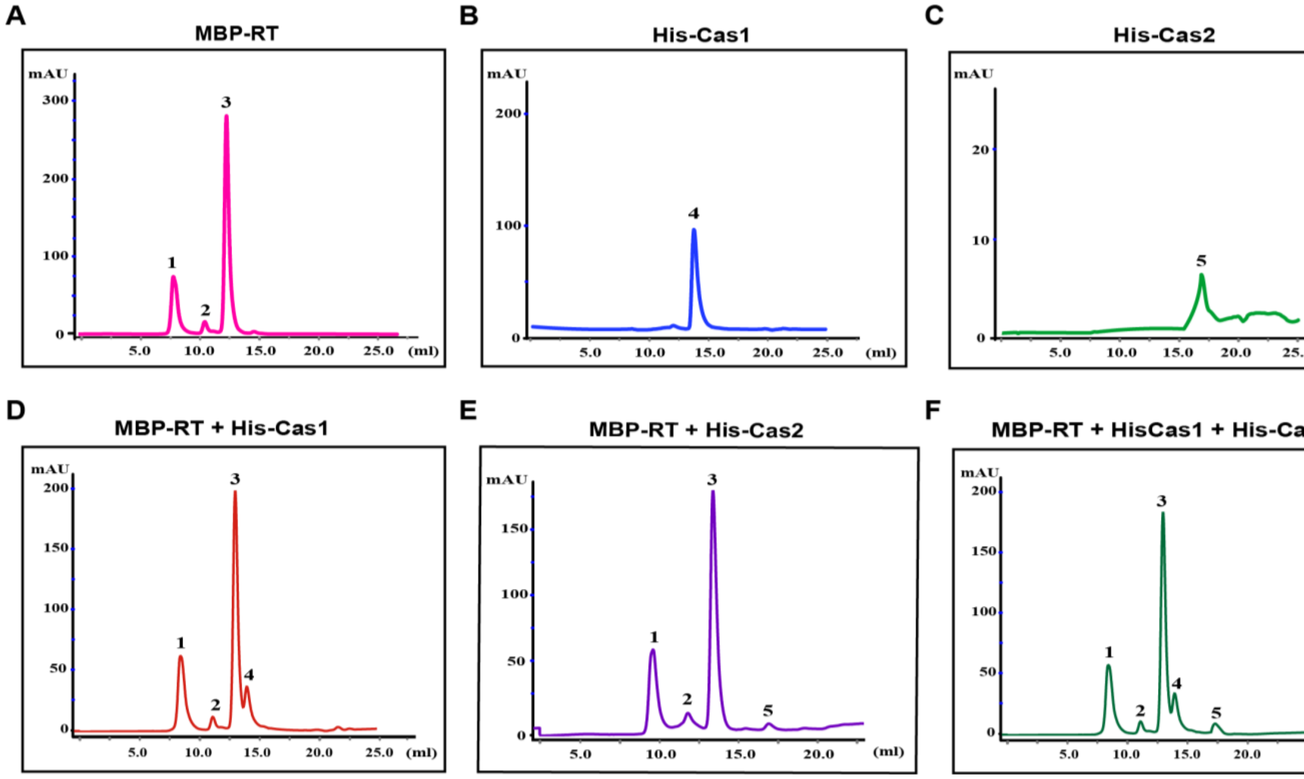


Figure R2.6 Oligomeric state of the RT, Cas1 and Cas2 proteins of *S. hofmanni* PCC 7110. (A) Oligomeric state of MBP-R
appear: aggregated protein (1), dimer (2) and a main peak corresponding to the monomer (3). **(B)** Oligomeric state of His-Cas1
corresponding to the monomer (4). **(C)** Oligomeric state of HisCas2. A single peak corresponding to the dimer (5). **(D)**, **(E)** and
state of MBP-RT + HisCas1, MBP-RT +His-Cas2 and MBP-RT + His-Cas1 + His-Cas2, respectively. The mixed proteins were i
minutes at 4ºC. In any case a complex formation was observed. All proteins were analysed running 300 ng of each protein (150 ng
Superdex 200 increase 10/300 GL.

Briefly, this method consists in the overexpression of the different proteins using IPTG in *E. coli* Rosetta. Then, the cell lysate of MBP-RT is incubated with amylose beads, that specifically binds this protein. The non-bound proteins are removed and, therefore, cell lysates containing the overexpressed His-Cas1, His-Ca2 or a mix of both are added to the tube containing the MBP-RT bound to the beads. Upon several rounds of washing, selective elution using maltose resulted in the co-elution of MBP-RT (used as bait) with Cas1, Cas2 and both proteins. The level of this interaction could be estimated by using anti-His antibodies that specifically detect the His-tagged Cas1 and Cas2 (Figure R2.7A). This assay indicates that the RT interaction is stronger with Cas2 than with Cas1. Nevertheless, when both, Cas1 and Cas2 are present in the cell lysate it seems that the interaction with Cas1 increase and that with Cas2 decrease (Figure R2.7A, lane 6). The enhanced interaction observed when the three proteins are present may suggest they form a protein complex. The interaction between RT and Cas1 and/or Cas2 is specific since no interaction is shown after the use of MBP-Flag as bait of the other proteins (Figure R2.7A).

The same approach was used to analyze the interaction between the RT and the two ancillary proteins located just upstream of the gene encoding the RT alone in the CRISPR-*cas* loci of *S. hofmanni* PCC 7110. As describe above, these two proteins are also present in close-related species/strains and may play a role in the adaptive module of this specific CRISPR-Cas system. With the aim of study the possible interaction, the ORFs encoding the WYL-domain-containing-protein and the putative phosphohydrolase were also cloned in pET16b vector, just as described for Cas1 and Cas2 (Table M5). The pull-down protocol described in the previous paragraph also reveals that the RT seems to interact with both proteins (Figure R2.7B). The interaction of the RT with the phosphohydrolase is stronger than with the WYL-domain-containing protein. Moreover, this interaction appears to be specific as neither of the proteins interact with the MBP-Flag control (Figure R2.7B).

In this chapter, it has been shown the interaction of the RT with the rest of the proteins that could be involved in the acquisition of novel spacers of the type III-B/C CRISPR-Cas system of *S. hofmanni* PCC 7110, specially with Cas1 and Cas2. However, more assays will be required to validate this interaction and to demonstrate the formation of an integrase complex that includes the RT. Unfortunately, in the *in vivo* spacer acquisition assays carried out using this particular CRISPR-Cas adaptive operon, no novel spacer acquisition events were detected using this particular system (data not shown).

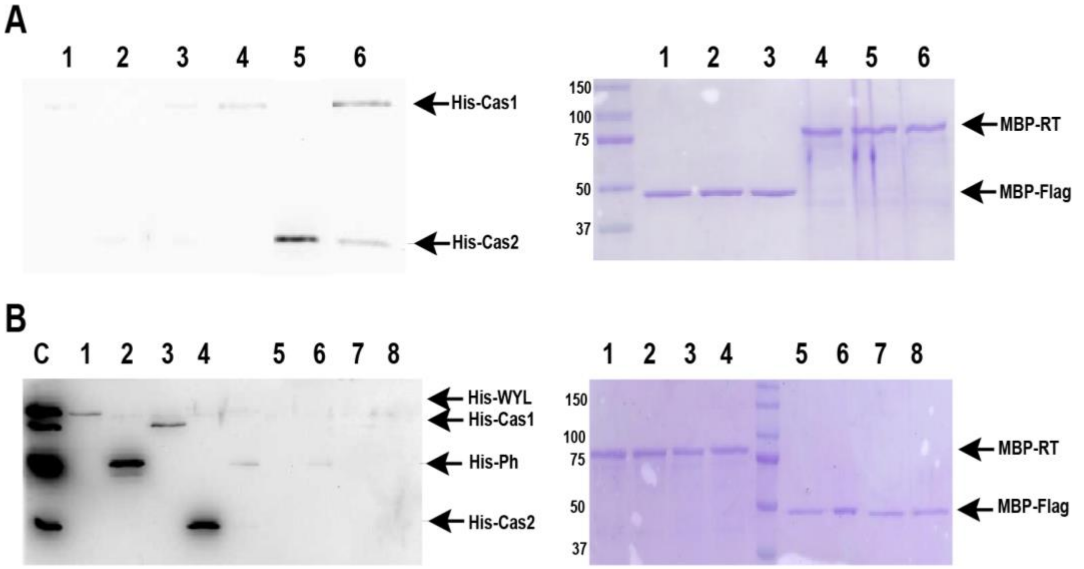


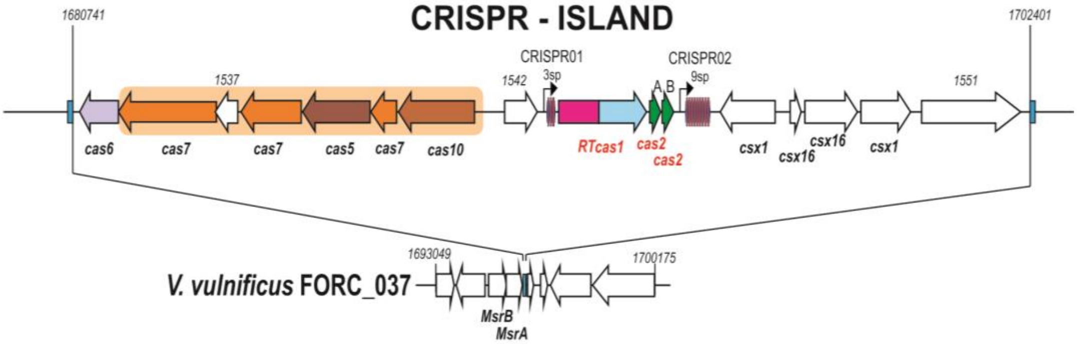
Figure R2.7 RT interaction with proteins of the adaptive module of *S. hofmanni* PCC 7110. (A) Interaction of the RT with Cas1 and Cas2. Pull-down assay of MBP-RT (4-6) and MBP-Flag as a negative control (1-3). Protein samples were assayed against supernatant containing His-tagged Cas1 (1,4), Cas2 (2,5) or a mix of both lysates (3,6). The proteins retained after elution with maltose were subjected to SDS-PAGE and stained with Coomassie blue (right), and a western blot was performed with antibodies against the His-tag (left), demonstrating the presence of both proteins, His-tagged Cas1 and Cas2, in the MBP-RT samples **(B)** Interaction of the RT with the rest of the putative members of the adaptive operon. Pull-down assay of MBP-RT (1-4) and MBP-Flag as a negative control (5-8). Protein samples were assayed against supernatant containing His-tagged WYL-domain-containing protein (1,5), phosphohydrolase (Ph) (2,6), Cas1 (3,7) and Cas2 (4,8) The proteins retained after elution with maltose were subjected to SDS-PAGE and stained with Coomassie blue (right), and a western blot was performed with antibodies against the His-tag (left), demonstrating the presence of all the proteins assayed in the MBP-RT samples. All the proteins assayed are indicated with a black arrow.

R.2.4 The RTCas1-Cas2A-Cas2B adaptive operon from *Vibrio vulnificus* YJ016

R.2.4.1 The type III-D CRISPR-Cas system of Vibrio vulnificus YJ016

Vibrio vulnificus strain YJ016 harbor a unique type III-D CRISPR-*cas* locus which is only present in a few representatives of the Vibrio genus, and some other close related Gamma-protobacterium clustered in clade 6 of the phylogeny of RTs associated with CRISPR-Cas systems (Appendix A1 and B4). However, strain YJ016 is the only one within *V. vulnificus* species that contains this particular RT-containing type III-D system. The genomic neighborhood of the RTCas1 locus from *V. vulnificus* strain YJ016 encoding genes was retrieved, as previously described in section M.1.3. This locus spans 21.6 kb and it is placed on the chromosome II (Figure R2.8). A blast search in the genome of other strains of *V. vulnificus* showed that this locus constituted a genomic island which is located downstream from a highly conserved operon encoding two peptide-methionine-S-oxide reductase genes (*msrA* and *msrB*) (Figure R2.8; McDonald *et al.*, 2019). This finding may suggest that this type III-D CRISPR-Cas system has likely been acquired by a lateral transfer event.

This RTCas1-containing type III-D locus harbor a gene-cassette encoding the type-III-D Csm effector machinery, which also includes a *csx19* gene that could act as the small unit of the interference complex (Figure R2.8). In addition, at its 5' end of this operon is encoded the *cas6* gene, responsible for the processing of the CRISPR Array to produce the mature crRNA. In the opposite orientation is encoded the adaptive operon which is constituted by the RTCas1 fusion protein and two different Cas2 (A and B) proteins. This adaptation module is flanked by two CRISPR Arrays: CRISPR01 and CRISPR02, which contains three and nine spacers, respectively (Figure R2.8). Moreover, both arrays contain identical DRs. Interestingly, the RTCas1 fusion protein (690 aa') displays a highly exogenous RT activity as previously described in this chapter (Figure R2.3B). Additionally, the VvYJ016 CRISPR-Cas system also contains a set of ancillary genes which are located downstream from the larger CRISPR Array (Figure R2.8). These ancillary genes include two *csx1* genes, encoding the RNAse that is activated once the interference complex cleavage the target, which function is to non-specifically degrade the RNA of the target region leading to cell dormancy and, therefore, protecting the bacterial population. Moreover, this locus also presents two *csx16* genes, which function remains uncharacterized, but that seems to belong to a highly diverse family of ring nucleases, which act regulating the Csx1 activity (Makarova *et al.*, 2020). Thus, it seems that the VvYJ016 type III-D CRISRP-Cas system harbor all the genes required to be completely functional.



R2.8 Architecture of the type III-D CRISPR-*cas* loci in VvYJ016. The VvYJ016 CRISPR-Cas locus consists of the one hand of five-gene cassette putatively encoding the type III-D effector complex (indicated by a beige background), including the putative small subunit (VVA1537). The interference module is followed by the gene encoding Cas6, which is responsible for crRNA processing and maturation. In the other orientation is encoded the adaptation module, which consists of three genes encoding a RTCas1 fusion protein and two Cas2 proteins (A and B), located between two CRISPR Arrays containing three and nine spacers (CRISPR01 and CRISPR02, respectively). Four ancillary genes (two csx1 and two csx16) and two genes of unknown function (VVA1542 and VVA1551) complete this CRISPR-island. Black arrows indicate the identified leader sequence promoters. The genomic coordinates of the CRISPR island in the chromosome II of *V. vulnificus* YJ016 are indicated, as well as the equivalent coordinates of strain FORC_037, which lacks the type III-D CRISPR-Cas system.

As performed with the adaptation module of *S. hofmanni* PCC7110, the first step to analyze the VvYJ016 integrase complex is the purification of all the proteins that constitute it. Here, the complex is formed by the RTCas1 fusion and the two different Cas2 proteins. Both, Cas2A and Cas2B present a 70% amino acid identity and a characteristic Cas2 structure based on 2 alpha helix and 5 beta strands, with a conserved aspartic residue at the end of the first beta strand in both proteins (Figure R2.9).

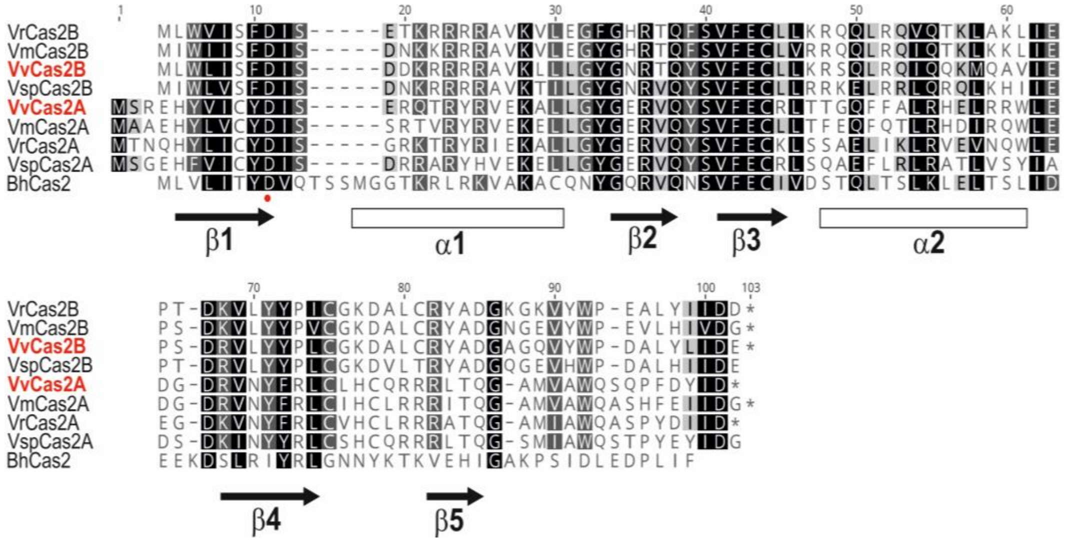

1
10
20
30
40
50
60
VrCas2B MLWVISFDIS-----ETKRRRRAVKVLEGFGHRTQFSVFECLLKRQQLRQVQTKLAKLIE
VmCas2B MIWIISFDIS-----DNKKRRRAVKVLEGYGHRTQFSVFECLVRRQQLRQIQTKLKKLIE
VvCas2B MLWLISFDIS-----DDKRRRRAVKLLLGYGNRTQYSVFECLLKRSQLRQIQQKMQAVIE
VspCas2B MIWLVSFDIS-----DNKRRRRAVKTILGYGNRVQYSVFECLLRRKELRRLQRQLKHIIE
VvCas2A MSREHYVICYDIS-----ERQTRYRVEKALLGYGERVQYSVFECRLTTGQFFALRHELRRWLE
VmCas2A MAAEHYLVCYDIS-----SRTVRYRVEKELLGYGERVQYSVFECLLTFEQFQTLRHDIRQWLE
VrCas2A MTNQHYLICYDIS-----GRKTRYRIEKALLGYGERVQYSVFECKLSSAELIKLRVEVNQWLE
VspCas2A MSGEHFVICYDIS-----DRRARYHVEKELLGYGERVQYSVFECRLSQAEFLRLRATLVSYIA
BhCas2 MLVLITYDVQTSSMGGTKRLRKVAKACQNYGQRVQNSVFECIVDSTQLTSLKLELTSLID
β1
α1
β2
β3
α2
70
80
90
100
103
VrCas2B PT-DKVLYYPICGKDALCRYADGKGKVYWP-EALYIIDD*
VmCas2B PS-DKVLYYPVCGKDALCRYADGNGEVYWP-EVLHIVDG*
VvCas2B PS-DRVLYYPLCGKDALCRYADGAGQVYWP-DALYLIDE*
VspCas2B PT-DRVLYYPLCGKDVLTRYADGQGEVHWP-DALHIIDE
VvCas2A DG-DRVNYFRLCLHCQRRRLTQG-AMVAWQSQPFDYID*
VmCas2A DG-DRVNYFRLCIHCLRRRITQG-AMVAWQASHFEIIDG*
VrCas2A EG-DKVNYFRLCVHCLRRRATQG-AMIAWQASPYDIID*
VspCas2A DS-DKINYYRLCSHCQRRRLTQG-SMIAWQSTPYEYIDG
BhCas2 EEKDSLRIYRLGNNYKTKVEHIGAKPSIDLEDPLIF
β4
β5

Figure R2.9 Sequence alignment and structural comparison of Cas2 A and B homologues. Sequence alignment of both *Vibrio vulnificus* Cas2A and B (in red) with other Cas2 homologs of the RT-CRISPR systems of clade 6 of the RT phylogeny. Conserved amino acid residues are highlighted in dark background. Secondary structures are indicated at the bottom of the alignment based on the *Bacillus halodurans* Cas2 protein structure (BhCas2; Q9KFX8.1). The critical conserved aspartic D8 (Nam *et al.*, 2012) at the end of the first beta-strand is indicated with a red dot. Bacterial species and the corresponding accession numbers of the Cas2 from the NCBI or PATRIC database are: *V. vulnificus* YJ106 (VvCas2A: WP_043877605.1, VvCas2B: WP_011152752.1); *V.rotiferianus* CAIM 577 (VrCas2A: WP_081641186.1, VrCas2B: WP_038884988.1); *V. mexicanus* strain CAIM 1540 (VmCas2A: fig|1004326.3.peg.1526, VmCas2B: fig|1004326.3.peg.1527; *Vibrio* sp. PID17 (VspCas2A: WP_099078970.1, VspCas2B: WP_099078971.1).

As describe for type I and II CRISPR-Cas system, the integrase complex is a heterohexamer comprised by a central Cas2 dimer that binds two distal Cas1 dimers (Nuñez *et al.*, 2014, Wright and Doudna, 2016). Thus, an important feature in the VvYJ016 system is to test whether Cas2A and Cas2B could form a heterodimer unit which could be required for a functional adaptive complex together with the RTCas1 fusion protein. The interaction between the genes that constitute the VvYJ016 adaptation module will be studied in detail in the following section.

R.2.4.2 Cas2A and Cas2B of V. vulnificus YJ016 form a heterodimer unit.

The first step to analyze the interaction between the proteins (RTCas1-Cas2A-Cas2B) that constitute the VvYJ016 adaptive module was to increase the yield of the purification process of each protein. As described for the RT of *S. hofmanni* PCC 7110 (section R2.3.2), the RTCas1 of *V. vulnifucus*, that was also cloned into pMal-Flag vector, was purified using a two-step protocol (section M.11.2). The use of two affinity columns (amylose and heparin) leads to the obtention of a high-purity protein (Figure R2.10). As describe for *S. hofmanni* Cas2, VvYJ016 Cas2A and Cas2B encoding genes were also cloned into pET16b vector (Table M5). However, no His-tagged Cas2A protein was obtained after the purification process with a His-Trap column (data not shown). Thus, in order to increase the solubility of both proteins, Cas2A and Cas2B encoding genes were cloned in the pMal-Flag vector as described previously for the different RTs used in this study (table M5; section R2.2). Both MBP-Cas2 proteins were also purified following an equivalent two-step protocol, yielding Cas2A and Cas2B proteins of over 95% purity (Figure R2.10).

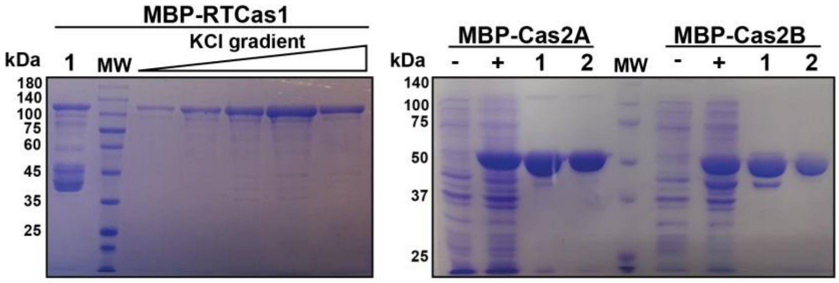


Figure R2.10 Purification process of proteins of the VvYJ016 adaptation module. The process consists of two chromatography affinity steps: amylose column (1) and heparin column (2) (section M.11.2). The MBP-RTCas1 (124 kDa) purification process is shown with a Coomassie Blue staining of a 10% SDS-PAGE gel. After a first step with an amylose column (1), a second step through a heparin column was performed. The MBP-RTCas1 is eluted using a KCl gradient (0.5-1.5 M) and the different fractions are collected. The MBP-Cas2A (57 kDa) and MBP-Cas2B (55kDa) follow a similar process as indicated in the 15% SDS-PAGE gel: total protein (-), induction of the recombinant protein with IPTG (+), eluted protein after the amylose column step (1) and eluted protein after the heparin column step (2).

In addition, the fusion proteins produced using pMal-Flag vector also contain a Xa Factor protease recognition sequence (Ile-Glu/Asp-Gly-Arg), which is located between the MBP and the Flag region of the recombinant protein and, therefore, is used to remove the MBP-tag. However, this protease cannot be used with RTCas1, since this protein also contains the above-mentioned recognition sequence. To avoid this issue, the Xa Factor recognition sequence was replaced in the pMal-Flag vector by the recognition sequence of the HRV 3C protease (Leu-Glu-Val-Leu-Phe-Gln-Gly-Pro), with higher specificity. Upon the cloning of *RTcas1, cas2A* and *cas2B* genes in the new vector, the cleavage efficiency of the 3C protease was tested on different protein samples. A protease gradient (0.05 to 2 units) over a 10 µg MBP-RTCas1 sample revealed that only 0.4 units of the protease are enough to cleavage more than the 95% of the sample (Appendix B5.A). Similarly, 1 unit of the enzyme is sufficient to cleavage more than 90% of a 20 µg MBP-Cas2A or MBP-Cas2B sample (Appendix B5.B). Nevertheless, the removal of the MBP-tag greatly decreases the solubility of all proteins, resulting in protein aggregates, particularly in the case of Cas2A (data not shown).

To explain aggregates formation, one hypothesis could be that the proteins are not correctly folded and only remain soluble due to the presence of the MBP-tag. To test this possibility, the oligomeric state of MBP-RTCas1 was analyzed by size exclusion chromatography (SEC) (section M.11.3). A sample of the purified MBP-RTCas1 was loaded and run into the chromatography column revealing that all protein elutes before the void volume (Appendix B6). This finding could suggest that the MBP-RTCas1 is forming and soluble aggregate that may block the formation of an integrase complex together with Cas2A and Cas2B. It is important to note that in this state the protein also present exogenous RT activity as described above (Figure R2.3B). Different strategies were carried out to avoid the formation of the MBP-RTCas1 soluble aggregate, including the use of different bacterial strains, protein-tags or buffers among other variants of the purification process. However, in all the conditions tested the protein still eluted before the void volume (data not shown).

On the other hand, Cas2A and Cas2B solubility problems might be solved by co-expression of both proteins simultaneously. Thus, a plasmid containing an MBP-tagged *cas2A* followed by a His-tagged *cas2B* were constructed using the pMal-Flag-3C vector as backbone (Table M5). Upon protein overexpression using IPTG, Cas2A and Cas2B were firstly purified using an amylose column, that leads to the co-purification of MBP-Cas2A and His-Cas2B (Figure R2.11), showing a first evidence of interaction between both proteins. Further purification of both Cas2 using a His-Trap column also demonstrate that MBP-Cas2A and His-Cas2B eluted together (Figure R2.11), supporting the close interaction of both proteins.

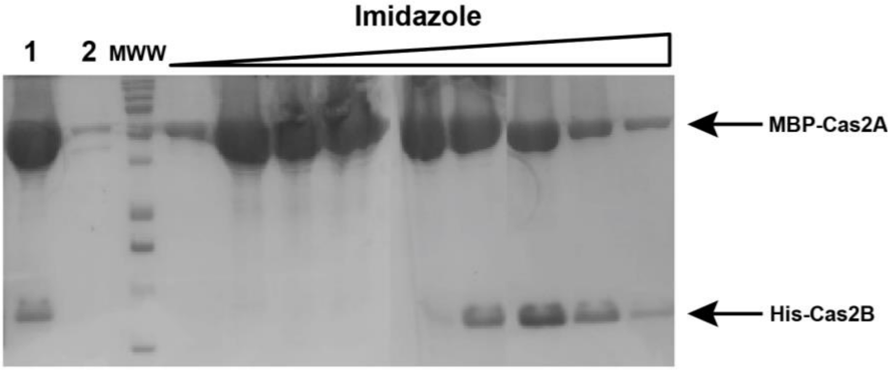


Figure R2.11. Purification process of MBP-Cas2A-His-Cas2B. SDS-PAGE electrophoresis (15%) with the purification process of the MBP-Cas2A-His-Cas2B complex. The use of the amylose columns results in the elution of both proteins after the addition of maltose (1). This protein is further purified with a His-Trap column. The flow-through of His-Trap column is shown (2). The elution with an imidazole gradient (0 to 1 M) reveals that several fractions contain MBP-Cas2A and His-Cas2B, suggesting protein interaction.

Finally, the fractions containing MBP-Cas2A and His-Cas2B after the two purifications steps were pooled and loaded into a gel filtration column to determine their oligomeric state. The elution profile reveals that MBP-Cas2A and His-Cas2B constitute a heterodimer unit of 67 kDa (Figure R2.12). Indeed, despite a excess of MBP-Cas2A is found forming a homodimer (110 kDa), most Cas2A protein is part of the heterodimer (Figure R2.12), suggesting that a Cas2A-Cas2B complex is the most stable state of these proteins. The heterodimer-containing fractions (D3 to D9 in Figure R2.12) were pooled and stored to perform some of the *in vitro* spacers acquisition assays that will be describe in the following section.

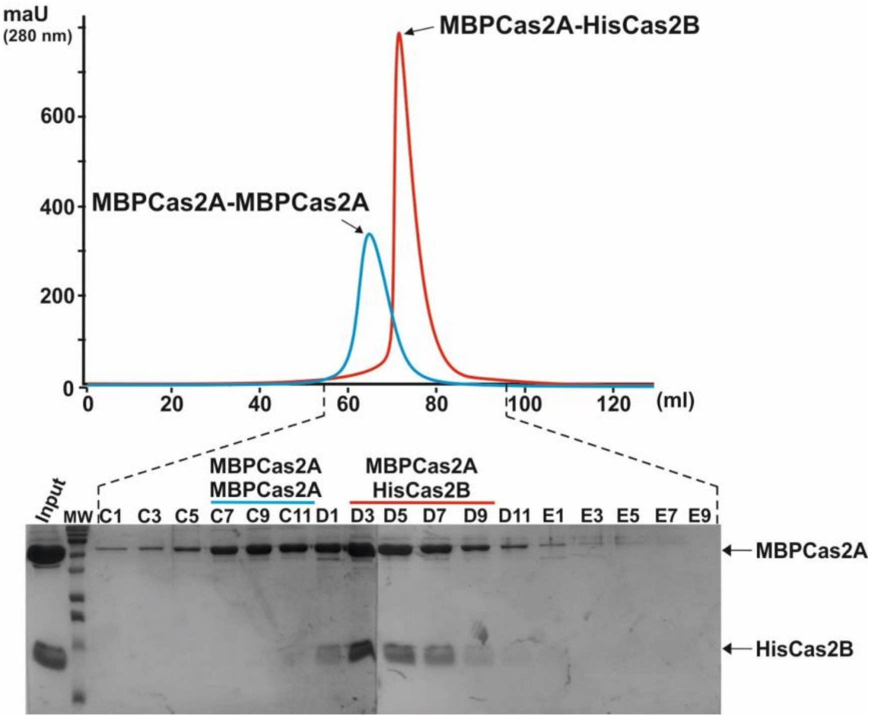


Figure R2.12 VvYJ016 Cas2A and Cas2B form a heterodimer unit. Elution profile of Cas2A-Cas2B co-purification by Superdex Hiload 16/60 gel filtration showing Cas2A-Cas2B complex formation (67 kDa) and its separation from MBP-Cas2A dimer (110 kDa), visualized by Coomassie blue staining of fractions analyzed by SDS-PAGE gel (15%).

The detection of the Cas2A-Cas2B complex after co-expression of both proteins using a single plasmid suggest the possibility that a similar approach may be used to purify the entire YJ016 adaptive module, including the RTCas1 protein. With this aim, a plasmid containing the three proteins was constructed (MBP-tagged RTCas1, Cas2A and a His-tagged Cas2B was constructed (Table M5). Thus, to test whether the three proteins interact, MBP-RTCas1 could be used as bait to purify both Cas2 and then His-Cas2B to further purify MBP-RTCas1 and Cas2A. The co-expressed proteins were loaded firstly in an amylose column and then in a His-Trap-column. The elution of the proteins bound to the latter column using an imidazole gradient reveals that MBP-RTCas1, Cas2A and His-Cas2B are present in the same fractions indicating the potential formation of a complex (Figure R2.13A). Indeed, His-Cas2B was detected in either samples, after induction or after the entire purification process, even though an additional size exclusion chromatography (SEC) step was used (Figure R2.13B) Although its presence cannot be verified, these results suggest that Cas2A is also present in the samples. Despite the fact that the three proteins are co-purified using this protocol, the SEC revealed that the MBP-RTCas1, Cas2A and His-Cas2B are eluting before the void volume as a soluble aggregate (data not shown) and, therefore, the specific RTCas1-Cas2A-Cas2B complex could not be confirmed.

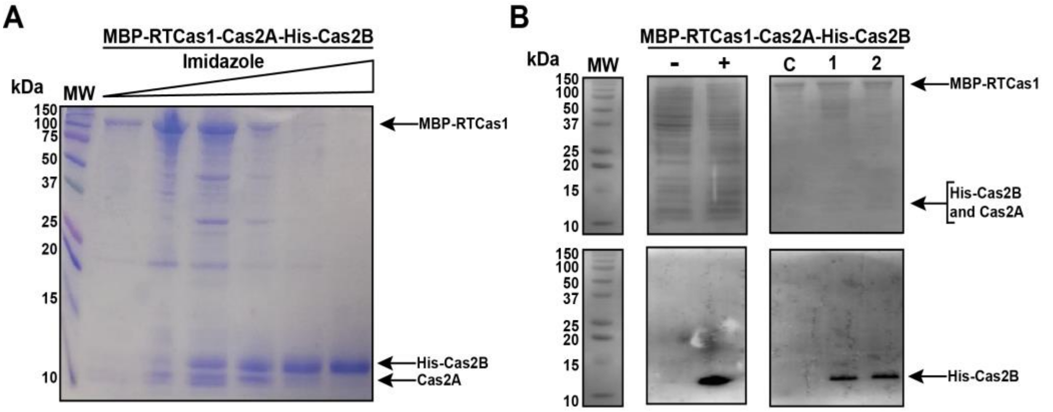


Figure R2.13 Purification of VvY016 adaptive operon. (A) SDS-PAGE gel (15%) stained with Coomassie blue showing the elution profile of MBP-RTCas1, Cas2A, His-Cas2B co-purification after an imidazole gradient in an His-Trap Column (a previous purification step in an amylose column was carried out). **(B)** Western-blot showing that His-Cas2B is co-purified together with MBP-RTCas1. His-Ca2B is detected after overexpression of the entire operon (- and + lanes). His-Cas2B is also detected in fractions containing the MBP-RTCas1 after amylose and His-Trap columns (1) and after an additional size exclusion chromatography step using Superdex 200 Increase 10/300 GL (2). MBP-RTCas1 alone is purified following the same protocol and use as control (C).

The finding that Cas2A and Cas2B form a stable heterodimer unit together with the fact that the entire adaptive operon could be co-purified suggest that the VvYJ016 adaptive module might be completely functional. In order to demonstrate the activity of the potential complex, *in vitro* spacer integration assays (see section M.12.3) will be carried out *in vitro* by mixing the different protein samples purified in this section.

R.2.4.3 Spacer acquisition assays in vitro reveals the presence of a non-specific nuclease activity

To determine the requirements for the integration of new spacer into the CRISPR Array, RTCas1, Cas2A and Cas2B proteins purified as described in the previous section were tested using *in vitro* spacer integration assays. With this aim, the proteins were mixed with a synthetic protospacer (34-nt dsDNA) and a target plasmid consisting of the pGEM-T Easy backbone with the CRISPR Array 2 of *V. vulnificus* YJ016 (pCRISPR-439; table M7) (Figure R2.14A; section M.12.3.1). The successful protospacer integration into the target plasmid generates two main products: linear (full-site integration), in which the protospacer is totally integrated into the plasmid, and relaxed (half-site integration), when one transesterification reaction has occurred and, therefore, the plasmid is only nicked in one of the strands (Figure R2.14A). Thus, protospacer integration could be monitored by plasmid topology. As previously described for type I and II CRISPR-Cas adaptation modules (Nuñez *et al.*, 2014; Wright and Doudna, 2016; Grainy *et al.*, 2019), half-site and full-site product formation requires the presence of magnesium together with the protospacer and all proteins (MBP-RTCas1 and MBP-Cas2A-HisCas2B mixed and incubated for 1 hour at 4ºC) (Figure R2.14B). No improvement in the reaction is detected after the addition of 3C protease to the protein mix in order to remove the MBP-tags and facilitate complex formation (Figure R2.14B).

However, RTCas1 alone show a similar activity on the plasmid as that generated by all mixed proteins (Figure R2.14C). Cas2A and Cas2B (alone or together) cause plasmid nicking detectable by agarose gel electrophoresis (Figure R2.14C). Thus, the products detected when all proteins are present might be the result of the synergistic activity of the individual proteins added to the reaction. Furthermore, the observed activity is independent of protospacer addition (Figure R2.14C). This fact could be explained by the presence of contaminant nucleic acids in protein sample preparations that can also serve as a substrate for integration into the plasmid.

Firstly, to analyze whether the activity detected in the above integration assays was specific of RTCas1, a mutant in the Cas1 active site (E597A) was purified following the two-step protocol (amylose and heparin column) described in the previous section. The purity level of both MBP-RTCas1 proteins, Wild-Type (WT) and E597A mutant, was highly similar (Appendix B7). An integration assay comparing the nuclease activity of both proteins reveals that wild-type Cas1 domain is required for product formation (Figure R2.15A).

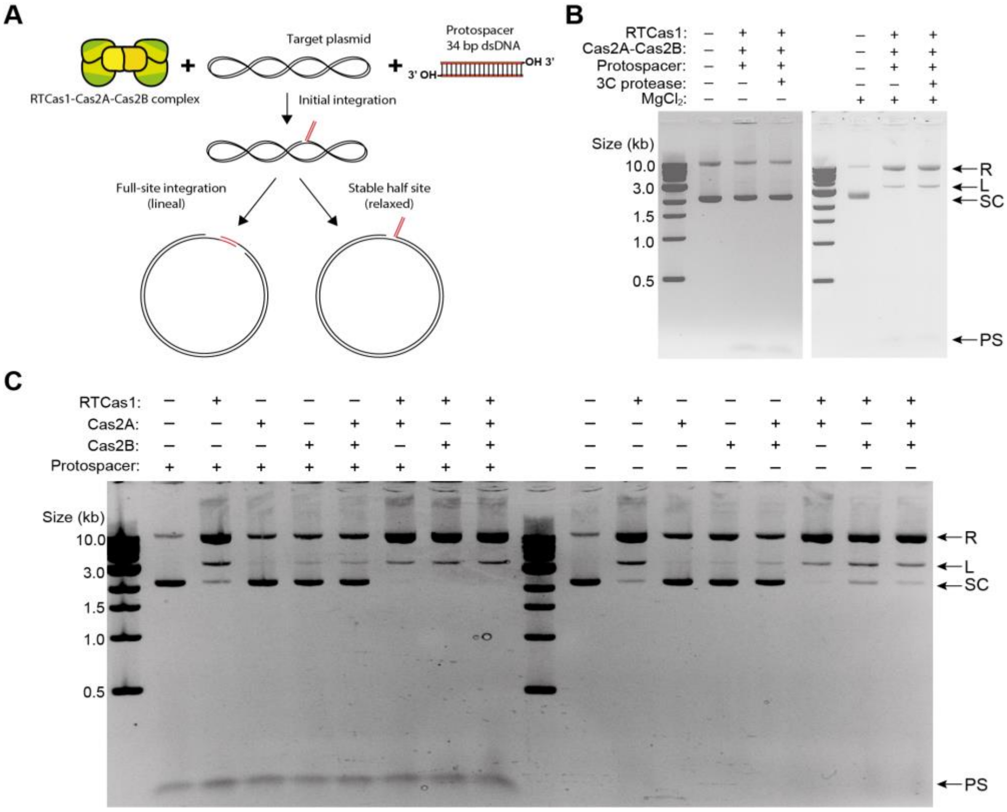


Figure R2.14 Analysis of spacer integration assays *in vitro*. (A) Scheme of RTCas1-Cas2A-Cas2B integration into a supercoiled plasmid (pCRISPR-439) using a 34-nt dsDNA protospacer. Integration of the protospacer can yield to a full-site integration (linear plasmid) or stable half-site integration (relaxed plasmid). **(B)** Nuclease activity detected in integration assays is metal-dependent. MBP-Cas1 and MBP-Cas2A-His-Cas2B purified proteins were mixed an incubated with the protospacer (section M.12.3.1). 3C protease was added to the reaction to remove MBP-tags. A nuclease activity is observed after the addition of magnesium. **(C)** Nuclease activity detected in integration assays is not dependent on the addition of protospacer. The activity of MBP-RTCas1, MBP-Cas2A and MBP-Cas2B (alone or mixed) over pCRISPR-439 are shown. Products are separated by agarose gel electrophoresis (1%). Relaxed (R), linear (L), and supercoiled (SD) pCRISPR-439, as well as protospacer (PS) are indicated.

On the other hand, since nucleic acids were detected in purified MBP-RTCas1 samples (Appendix B8), a new purification using an additional polyethylenimine (PEI) step in order to remove nucleic acids was performed. The comparison of MBP-RTCas1 samples purified with and without PEI shown that product formation was undetectable when no nucleic acids are present in the sample (Appendix B.7.B). Thus, the purified MBP-RTCas1 using PEI was tested in a new integration assay which reveals an increment of half and full-site products when all proteins (MBP-RTCas1 and MBP-Cas2A-His-Cas2B heterodimer) were mixed (Figure R2.15B). Nevertheless, no integration products were detected using a radiolabeled protospacer to confirm specific integration events (Figure R2.15B; section M.12.3.2). The results of cleavage/ligation assays (M.12.3.3), in which MBP-RTCas1 is used alone or mixed with MBP-Cas2A and/or MBP-Cas2B on an internally radiolabeled substrate containing the leader sequence, the first two direct repeats and the first two spacers, also shown a band smear (Appendix B9). This finding supports the existence of a non-specific nuclease activity, which could be derived from RTCas1 or from a putative contaminant present in the protein samples.

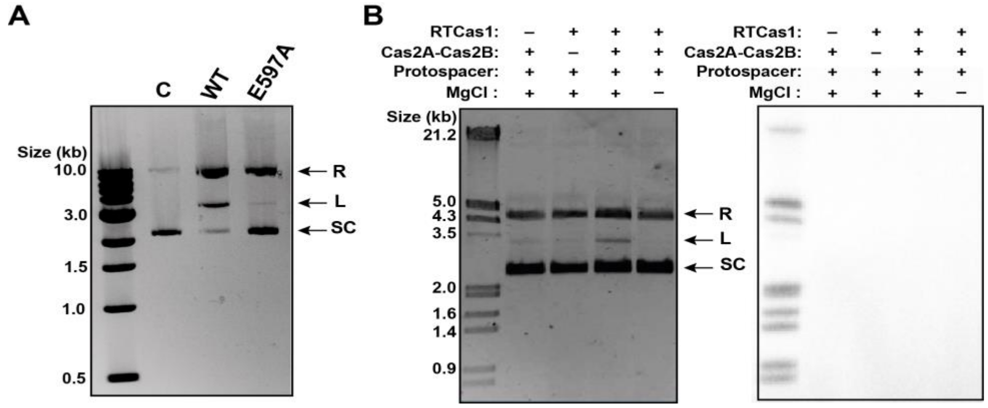


Figure R2.15 Non-specific nuclease activity detected in *in vitro* integration assays. (A) Integration assay comparing the nuclease activity of MBP-RTCas1 wild-type (WT) versus MBP-RTCas1 mutant in the Cas1 domain (E597A) samples over pCRISPR-439. A negative control (C) is also shown. **(B)** Integration assay with unlabeled (left) and labeled (right) protospacer on pCRISPR-439 using MBP-Cas2A-HisCas2B, MBP-RTCas1 or a mix of the three proteins (incubated 1 hour at 4ºC). The products detected by GelRed (left) are not detected by autoradiography (right). Products are separated by agarose gel electrophoresis (1%). Relaxed (R), linear (L), and supercoiled (SD) pCRISPR-439, as well as protospacer (PS) are indicated.

Chapter 3: Spacer acquisition and interference in the type III-D CRISPR-Cas system from **Vibrio vulnificus** *YJ016*

R.3.1 Background

The demonstration that the *in vivo* acquisition of RNA molecules is facilitated by the RT domain of a Cas6RTCas1 fusion protein associated with a type III-B CRISPR-Cas system present in the genome of the marine bacterium *Marinomonas mediterranea* (MMB-1) has raised the attention on this particular subtype of CRISPR-Cas systems (Silas *et al.*, 2016). This RT-dependent acquisition of RNA spacers has been shown to occurs through a mechanism displaying several similarities to group II intron retrohoming mechanism (Silas *et al.*, 2016). More recently, an RTCas1 fusion protein linked to a type III-D system found in the bacterium *Fusicatenibacter saccharivorans* was shown to acquire RNA spacers efficiently in *E. coli* host. Furthermore, this adaptive module has been used in a novel biotechnological application, termed Record-seq, in which the RT-containing adaptation machinery works as a transcriptional recorder, describing both continuous and transient complex cellular behaviors (Schmidt *et al.*, 2018). As described in Chapter 1, a total of 15 clades of RTs associated with CRISPR-Cas systems have been found to date, 13 of which evolved from group II intron. The two RT domains from *M. mediterranea* and *F. sacharivorans* belongs to clades 8 and 12, respectively (Figure R1.10).

Since only two examples of RT-CRISPR systems has been extensively characterized, the aim of this chapter is to study the acquisition of novel spacers carried out by the adaptation machinery linked to a type III- D system in *Vibrio vulnificus* YJ016 (Figure R2.8). RTCas1 fusion protein present in the adaptation module of this Gammaproteobacteria belongs to clade 6 of the phylogeny of RTs associated with CRISPR-Cas systems and it has been shown to present exogenous RT activity (Figure R2.3B). Furthermore, in the adaptation loci of clade 6 systems is characteristic the presence of two different Cas2 (Cas2A and Cas2B), which has been demonstrated to form a heterodimer (Figure R2.12). Thus, the *in vivo* characterization of this particular type of RT-containing adaptation module could reveal novel properties and lead to expand the CRISPR-Cas toolbox.

R.3.2 *In vivo* spacer acquisition by the *Vibrio vulnificus* YJ016 adaptation module

To investigate whether the adaptation operon of *V. vulnificus* YJ016 is able to acquire new spacers in a heterologous host (*E. coli*) both the adaptation module and the two CRISPR Arrays were cloned in different vectors (Table M5). On one side, the adaptation machinery of *V. vulnificus* YJ016 (VvYJ016), comprised by the RTCas1 fusion and the two Cas2 proteins, Cas2A and Cas2B, were cloned under T7 promoter in a pGEM-T Easy vector (Table M4). On the other side, a compatible pMP220 vector (Table M4) was used to clone a reduced version of both CRISPR Arrays 1 and 2 (hereafter referred as CRISPR01 and CRISPR02 arrays, respectively) containing the leader sequence, the first direct repeat (DR) and the first spacer. After co-transformation with the two plasmids, spacer acquisition in *E. coli* was assessed by overexpressing the RTCas1, Cas2A and Cas2B operon (Figure R3.1A; section M.13.2.1). New spacer acquisition by the two arrays was evident after two rounds of PCR purification of the expanded array band, corresponding mostly with an acquisition event (Figure R3.1A and B; section M.13.2.2). Finally, the purified expanded band was prepared and sequenced with Illumina-MiSeq (Figure R3.1A).

A data processing pipeline was designed to analyze the reads corresponding with spacer acquisition assay (section M.13.2.3). Briefly, the reads were trimmed to obtain the sequence of the spacers between the two direct repeats. Then, spacers were grouped based on unique start and end coordinates, and, finally, mapped on the used plasmids and the *E. coli* HMS 174 (DE3) genome. The results of this pipeline reveal that most acquired spacers derived from the *E. coli* genome (~95%) with the rest being derived from plasmids DNA (~5%) (Appendix C1). Thus, the experimental procedure designed in the present work serves to demonstrate that the CRISPR-Cas adaptation module from *V. vulnificus* YJ016 is capable of acquiring new spacers in a heterologous host. Curiously, comparing the number of novel spacers acquired in both CRISPR Arrays it is important to note that spacer acquisition rate is 20 times higher in CRISPR02 than in CRISPR01 array (Figure R3.2A and B).

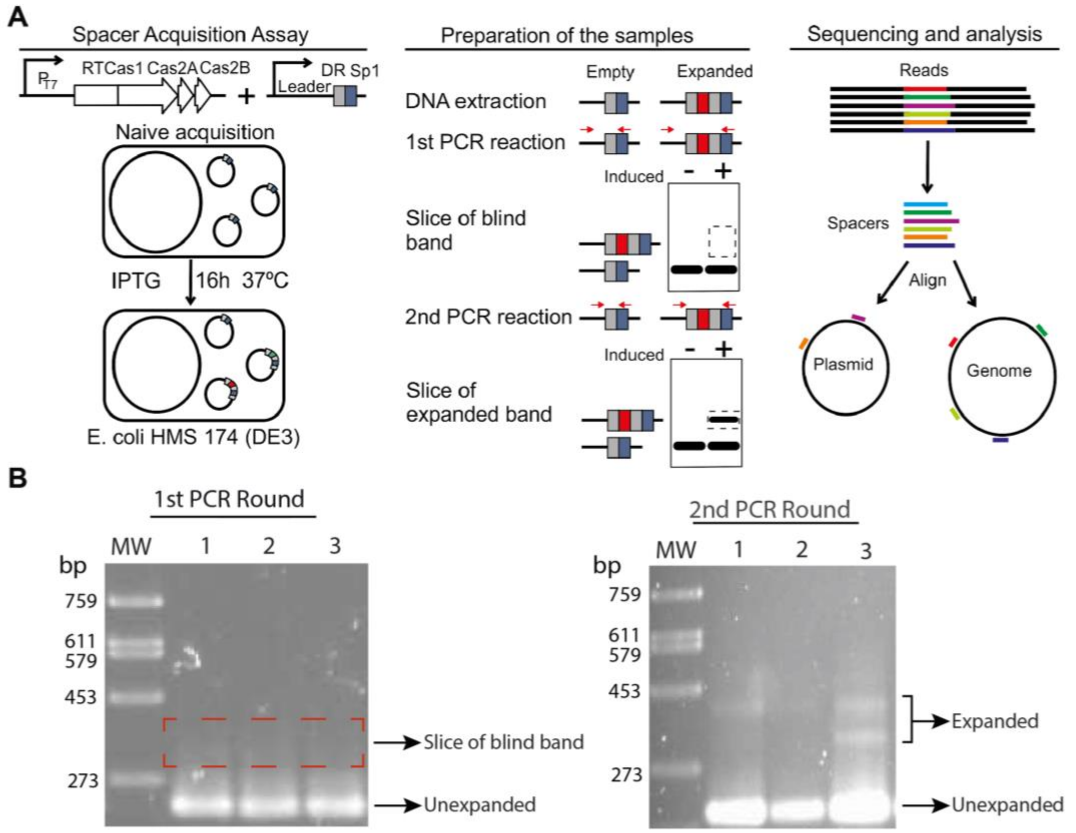
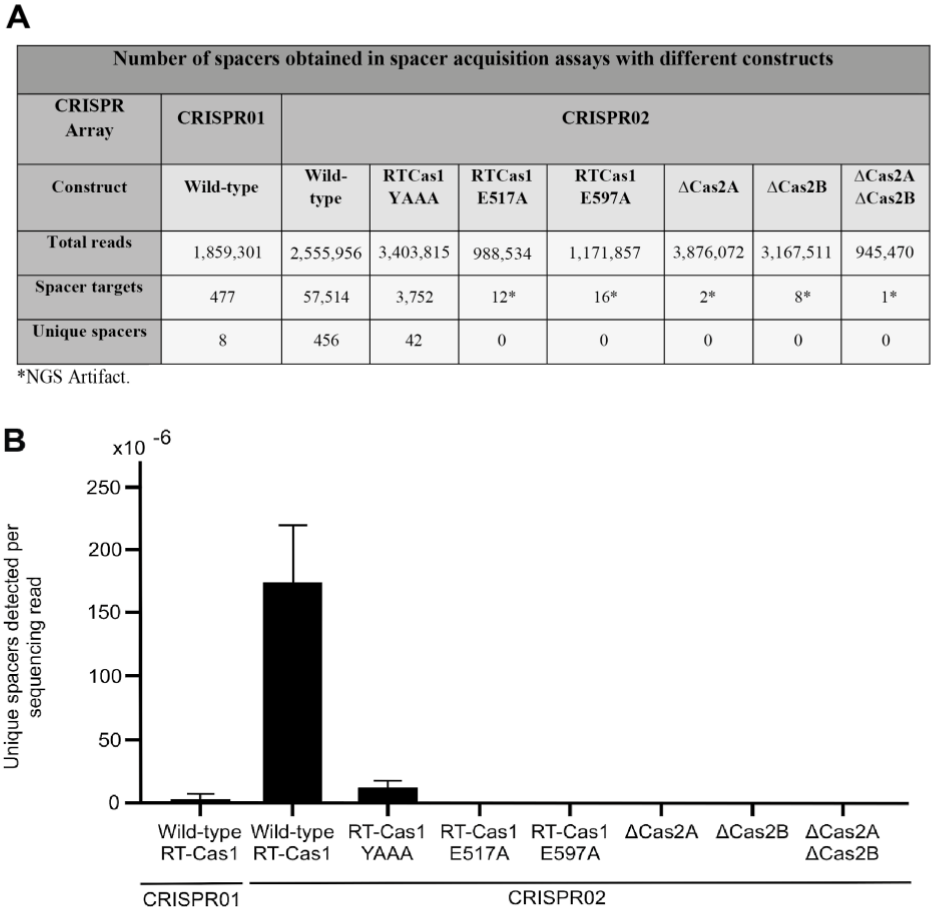


Figure R3.1 Spacer acquisition assay using the VvYJ016 adaptation module. (A) Schematic diagram of the high-throughput spacer acquisition assay. Overexpression of the adaptation operon in *E. coli* HMS 174 (DE3) followed by the extraction of plasmid DNA, two rounds of PCR/purification of the expanded CRISPR Array, and Illumina-MiSeq sequencing, analysis and characterization of the spacers identified (section M.13.2). **(B)** Expanded CRISPR Array purification process. Representative examples of 2% agarose gel from the first and second PCR round of three independent biological replicates of spacer acquisition assay. The red dashed box in the 1st PCR round gel indicate the size of the band sliced and used as substrate for the 2nd PCR round. Expanded and unexpanded bands are indicated.

The nucleotide sequence of the leader region and the characteristic of both CRISPR Arrays was analyzed in detail to understand the differences observed in the acquisition process. Although both arrays share the same direct repeat (35 nt), slight differences are observed in the leader region (Figure R3.3A). To analyze whether the differences observed in these regions affect the promoter activity of the leader sequences, a β-galactosidase assay was designed (sectionM.13.1) to evaluate this activity in *E. coli* using a transcriptional fusion of CRISPR01 and CRISPR02 with a *lacZ* reporter gene (Table M7).



Number of spacers obtained in spacer acquisition assays with different constructs								
CRISPR Array	CRISPR01	CRISPR02						
Construct	Wild-type	Wild-type	RTCas1 YAAA	RTCas1 E517A	RTCas1 E597A	ΔCas2A	ΔCas2B	ΔCas2A ΔCas2B
Total reads	1,859,301	2,555,956	3,403,815	988,534	1,171,857	3,876,072	3,167,511	945,470
Spacer targets	477	57,514	3,752	12*	16*	2*	8*	1*
Unique spacers	8	456	42	0	0	0	0	0

*NGS Artifact.

Figure R3.2 Spacer acquisition by the VvYJ016 adaptation operon in the heterologous *E. coli* system. (A) Summary of total reads, total spacers and unique spacers detected in the spacer acquisition assays carried out with the different plasmids constructed. **(B)** Frequency of new spacer detection per million reads for the RT-Cas1-Cas2A-Cas2B wildtype operon for both CRISPR01 and CRISPR02, RT active site mutant (YAAA), Cas1 domain mutants E517A and E597A and the ΔCas2A, ΔCas2B and ΔCas2A-B mutants. All point and defective mutants were assayed only for the CRISPR02. The bars indicate the range for three biological replicates.

The results of the β-galactosidase assay show only a constitutive expression for the leader sequence of the CRISPR01 (Figure R3.3B). This data is consistent with the transcriptional RNA-seq data for this *V. vulnificus* strain in which the transcription rates for the CRISPR01 array are several times higher than those for the CRISPR02 array in two different conditions assayed (Figure R3.3C; Williams *et al.*, 2014).

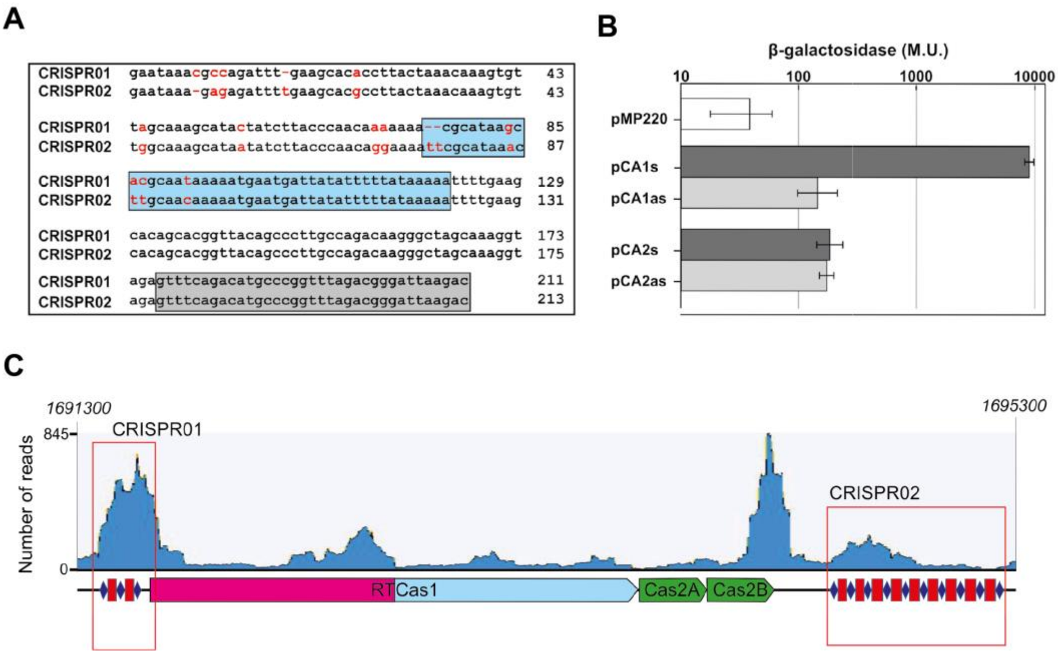


Figure R3.3 Promoter activity of VvYJ016 CRISPR01 and CRISPR02 arrays. (A) Alignment of the leader sequence and first direct repeat of the VvYJ016 CRISPR01 and CRISPR02 arrays. A gray box indicates the nucleotide sequence of the first repeat. A blue box shows the putative promoter of the two sequences. The red letters represent the bases differing between the two sequences. **(B)** Determination of the level of transcription of the leader sequence of CRISPR01 and CRISPR02. β-galactosidase activity (Miller Units) was measured for the empty plasmid (pMP220) and the two complete arrays, in both orientations with respect to the *lacZ* gene (sense: pCA1s; pCA2s for CRISPR01 and CRISPR02, respectively; antisense: pCA1as; pCA2as for CRISPR01 and CRISPR02, respectively). The standard deviation for three biological replicates with errors bars. **(C)** Expression of CRISPR01 and CRISPR02 in RNA-seq data from *Vibrio vulnificus* YJ016. Coverage of reads from RNA-seq data obtained from Williams *et al.*, 2014 against coordinates 1691300-1695300 from Chromosome II (NC_005140). This region contains both arrays (framed in red) together with the adaptive module. CRISPR01 is expressed several times more strongly than CRISPR02. *RTCas1* is shown in pink (*RT* domain) and blue (*cas1* domain), and the two different *cas2* (*cas2A* and *cas2B*) genes are shown in green. Direct repeats (DRs) are indicated in purple and spacers in red.

Due to the higher acquisition rate observed in CRISPR02, this array was selected to perform a more extensive study of the adaptation process in order to analyze the role of the different proteins/domains involved in the acquisition of novel spacers in the adaptive module of VvYJ016. Firstly, to evaluate the function of both RT and Cas1 domains of the RTCas1 fusion protein in spacer acquisition, a series of mutants were constructed. On one side, in the RT motif 5 (YADD) responsible for the reverse transcriptase activity, the two catalytic aspartic amino acid residues were substituted by two alanine residues, generating the YAAA mutant. On the other side, a mutant in the active site of the Cas1 domain was constructed, the E597A mutant, in which a catalytic glutamic residue is substituted by an alanine residue. The RT *in vitro* assays carried out (section M.12.1) show that the YAAA mutant lacked RT activity, whereas the E597A mutant had a level of RT activity comparable to that of the wild-type protein (Figure R3.4).

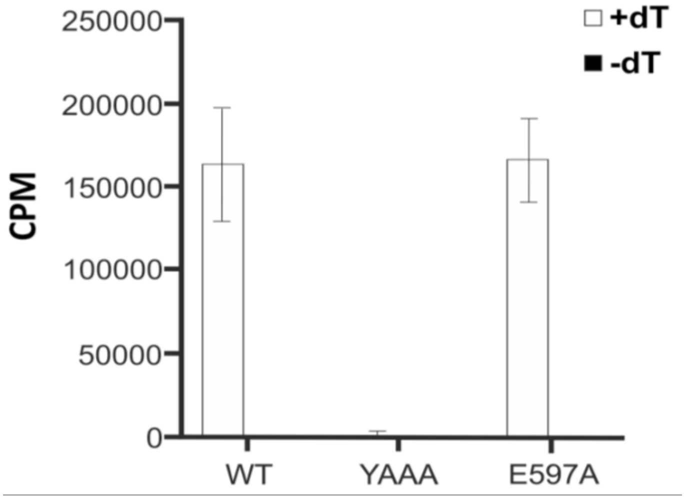


Figure R3.4 Exogenous RT activity and spacer acquisition of wild-type RTCas1 and mutants. 200 ng of Wild-type (WT) and mutant RT-Cas1 proteins were analysed for RT activity by assessing the polymerization of radiolabelled [α-32P]-dTTP for 10 minutes with either the artificial template-primer substrate poly(rA)/oligo $[(dT)]_{18}$ or the template poly(rA) alone (-dT). The bar graphs show RT activity, measured as counts per minute (CPM). Three independent protein preparations were assayed. Mutation of the RT active site abolished RT activity (YADD to YAAA), whereas the E597A Cas1 mutant displayed levels of RT activity similar to those of the wild-type RTCas1.

In addition, after protein purification, wild-type and both mutants yield similar amount of fusion protein, indicating that no effect of the mutation during the protein production process (Appendix C2.A). After analyzing the spacer acquisition rates of the distinct mutants in comparison with the wild-type, it was shown that the mutation in the RT active site decreased the acquisition of new spacers by ~ 90%, while the mutation of the Cas1 domain abolished spacer acquisition (Figure R3.2A and B; Appendix C2.B). These findings reveal that the catalytic activity of the RT domain is important for *in* vivo spacer acquisition in the heterologous *E. coli* host. On the other hand, the Cas2A and Cas2B requirement for the formation of an active acquisition complex *in vivo* was investigated. For this aim, deletion mutants of Cas2A, Cas2B or both were constructed. In all the cases, the deletion completely abolished the acquisition of new spacers, proving that both Cas2 proteins were required for a functional VvYJ016 adaptation module (Figure R3.2B).

This data, together with the finding that Cas2A and Cas2B form a heterodimer complex as show in the previous chapter (section R.2.4.2) suggest that the Cas2 dimer which forms part of the integrase complex required for *in vivo* spacer acquisition is formed by two different Cas2 proteins in VvYJ016 CRISPR-Cas system. Indeed, these arguments are supported by the fact that both proteins present a conserved Cas2 structure as previously described (Figure R2.9).

R.3.3 Features of the spacers acquired by the VvYJ016 adaptation module

To further investigate the particularities of the RT-containing VvYJ016 acquisition machinery, the pool of newly acquired spacers was characterized. Although, the acquired spacers matched regions throughout the host genome, it is worth to highlight that spacers complementary to rRNA genes were the most abundant ones (Figure R3.5A). Additionally, a bias of acquired spacers towards to the antisense orientation of coding sequences was detected, reflecting the complementarity among the newly acquired spacers and the predicted messenger

RNA (Figure R3.5A and B). Even in the absence of the effector module, these findings suggest a relationship between the acquired spacers and transcription.

Consistent with the length distribution of the natural spacers present in these arrays, the average length of the newly acquired spacers ranged between 34 and 38 base pairs (bp) long, being the spacers originating from plasmids were on average 1 bp longer than those spacers with a genomic origin (Figure R3.5C). Furthermore, the median 'GC' content of the spacers correlates with the 'GC' content of the 'template' used, independently this were the plasmid or the *E. coli* genome (Figure R3.5D).

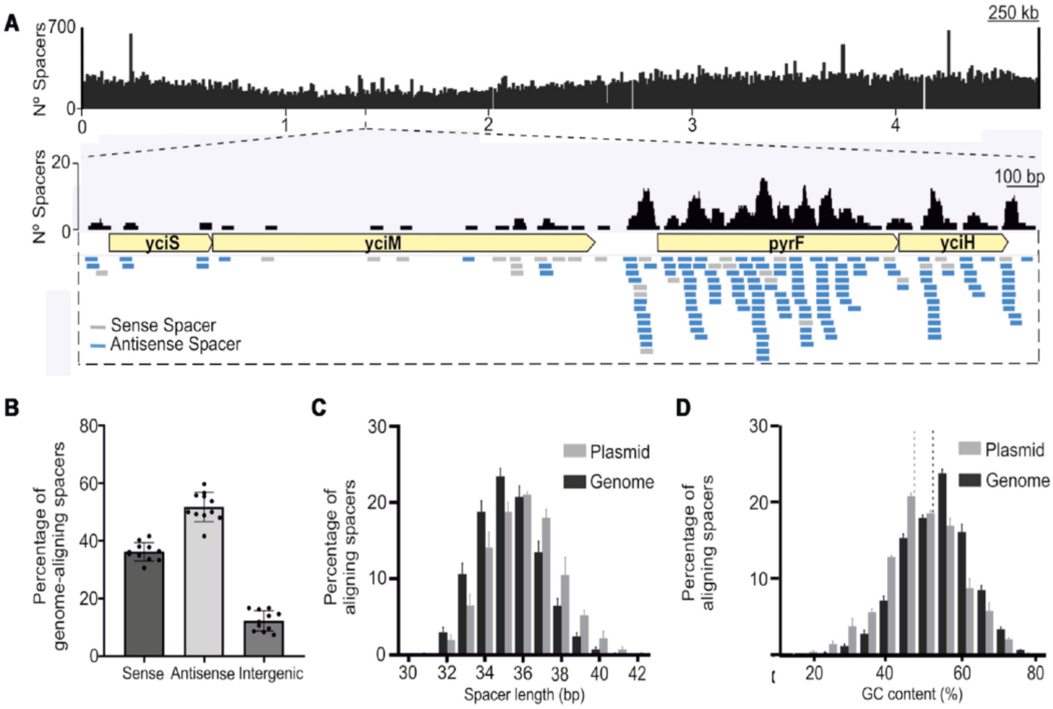
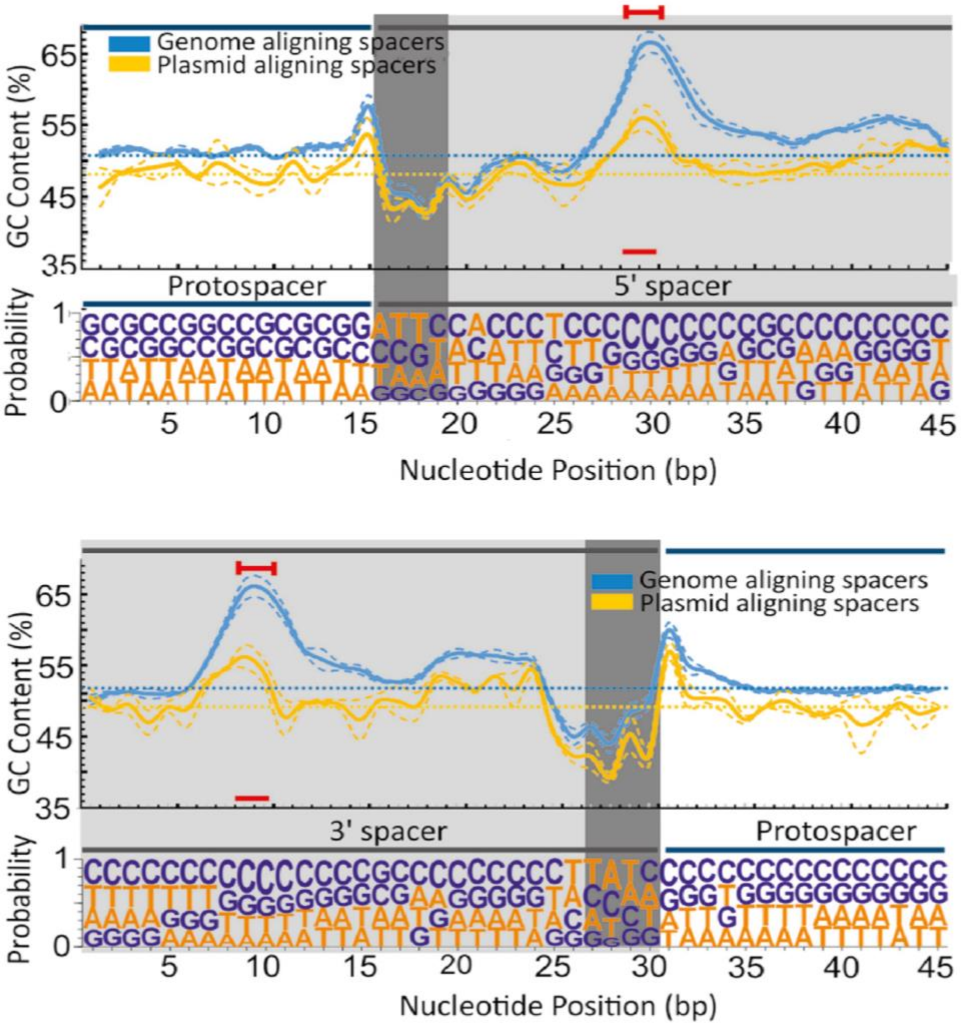


Figure R3.5 Characterization of the spacers acquired by the VvYJ016 adaptation module. (A) Coverage of spacers aligning with the *E. coli* HMS174 (DE3) genome and a representative locus. **(B)** Strand bias in pools of newly acquired spacers relative to the source transcript. Proportion of newly acquired spacers with the Wild-type RTCas1 in the sense or antisense strand of coding genes or in intergenic regions of the *E. coli* genome (*n* = 11). **(C)** Histogram showing normalized counts of *E. coli* genome or pAGDt-Op439 plasmid spacers, by length. **(D)** GC content distribution of genome- and plasmid-aligned spacers. The dotted lines represent the GC content of the plasmid (light gray) and the genome (dark gray). For C and D, the bars indicate the range for the three assays in which the largest numbers of spacers were detected (>10,000 newly acquired spacers per experiment).

Along to the particular characteristics observed in the integrated sequences by the VvRTCas1-Cas2A-Cas2B adaptation machinery, it has been generally observed that spacers acquired by type III CRISPR–Cas systems lack a protospacer-adjacent motif (PAM) (Pyenson and Marraffini, 2017). After analyzing the flanking regions of the protospacer (the original plasmid or genomic sequence where spacers come from) no conserved PAM was observed (Figure R3.6), even if the analysis was based on frequency of the possible pairs of dinucleotides (data not shown, Kieper *et al.*, 2018).



(Figure Legend in the next page)

Figure R3.6 Spacer sequence composition. GC content (above) and nucleotide probabilities (below) at each position along the sequence of the acquired protospacers. Given the variation of protospacer length, two panels are shown, with the spacer anchored 5′and 3′ at positions 15 and 35, respectively. Spacer (gray background) and flanking (white background) nucleotides are shown. The dark gray background indicates an 'AT' rich region at the two ends of the spacers. The particular bias towards 'C' within the spacer observed is indicated by a red line. 'GC' content of the spacers aligning with the genome and the plasmid are shown in blue and yellow, respectively.

Nevertheless, given the vast numbers of spacers obtained in the distinct spacer acquisition assays performed in this study, a significant deviation of the expected 'GC' content was observed at different positions within the spacer sequence. This observed deviation consists in a symmetric bias emerged at the beginning and the end of the spacers, at which an 'AT' rich region was observed stretch of four to five positions. These 'AT'-rich positions contrast with the bias towards 'GC' enrichment observed at the first nucleotide of the protospacer flanking the spacer (Figure R3.6).

A detailed analysis of the sequence of the spacers reveals strong bias towards 'GC' was observed at positions +14 and +15 of the spacer (present at ~67% and ~55% of the spacers acquired from the genome and from the plasmid, respectively). Furthermore, it is worth to specially highlight the high frequency of cytosine at these positions, which represents about ~ 40% of the total, when in a normal distribution about 25% should be expected (Figure R3.6). This specific bias was observed not only for spacers originating from the genome, but also for those originating from plasmids, suggesting an inherent preference of the acquisition complex when selecting the sequences that will be integrated into the CRISPR Array.

The relative position of the acquired spacers was analyzed in order to explore the existence of a bias of the newly integrated spacers towards a specific region of an opening reading frame (ORF). However, this analysis reveals that no bias exists in the acquired spacers to any region of the genes within the *E. coli* genome (Figure R3.7). This finding is in contrast with that observed in the acquisition events carried out by the adaptive machinery of *Fusicatenibacter saccharivorans* in the same heterologous host, where newly integrated spacers present a strong bias towards the beginning and the end of the ORFs (Schmidt *et al.*, 2018). Taken together, these findings suggest that several sequence-specific requirements for integration of novel spacers are required by the VvYJ016 adaptation module, regardless of their plasmid or genome origin or their relative position within an ORF.

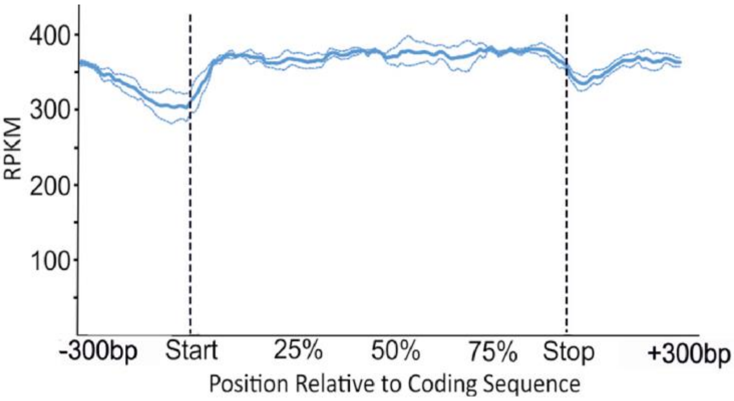


Figure R3.7 Gene body coverage of spacer alignments along the length of transcripts. The relative position corresponds to the percentile of coding sequence length ± 300 bp of the adjacent genomic regions. Dotted lines in A and B represent the mean error of alignment for the three assays in which the largest numbers of spacers were detected (>10,000 newly acquired spacers per experiment).

Contrary to the other RT-containing CRISPR-Cas adaptation modules analyzed to date, in which spacer acquisition assays were carried out in *E. coli* strain BL21 (DE3) or in close relative strains (Silas *et al.*, 2016; Schmidt *et al.*, 2018), in the assays performed in this work the *E. coli* strain used was HMS 174 (DE3). One of the main differences between both *E. coli* strains is that BL21 is a *recA*+ strain, whereas HMS 174 is a *recA*- strain (see Table M1). RecA together with RecBCD is involved in homologous recombination in *E. coli* (Anderson and Kowalczykowski, 1997). Additionally, RecBCD is involved in the production of short ssDNA fragments which have been shown to be the substrate captured by Cas1-Cas2 integrase complex for their subsequence integration within the CRISPR Array (Levy *et al.*, 2015; section I.5.1). Thus, it could be reasonable to suggest that the lack of RecA, which results in DNA repair failures during the homologous recombination process, might also lead to changes in the observed acquisition events after analyzing the results of integration assays.

To further study this hypothesis, a spacer acquisition assay was carried out in order to observe variability related to bacterial host using the following *E. coli* strains: BL21 (DE3), HMS 174 (DE3), and JM109 (DE3) (Table M1). The results of the spacer acquisition assays show a similar rate of novel acquired spacers in the three strains (Figure R3.8A). In addition, the analysis of the acquired spacers reveals no significant differences and although all the strains show a conserved bias towards transcription, slight changes in the percentage of the strand from which the spacers originated were observed (Figure R3.8B). However, more biological replicates would be necessary to confirm these differences.

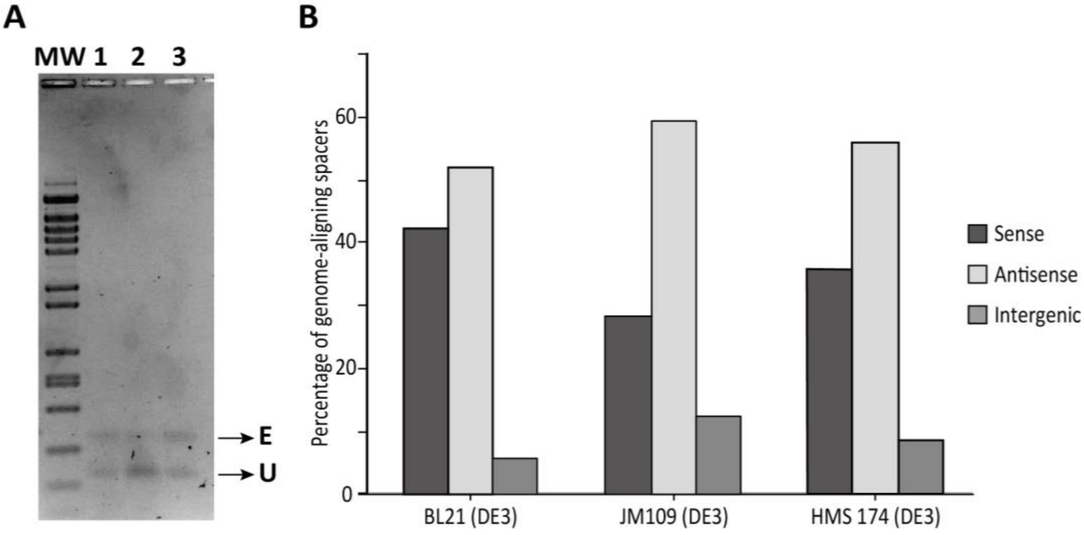


Figure R3.8 Spacer acquisition assay in different *E. coli* strains. (A) 2% agarose gel showing unexpanded (U) and expanded (E) bands after the second round of band purification of the spacer acquisition assay carried out in *E. coli* strains BL21 (DE3) (1), JM109 (DE3) (2), and HMS 174 (DE3) (3). **(B)** Strand bias in pools of newly acquired spacers relative to the source transcript. Proportion of newly acquired spacers with the Wild-type RTCas1 in the sense or antisense strand of coding genes or in intergenic regions of the *E. coli* genome of strains BL21 (DE3), JM109 (DE3), and HMS 174 (DE3).

The minor differences described above could be consequence of host factors involved in the process of acquisition of novel spacers within the CRISPR Array, such as the integration host factor (IHF) involved in type I and II CRISPR-Cas spacer acquisition (Nuñez *et al.*, 2016; Wright *et al.*, 2017). Nevertheless, in type III systems host factors implicated in CRISPR adaptive step has not been described yet.

In order to detect novel host factors that may participate in prespacer generation or integration of new spacers in type III CRISPR-Cas systems a protein-protein interaction approach will be carried out. With this objective, RTCas1 fusion protein was N-terminally tagged with Flag epitope which will allow the purification of RTCas1 directly from *V. vulnificus* YJ016 using a resin containing Anti-Flag antibody (section M.12.2.2). Then, proteins stably associated with Flag-RTCas1 will be identified using liquid chromatography mass spectrometry (LC-MS). The first part of this approach consists in the cloning of Flag-tagged RTCas1 in pVSV-105-OmpC plasmid (Table M9) just downstream of the *ompC* constitutive promoter to maximize protein production. Upon plasmid mobilization to *V. vulnificus* YJ016 (section M.7.3), the use of anti-Flag antibodies has enabled the detection of RTCas1 by western blotting in exponential samples of this strain (Appendix C3; section M.11.6). Thus, these samples will be used to characterize the putative interaction of RTCas1 with other host factors that could perform an important role in the adaptive stage of RT-containing type III CRISPR-Cas systems.

R.3.4. The RT-containing VvYJ016 adaptation module acquires spacers directly from RNA

To fully investigate whether the newly acquired spacers within the CRISPR Array by the VvRTCas1-Cas2A-Cas2B system were originated from RNA molecules, an acquisition assay was designed in order to look for spacers that certainly came from an RNA source. With this aim, a technology based on the self-splicing *td* group I intron (*td*I) was used. This intron is a functional ribozyme which catalyses its own excision from the original transcript resulting in the joining of the exons, a region which is not present in the host DNA (Silas *et al.*, 2016; Schmidt *et al.*, 2018). Thus, as long as a new integrated spacer contains the exon junction it can be stated with confidence that this spacer has an RNA origin. To perform this assay, a plasmid was constructed with *td*I cloned just downstream from the adaptation operon, which is expressed under the control of the T7 promoter (section M13.2.4; Table M7).

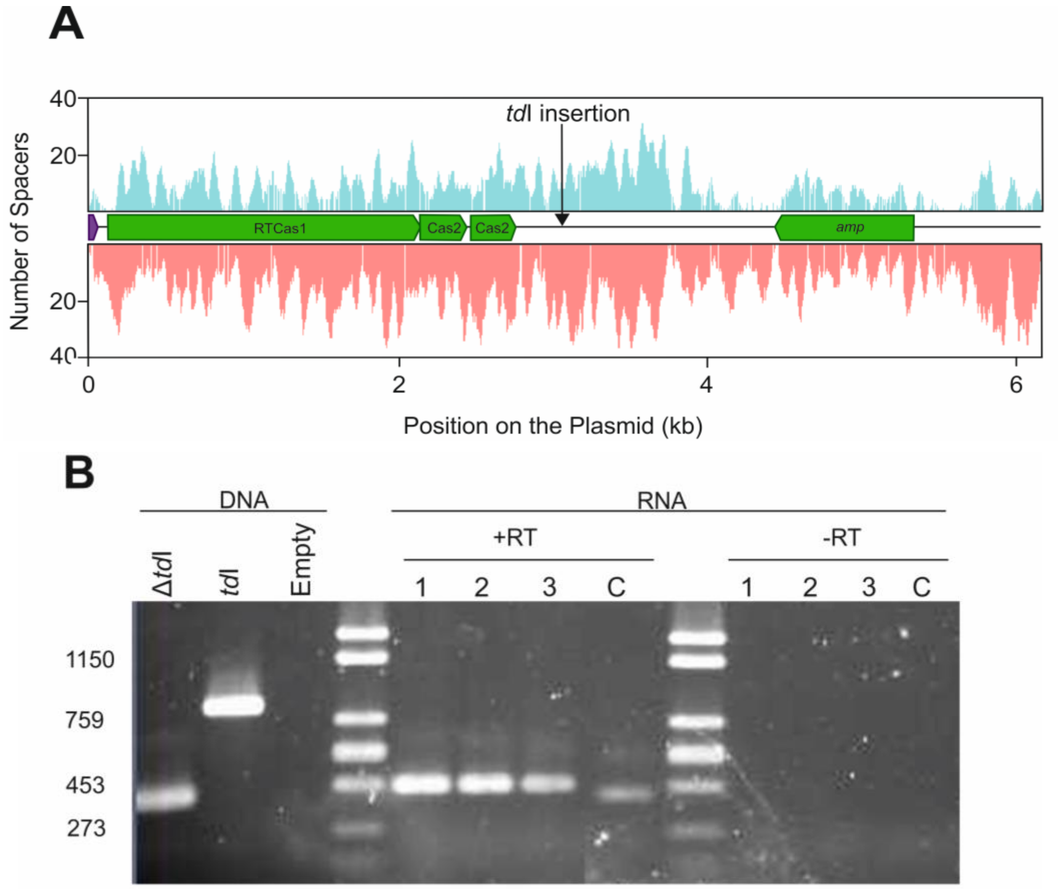
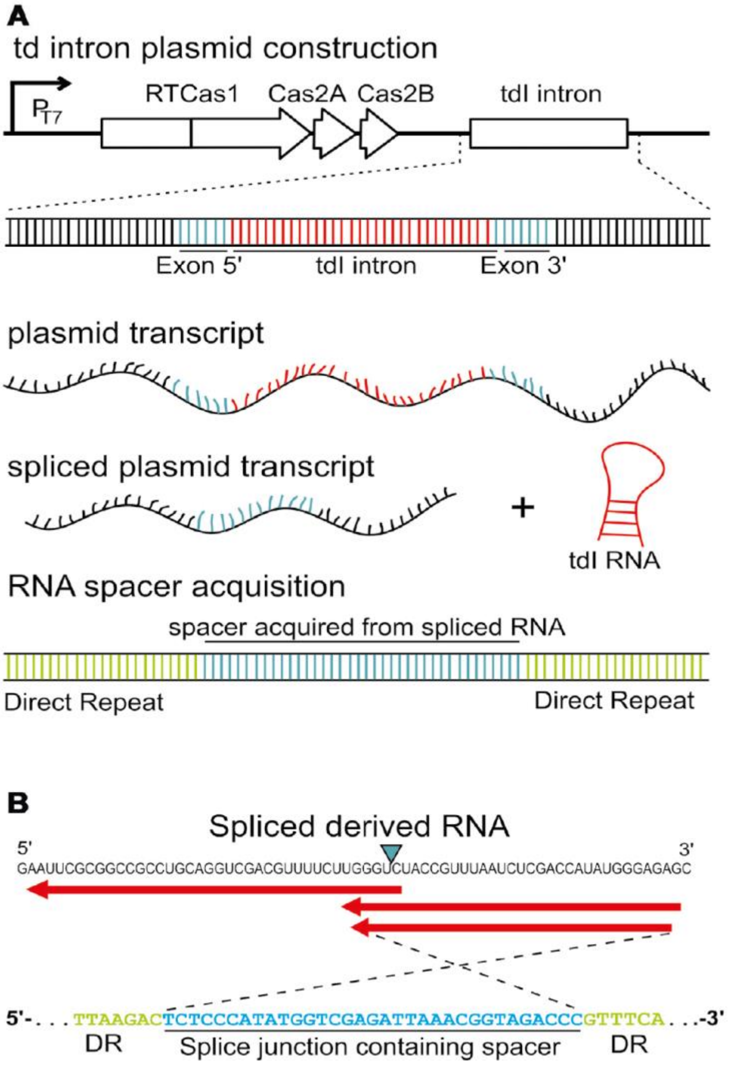


Figure R3.9 Construction and verification of *td*I splicing. (A) Protospacer mapping on the pAGDt-439 plasmid, based on several experiments, suggests that the *td*I insertion site chosen (black arrow) is valid for the acquisition of spacers by the array. Protospacers on the plus and minus strand are indicated in blue and red, respectively. Genes (green) and the T7 promoter (purple) are schematically depicted between the plus and minus graphs. **(B)** Electrophoresis of spliced *in vitro* transcripts from the *td* intron showing highly efficient splicing activity. The DNA samples show the pAGDt-439 vector (Δ*td*I), pAGDt-439- *td*I (*td*I) and the empty vector (Empty) amplified with 439-tdIF and SP6 primers. The RNA samples correspond to three biological replicates of the cDNA or non-reverse-transcribed RNA (+RT or –RT) of the transcript of *td*I (1, 2, 3) and a control with no intron (C).

The position of the introns in this plasmid is optimal to detect spacer originated from RNA, as *td*I is clone in one of the regions from which a large number of spacers were detected in the antisense orientation (bias source transcription) (Figure R3.9A). The sequence corresponding to the exon-junction was checked to see that it was not present elsewhere in the *E. coli* genome. Then, it was confirmed that efficient self-splicing of *td*I occurred *in vivo* (almost 100%) by extracting the RNA from three biological replicates of *E. coli* transformed with the *td*I-containing plasmid (Figure R3.9B; section M.5.3 and M.13.2.5).


A
td intron plasmid construction
P_T7
RTCas1
Cas2A
Cas2B
tdI intron
Exon 5'
tdI intron
Exon 3'
plasmid transcript
spliced plasmid transcript
+
tdI RNA
RNA spacer acquisition
spacer acquired from spliced RNA
Direct Repeat
Direct Repeat
B
Spliced derived RNA
5'
3'
GAAUUCGCGGCCGCCUGCAGGUCGACGUUUUCUUGGGUCUACCGUUUAAUCUCGACCAUAUGGGAGAGC
5'-. . . TTAAGACTCTCCCATATGGTCGAGATTAAACGGTAGACCCGTTTCA. . .-3'
DR
Splice junction containing spacer
DR

(Figure Legend in the next page)

Figure R3.10 Spacer acquisition from RNA in the VVYJ016 type III-D system. (A) Schematic diagram of *td* intron-containing constructs. We determined whether the spacers originated from RNA, using a self-splicing transcript that produces an RNA sequence junction not encoded by DNA. Newly acquired spacers containing this exon junction may be considered to have been acquired from an RNA target. **(B)** RNA derived from the newly acquired exon junction-spanning spacer (blue). The splice site is indicated by a blue triangle. Red arrows indicate that the spacer is in the antisense orientation relative to the direction of transcription of the *td* intron. At the bottom, the highlighted sequence of one of the splice junction-containing spacers located in the CRISPR Array is indicated.

Finally, an optimized spacer acquisition assay was carried out to detect newly integrated spacers in plasmid copies of CRISPR02 (section M.13.2.5), which allows to detect more than 100,000 new spacers matching to plasmids or to the *E. coli* genome (high-throughput sequencing data presented in this thesis is listed in Appendix C4). After the analysis of the data, three unique spacers contain the splice junction (Figure R3.10A and B), confirming that the adaptation module of *V. vulnificus* YJ016 is able to acquire spacers directly from RNA molecules in a heterologous host.

R.3.5 The VvYJ016 type III-D CRISPR-Cas interference module is functional

One of the main differences between type III and the rest CRISPR-Cas systems is that type III systems target both DNA and RNA substrates. Indeed, in these systems the DNA requires to be transcriptionally active to be cleavage (to a more detailed description of type III CRISPR-Cas systems see section I.4.1.2). On the other hand, the expression stage, which involves processing of the CRISPR Array to generate mature crRNA guides, is quite similar to that carried out by type I CRISPR-Cas systems. Furthermore, in both systems Cas6 is the enzyme required for crRNA maturation (sections I.4.1.1 and 2). As described in the previous chapter, *V. vulnificus* YJ016 harbors a whole type III-D CRISPR-Cas system, and since the adaptation module works in a heterologous host it is plausible to think that the effector module of this system could be functional as well.

In fact, the analysis of the RNA-seq data used to see the differential expression of the two CRISPR Arrays encoded by the type III-D CRISPR-*cas* locus in *V. vulnificus* YJ016 also allow an interesting observation (Figure R3.3C; Williams *et al.*, 2014). The selection of those reads (>100 nt) in the same orientation of the leader sequence containing at least the direct repeat (DR) and the first spacer of each array (CRISPR01 and CRISPR02) showed that most reads began in a specific nucleotide of the DR. More specifically, the adenine +10 of the DR was the starting nucleotide in the 80% of the reads corresponding with the CRISPR01 and 60% of those of the CRISPR02 (Figure R3.11A). This finding may indicate a pre-processing of the CRISPR Array in order to produce functional crRNA guides (Figure R3.11B).

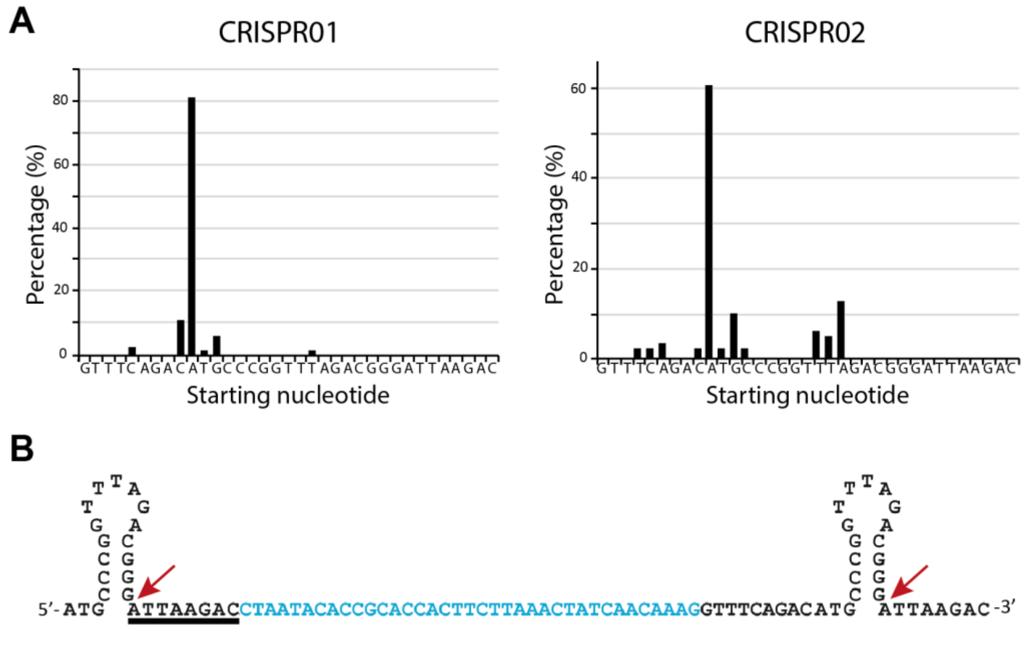


Figure R3.11 Pre-processing of the VvYJ016 CRISPR01 and CRISPR02 observed *in vivo*. (A) Percentage of reads that begin in each nucleotide of the DR of CRISPR01 (left) and CRISPR02 (right) (William et al., 2014). Most reads of both arrays start at the adenine +10 of the DR (80 and 60%, respectively). **(B)** Structure of the most abundant read. The read starts in the adenine +10 of the first DR. The first spacer of the CRISPR01 is indicated in blue as example. DRs and their characteristic stem loop structure are indicated in black. The 8 nts underlined just upstream of the spacer represent the putative region that allow the type III interference module to distinguish self-versus non-self in order to avoid auto-immunity. The red arrows indicate the putative cleavage sites of Cas6 to produce the mature crRNA.

To confirm that the type III-D effector module of *V. vulnificus* YJ016 is functional in their native host, an interference assay was carried out as described in section M.13.3. Briefly, the experimental procedure consists of mobilize a plasmid containing the complementary sequence (DNA target) to one of the natural spacers of the CRISPR Array to YJ016 strain. As long as the DNA target is transcriptionally active, the interference machinery would cleavage both DNA and RNA driving cell to die. Since the spacer 1 of the CRISPR01 is the highly expressed spacer in the RNA-seq data (William *et al.*, 2014), it was selected as target for the effector module. Thus, this spacer was cloned in both orientations between *Sph*I/*Kpn*I restriction sites of pVSV-105 vector (Table M4). To test whether there is an interference effect against the target plasmid without adding an external promoter, an empty plasmid and the plasmid with the target in one orientation (pVSV-105-sk) or in the other (pVSV-105-ks) was used to transform two strains of *V. vulnificus*: YJ016, which contains the type III-D CRISPR-Cas system, and R99, that lacks CRISPR-Cas systems and serves as negative control (Table M1). Despite the fact that when the plasmid contains the target the transformation efficiency decreases by an order of magnitude, the results show no significant effect between both strains (Figure R3.12).

Thus, in order to guarantee the expression of the target, a constitutive promoter, the OmpC promoter, was cloned in the *Sph*I restriction site of pVSV-105 vector (Table M9). Due to promoter activity only the spacer cloned in *Kpn*I-*Sph*I orientation (ks) would be complementary to the mature crRNA guide containing the first spacer of CRISPR01. As described above, *V. vulnificus* strains YJ016 and R99 were transformed with the three plasmids: pVSV-105-OmpC, pVSV-105-OmpC-sk and pVSV-105-Ompc-ks. The results of the assay reveal that only when the expressed target is complementary to the crRNA guide (ks construction) a difference of three orders of magnitude between the strain containing the CRISPR-Cas system (YJ016) and the one lacking it (R99) was observed (Figure R3.12). Then, the type III-D effector module of *V. vulnificus* YJ016 works similarly to other already described type III CRISPR-Cas system by requiring DNA transcriptionally active to properly perform the interference stage.

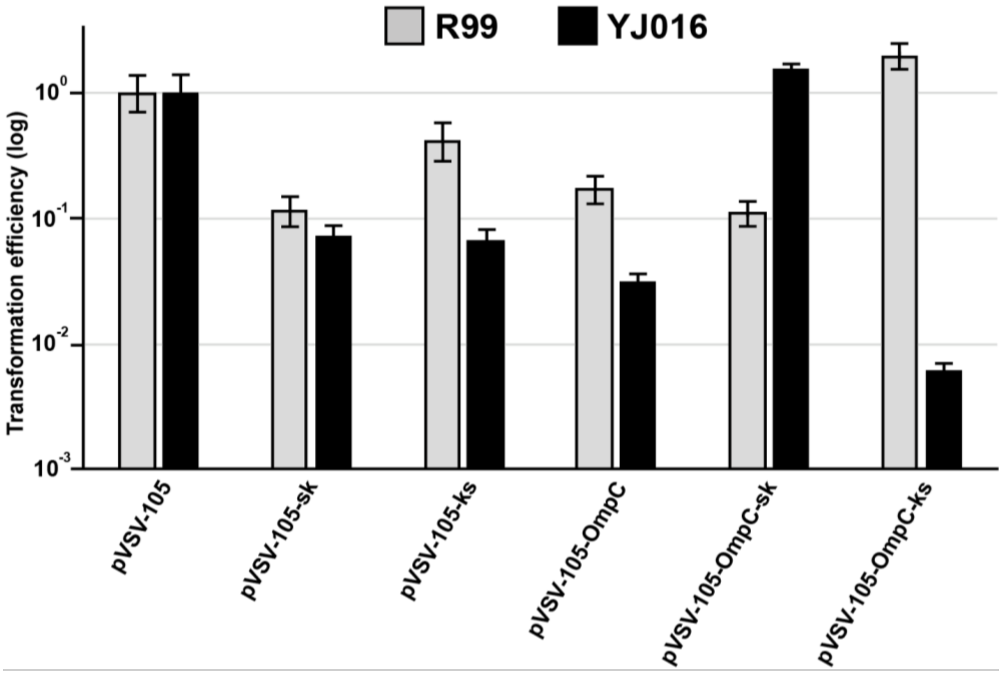


Figure R3.12 Detection of type III-D interference activity *in vivo*. Histogram showing the relative transformation efficiencies of *V. vulnificus* strains R99 (grey) and YJ016 (black) with the different constructs for spacers 1 of CRISPR01 with or without the OmpC promoter. Transformations were performed three times and average relative transformation efficiencies are given with the standard deviations. The transformation efficiency of the empty vector pVSV-105 is taken as 100% ($\mathbf{10^0}$).

Discussion

In the first chapter an overview of the current knowledge about prokaryotic RTs is provided. Group II introns and retrons are the most vastly distributed RTs in bacterial phyla, while DGRs and group II introns as well are the widespread RTs in archaeal genomes. Overall, the data shown here are in agreement with previous RT diversity surveys (Kojima and Kanehisa, 2008, Simon and Zimmerly, 2008; Toro and Nisa-Martínez, 2014). Interestingly, the abundance of DGRs in DPANN group as well as in bacterial CPR group has been previously described (Paul *et al.*, 2017; Roux *et al.*, 2020), but here we provide new insights that DGRs are by far the main type of RTs found in these phylogenetic groups. DPANN group is the most divergent of all archaeal groups, analogous to what happens with CPR group in Bacteria (Castelle *et al.* 2018). Indeed, these groups share similar characteristics, differing from the rest of prokaryotic organisms because of their small cell and genome size, their episymbiotic relations with other organisms, and their limited metabolic capacities. In this context, a reason why DGRs are so abundant in DPANN and CPR groups could be the role this type of RTs plays in enabling the adaptation of these organisms to their biological niches through changes in membrane proteins that allow better interactions with their hosts (Castelle *et al.*, 2018). In summary, the up-to-date prokaryotic phylogeny presented here provides a basis for future studies of RT properties and function in a broad range of organisms.

Data shown in chapter 1 indicates that the association between RTs and CRISPR-Cas systems probably occurred first in bacteria. Furthermore, most of the RT sequences are associated with type III CRISPR-Cas system. It is extensively shown that type III CRISPR-Cas systems target DNA transcriptionally active (section I.4.1.2). The acquisition of an RT domain through evolution could provide an evolutive advantage allowing type III systems to acquire spacers from RNA or from highly transcribed regions in order to a better performance of the interference stage Silas *et al.*, 2016). Interestingly, in this study RTs are found in all type III subtypes (III-A to III-E), with the exception of subtype III-F. Indeed, despite been reported to lack *cas1* and *cas2* genes (Makarova *et al.*, 2015), numerous examples of subtypes III-C and III-D harboring an adaptive unit with an RT domain (adjacent or fused to Cas1) are described in the current analysis (Appendix A1).

Discussion

Two hypothesis of the origin of RTs linked to CRISPR-Cas systems are discussed. In contrast with the "single point origin" model (Silas *et al.*, 2017a), the results of the phylogenetic analysis carried out in the first chapter show several examples indicating that RTs may have become associated with type III CRISPR-Cas systems on various occasions. According to their phylogenetic positions relative to group II intron classes, they appear to form two major lineages in bacteria and one minor lineage in archaea. The archaeal RTs associated with CRISPR-Cas systems (clade 1) appears to be the most recent acquisition event from a class F group II intron. This fact suggests that the association between RTs and CRISPR-Cas systems may often occur as the result of group II intron retrotransposition events. Accordingly, based on the topology of the tree, the identified clades do not all branch off from a single supported node, and there is, therefore, no consistent evidence in favor of their monophyly. The two clearly independent events of acquisition of the Cas6 domain by RTCas1 fusion proteins in clades 8 and 11 also support the hypothesis of the "various origins" of RTs linked to CRISPR-Cas systems. Nonetheless, because of the difference in the branch length and branching patterns these conclusions should be carefully interpreted.

Indeed, the Cas6RTCas1 fusions of clade 11, which are harbored in *Porphyromonas* species, has been argued to be incorrectly positioned in the RT phylogeny (Mohr *et al.*, 2018). This argument is based on the fact that the RTs of this clade forms a distinct long branch, which is not closely related with any stand-alone Cas6 sequences. However, the phylogeny of Cas6 domains shows that Cas6 present in clade 11 forms a separate branch (Branch 17) different from that of clade 8 (Branch 11), indicating that not only the RT but also the Cas6 domain have high rates of amino-acid replacement (Mohr *et al.*, 2018). Furthermore, to avoid the influence in the topology of RT domains from clade 11 within the phylogenetic trees constructed during this work, the larger linker located downstream from the RT7 domain (Mohr *et al.*, 2018) that differentiates these fusions from those of clade 8 was removed from the alignments. In addition, a new member at the base of the clade 11 (*Bacteroidetes bacteria*, PID94761.1) does not influence its topology. Thus, the distant phylogenetic position between Cas6RTCas1 fusion from clades 8 and 11 could be consequence of a faster rate of molecular evolution within clade 2/11 and not necessarily imply a long branch attraction (LBA) effect.

Further support to the "multiple origin" model of RTs linked to CRISPR-Cas systems is also observed in both clades found in the phylum Cyanobacteria (clade 3 and 5), which may have two independent evolutionary origins (Figure R1.4 and R1-5; Appendix A2). RT-containing CRISPR-Cas systems are widespread among cyanobacteria of various developmental patterns and complexities. However, the phylogenetic relationships between the distinct cyanobacterial clades and their morphological complexity remain largely unknown (Shih *et al.*, 2013). Cyanobacteria phylum is classified into five subsections and examples of organisms from almost all subsections harboring RTs associated with CRISPR-Cas systems are found in clades 3 and 5 (subsection I, Chroococcales; subsection III; Oscillatoriales, subsection IV; Notococales, and subsection V, Stigonematales). Nevertheless, none of the five full genome sequences available (*Stanieria cyanosphera*, *Myxosarcina* sp. Gl1, *Pleurocapsa* sp. PCC 7327, *Pleurocapsa* sp. 7319 and *Xenococcus* sp. PCC7305) from subsection II (Pleurocapsales), contain genes encoding RTs associated with CRISPR-Cas systems. Subsection II corresponding to unicellular coccoids reproducing by multiple fission events to generate small cells (baeocytes). It could be suggesting that the particular lifestyle of these bacteria may restrains the presence of RT-containing CRISPR-Cas systems. However, only a few genomes of subsection II are sequenced to find whether exist some limitation of these organism to present RT-CRISPR systems due to their biology, thus, the significance of this fact is currently uncertain.

To give more arguments in favor of the "multiple origins" hypothesis, the comparison between the RT and Cas1 phylogeny also suggests that the association between both domains may have occurred independently on a number of occasions. However, given the diversity of *cas1* loci within type III systems (Makarova *et al.*, 2015), the existence of only two major lineages of Cas1 proteins associated with RTs suggests that these particular CRISPR-Cas systems may also be subject to functional constraints dependent on unknown features of the associated Cas1 protein subtype. Thus, efforts to unravel the biochemically interactions between RT and Cas1 proteins are required to fully understand the biological relationship between both domains.

On the other hand, the RT sequences fused to AEP domains found at the base of clade 12 imply a really interesting finding, since a single protein could contain both reverse transcriptase and primase activity. Thus, a hypothesis of their biological role could involve this primase activity in the acquisition of spacers within the CRISPR Array in these particular systems facilitating the conversion of RNA molecules into cDNA in the absence of a primer. This possibility it is worthy to be explored since the characterization of these singular group of RTs could also provide useful biotechnological tools, such as in RNA high-throughput technologies.

As it is repeatedly observed during chapter 1, all the data presented here suggests that RTs linked to CRISPR-Cas systems could have evolved from an ancestral group II intron retrotransposition event which may occur several times during evolution. Likewise, the close relationship between G2L and RT-CRISPR (Figure R1.8 and R1.10) might suggest that G2L-RTs have also evolved from a common ancestral group II intron. However, the biological function of these RTs remains unknown to date. One possibility could be that G2L-RTs 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, such as the adaptive immunity mediated by CRISPR-Cas systems. It would be interesting the biological characterization of G2L-RTs to shed light on this hypothetical scenario.

Apart from RTs which evolved from group II introns, a really interesting finding is the presence of RTs associated with CRISPR-Cas systems that have their origin in retron/retron-like or Abi-P2 sequences. These RT sequences constitute the novel clades 14 and 15, respectively (Figure R1.8 and R1.10). RTs from clade 14 have the characteristic retron signatures: the “VTG” motif within RT domain 7, and the conserved NAXXH motif, located between domains 2 and 3 (Simon *et al.*, 2019). However, no msr and msd regions (the substrate for the msDNA production) were identified close to the RT locus. This finding may suggest that retron/retron-like RTs from clade 14 were recruited relatively recently by type III CRISPR-Cas systems. Unlike the other clades, sequences from clade 15 are associated with type I-C CRISPR-Cas systems. Type I-C is the second most abundant CRISPR-Cas system type (Makarova *et al.*, 2015). However, current data reveals that RT-CRISPR emerging from Abi-P2 are only harbored by bacteria from order Pasteurellales. The clustering by host/environmental niche rather than by vertical inheritance in these species indicate that they dissemination has likely occur recently, possibly by lateral transfer between bacteria which are found in the same microniches.

The recent findings that multiple defense systems are found in genomic islands (Doron *et al.*, 2018), as well as, that both Abi-P2 and retrons confers resistance against a broad range of phages (Odegrip *et al.*, 2006; Gao *et al.*, 2020; Millman *et al.*, 2020) could suggest the possibility that RTs from clades 14 and 15 constitute a single immunity system co-locating with other defense mechanism (in this case a type III or type I-C CRISPR-Cas system, respectively) instead of a novel functional association between RTs and CRISPR-Cas systems. Thus, experimental characterization of these particular RTs groups is required in order to unravel their biological role.

In the case of type VI CRISPR-Cas systems, although they usually lack the *cas1* and *cas2* genes of the adaptation module (Makarova *et al.*, 2020), the identification of two different type VI-A systems including RTCas1-Cas2 proteins could indicate a higher specificity and efficiency of associated Cas13a effectors (Koonin and Makarova, 2019). Furthermore, due to type VI systems exclusively target single-stranded RNA (Shmakov *et al.*, 2015; East-Seletsky *et al.*, 2016), the presence of an RT domain may facilitate the acquisition of novel spacers from RNA molecules. However, the implications of this apparently specific association between RTs and type VI-A systems remains unclear. This association may reflect functional constraints of the adaptation and effector complexes of type VI-A CRISPR-Cas systems. Nonetheless, it cannot be excluded that RTs can be recruited by other Cas13 subtypes as a differentiated response to particular pressures on the host ecosystems due to changeable environmental conditions. Furthermore, most of the bacteria containing type VI-A systems have RT genes in their genomes (data not shown), raising the possibility that a *trans*-RT activity could allow type VI-A systems to acquire spacers from RNA.

As occurs with RTs associated with type III CRISPR-Cas systems, adaptation modules harboring RTCas1 fusions were recruited on several occasion by different effector Cas13a proteins, possibly from type III-D systems. These results are consistent with the proposed evolutionary scenario of multiple independent acquisitions of adaptation modules by type VI CRISPR-Cas systems (Koonin and Makarova, 2019). Within the RTCas1 fusion proteins found within clade 1 (Alphaproteobacteria) it is worth to highlight that in *E. rectale,* the type VI-A CRISPR-*cas* locus encoding the RTCas1 fusion protein was identified in several strains (AF25- 15, AF18-16LB, and AF18-18LB). Nevertheless, in the latter strain, a *cas2* gene is separated from the *RTCas1* locus by a genomic island of ~22 kb. Additionally, other *E. rectale* strains presented frameshift mutations in the RT sequence (strains AF19-4, AF19-3AC, and AM29-10), or had lost the RTCas1 fusion (strain TM10-3) or the complete adaptation module (strain T1-815). Interestingly, the Cas13a protein of the T1-815 strain (EreCas13a) the pre-cRNA processing and ribonuclease activities has been experimentally characterized (East-Seletsky *et al.*, 2017). Thus, the findings shown in chapter 1 suggest that the type VI-A CRISPR-Cas-associated genes encoding RTCas1 and Cas2 are frequently gained and lost, possibly due to the dynamics of microbial host-phage interactions which require further research.

All these findings paving the way for studying the RT-mediated acquisition of spacers for type VI CRISPR-*cas* loci. In addition, the search in metagenomes it is predicted to result in the expansion of RTCas1 fusion proteins linked not only to type VI, but also to type III CRISPR-Cas systems. Both type III and VI interference machinery has been engineered to provide useful biotechnological tools in fields such as gene knockdown, RNA editing as well as for diagnosis purposes (O’connell, 2019; Burmistrz *et al.*, 2020). In-depth investigation of RT-CRISPR systems could lead to novel and exciting opportunities in synthetic biology and engineering. Thus, it is worth to highlight that the comprehensive characterization of the origins and relationships of RTs associated with RNA-targeting CRISPR-Cas systems (type III and VI) carried out in this work provide the basis for experimental studies that have been performed in chapters 2 and 3.

In the second chapter, the biochemical activity of five RT homologs was used as first criteria for the selection and further study of putative functional RT-containing CRISPR-Cas systems. This strategy has served to define the study toward only two of the five selected examples of CRISPR-*cas* loci associated with RTs: the one from *S. hofmanni* PCC 7110 containing a RT alone and that from *V. vulnificus* YJ016, which harbor a RTCas1 fusion proteins. This approach represents a valid strategy as the first step to select candidates. Thus, apart from RTs associated with CRISPR-Cas system, this approach could be used to easily analyzed other RT groups which remains largely uncharacterized, including retrons/retron-like RTs or UG-RTs, both of which have been recently shown to participate in bacterial defense against phages (Gao *et al.*, 2020; Millman *et al.*, 2020).

The high exogenous RT activity shown by the selected proteins could suggest that these RTs might play an important role in the acquisition of novel spacers in the CRISPR-*cas* loci that harbor them. Interestingly, these two proteins represent different states of the evolution of RTs linked to CRISPR-Cas adaptation modules: on one side, a RT alone, that could suggest a more tightly association of this domain with the Cas1-Cas2 integrase complex and, on the other side, a RTCas1 fusion protein, in which the RT domain has been fully recruited by the CRISPR-Cas system. Thus, the RT-Cas1-Cas2 adaptive operon from *S. hofmanni* PCC 7110 and the RTCas1-Cas2A-Cas2B system from *V. vulnificus* YJ016 represent excellent examples to determine whether the different evolution stages of the association between RTs and CRISPR-Cas systems have functional implications in the mechanism of spacer acquisition.

However, only RT-containing CRISPR-Cas systems with a Cas6 domain fused to the N-termini of the RT has been biochemically characterized: one from *M. mediterranea* MMB-1 (Silas *et al.*, 2016; Mohr *et al.*, 2018) and the other from *Thiomicrospira* sp. (Wang *et al.*, 2020). In the *M. mediterranea* system it has been demonstrated that the additional Cas6 domain participates not only in crRNA biogenesis but is also required for RT activity and its regulation. This was the first evidence of a single protein (Cas6RTCas1) participating in the first two stages of CRISPR-Cas immunity, adaptation and expression (Mohr *et al.*, 2018). The crosstalk between the different domains of Cas6RTCas1 proteins has been also suggested by the cryo-EM structure of the *Thiomicrospira* sp. integrase complex (Wang *et al.*, 2020), which shows bidirectional crosstalk between the RT and Cas1 domains and unidirectional crosstalk from Cas6 to the other two domains. The close relationships of RT, Cas1 and Cas6 domains in these systems could suggest that Cas6 likely interact with RT and RTCas1 proteins as well. Thus, it is worthwhile to characterize whether the *S. hofmanni* RT alone and the VvYJ016 RTCas1 fusion interacts with the Cas6 present in different or the same CRISPR-*cas* loci, respectively in order to shed more light on Cas6-RT crosstalk.

Furthermore, the number of CRISPR-Cas systems of *S. hofmanni* and *V. vulnificus* hosts could also provide additional insight into the lifestyle of these organisms. *S. hofmanni* PCC 7110, a marine cyanobacterium (Rippka *et al.,* 1979), present a great variety of type I and type III CRISPR-Cas systems, suggesting that this strain require a battery of diverse systems to efficiently faced potential atackers present in their biological niche. This hypothesis is also support by the extraordinary number of CRISPR Arrays (21 in total; Appendix B2) containing an extensive repertoire of spacers. Furthermore, crosstalk between type I and type III CRISPR-Cas system has been already described (Silas *et al.,* 2017b). Thus, the type III CRISPR-Cas interference complexes from S. *hofmanni* PCC 7110 may also use the mature crRNA guides from type I CRISPR Arrays in order to target phages that are able to evade type I immunity through mutations in PAM sequence.

Curiously, strain PCC 7110 contains four RT-containing CRISPR-Cas, including two RTCas1 fusion and two RT alone which are clustered in clade 3 and 5, respectively. This finding may indicate that RT activity could be essential to protect this cyanobacterium against ssRNA phages or transcriptionally active DNA phages. Nevertheless, the deeper characterization of the spacer repertoire of RTCas1-containing type III CRISPR-Ca system from the commercial cyanobacterium *Arthospira platensis* revealed that the source(s) of RT-related spacers remain enigmatic and only a few matches to DNA phage-like sequences have been found (Silas *et al.*, 2017a). Due to the limited abundance and distribution of RNA phages, only one natural example of an RT-related spacer targeting an RNA phage has been described (Wolf *et al.*, 2020), 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 more than 15,000 near-complete genomes (Callanan *et al.*, 2020), might help to shed light on this issue and improve our understanding of the biological role of RT-CRISPR-Cas systems in the ecological niche of their host.

An interesting fact is that the RT alone from *S. hofmanni* form a dimer in solution, which could represent the active state of this protein as describe for other RT proteins such as that from the human immunodeficiency virus (HIV) (di Marzo-Veronese *et al.*, 1986). Despite Cas1 from *S. hofmanni* appears as a monomer in the conditions tested in the present work, Cas1 from other CRISPR-Cas systems has also been shown to form a dimer in other systems (Nuñez *et al.*, 2014; Wright and Doudna, 2016). Thus, the RT dimer could be required for an appropriate interaction with a predicted Cas1 dimer. This fact could be supported by the existence of RTCas1 fusion proteins, that indicates that RT and Cas1 domains should be present in a 1:1 ratio. In agreement with other integrase complexes described, the analyzed *S. hofmanni* Cas2 form also a dimer, which might act as a scaffold of the final complex. Thus, every Cas2 unit could bind to a dimer of a Cas1-RT unit, constituting the canonical heterohexamer architecture of the integrase complex (Nuñez *et al.*, 2014; Wright and Doudna, 2016) together with the four additional RT proteins. This fact is supported by the Cryo-EM heterohexamer structure of the Cas6RTCas1-Cas2 complex from *Thiomicrospira* sp, consisting of a central Cas2 dimer and two distal Cas6RTCas1 dimers (Wang *et al.*, 2020). This Cryo-EM structure also reveals interactions between the RT domain an Cas2, in agreement with the RT-Cas2 interactions described in chapter 2 (Figure R2.7).

In addition, the RT alone from *S. hofmanni* PCC7110 has been shown to interact with two other ancillary proteins: a WYL-domain-containing protein and with a putative phosphohydrolase. Although more experiments are required to validate these interactions, these proteins could play a role in the regulation of the *S. hofmanni* adaptation module. Indeed, WYL domains are frequently found in transcription factors that may be activated by binding RNA molecules (Muller *et al.*, 2019). In the cyanobacterium *Synechocystis* 6803 a WYL-domain containing protein act as transcriptional repressor of crRNA biogenesis (Hein *et al.*, 2013). As long as the CRISPR Array transcription is blocked, the RT-containing adaptive complex could integrate novel spacers within the Array. Thus, is reasonable to hypothesize that RT-WYL-domain-containing interaction may lead to a spatio-temporal regulation of adaptive and expression stages.

On the other hand, *V. vulnificus* YJ016, an opportunistic pathogen isolated from a human sample in a hospital (Chen *et al.*, 2003), only harbor a unique type III-D CRISPR-Cas system. This CRISPR-*cas* locus is located in a genomic island only present in a few representatives of the genus *Vibrio*. Indeed, YJ016 is the unique strain containing this singular system within the sequenced *Vibrio vulnificus* strains. As maintaining a CRISPR-Cas system required a lot of cellular resources and their presence can lead to autoimmunity (Stern *et al.*, 2010), an explanation for the fact that this particular CRISPR-Cas system is only present in YJ016 strain might be provided by the recent proposed “pan-immune system” model (Berhmeim and Sorek, 2019). This model hypothesized that a mixed population of strains potentially encodes a battery of defense systems that protects the whole population against a broad range of phages and MGEs. Thus, the effective immune system is not the one encoded by a particular genome of a single microorganism (e.g. *Vibrio vulnificus* YJ016) but rather by its pan-genome.

Apart from the RTCas1 fusion protein with a strong exogenous RT activity, the other particular feature of the VvYJ016 adaptive module is the presence of the two different Cas2 proteins, which form a heterodimer unit (Figure R2.12). This distinctive fact could have implications in the mechanism of spacer acquisition carried out by the entire adaptive operon as the central part of a protospacer has been shown to lies on the surface of the Cas2 dimer (Wang *et al.*, 2015). The binding its stabilized by charge-charge interactions via the phosphate backbone of the protospacer with the arginine residues (positively charged) of the Cas2 surface.

Interestingly, VvYJ016 Cas2A and Cas2B present some conserved arginine residues in different positions of their sequences (Figure R2.9). In addition, Cas2A present three arginine residues in a row in positions 75, 76 and 77 (80, 81 and 82 in the alignment shown in Figure R2.9), whereas Cas2B only contain the arginine in the position 77, lacking the other two residues. Then, each Cas2 might have a preference for binding specific nucleotides within a putative protospacer sequence. This fact may affect the selection and capture of the sequences that will be integrated in the CRISPR Array. Spacer integration assays using the VvYJ016 adaptive operon could help to determine the existence of a putative effect of Cas2A-Cas2B heterodimer in protospacer selection and integration.

Nevertheless, the *in vitro* integration assays performed in the second chapter have revealed a non-specific nuclease activity that could be derived from the RTCas1 protein. Indeed, Cas1 proteins of other CRISPR-Cas systems has been shown to have this non-specific nuclease activity as well (Babu *et al.*, 2011; Grainy *et al.*, 2019; Wang *et al.*, 2020). This activity could interfere in studies of spacer integration *in vitro* suggesting that the properly formation of the integrase complex appears to be an indispensable requirement to study spacer integration *in vitro*. In all integrase complex studied, the addition of a DNA substrate to mimic a genomic substrate facilitates the assembling of the (Cas6RT-) Cas1-Cas2 integrase complex (Nuñez *et al.*, 2015a, Wright and Doudna, 2016, Wang *et al.*, 2020). Although the MMB-1 complex has been shown to integrate spacers *in vitro* even with an MBP-tagged Cas6RTCas1, the removal of the protein tags (MBP and His) appear to be critical for the formation of the complex, especially in the case of the Cas2. Then, found the conditions in which the proteins are soluble after the cleavage of the tag with the 3C protease is also a necessary fact to further characterize the activity of VvYJ016 RTCas1-Cas2-Cas2B complex *in vitro.*

In the chapter three a characterization of the process of spacer acquisition mediated by a unique adaptation module containing an RTCas1 fusion protein associated with a type III-D CRISPR-Cas system from *V. vulnificus* YJ106 has been carried out. As mention above, this adaptation operon also includes two different Cas2 proteins (Cas2A and Cas2B) and two CRISPR Arrays. The whole module has been shown to be functional in a heterologous host (*E. coli*). Furthermore, it was found that spacer acquisition presented a different efficiency between CRISPR01 and CRISPR02 and that it was strongly impaired by a lack of RT activity and abolished by mutations of the Cas1 domain or deletion of each or both Cas2. Regardless of their source of origin (genome or plasmid), the newly acquired spacers displayed a bias for the antisense strand within the coding sequences. Additionally, the nucleotide sequence of the spacers integrated within the CRISPR Array presented a series of features which would suggest a specific recognition of the prespacer by the VvRTCas1-Cas2A-Cas2B adaptation complex. Lastly, here it was also shown that this RT-containing system was capable of acquiring spacers from molecules with an RNA origin. Overall, the findings shown in chapter 3 demonstrate that the V. *vulnificus* YJ016 adaptation module associated with the type III-D CRISPR-Cas system constitutes a good model for in-depth analyzing the mechanism of acquisition of spacers directly from RNA molecules, which also has a potential value for expanding the CRISPR-Cas toolbox.

Several studies of spacer acquisition by CRISPR-Cas adaptation modules have shown that when several CRISPR Arrays are present in the same locus, only one of them presents a high naïve acquisition efficiency, whereas the other/s barely acquired (Staals *et al.*, 2016; Schmidt *et al.*, 2018,). The results of the last chapter reveal a similar pattern, as only one of the arrays, CRISPR02, appears to be fully functional in acquisition in the assayed conditions (Figure R3.2B). Indeed, the natural difference observed between the number of spacers of both Arrays in the genome of *V. vulnificus* YJ016 could be the explanation of why CRISPR02 (9 spacers) acquired more spacers than CRISPR01 (3 spacers). Nevertheless, the difference in spacers acquisition between CRISPR Arrays with the same DR remains an intriguing question. Only small differences between CRISPR01 and CRISPR02 were observed in the sequence of the leader region but there were placed more than 119 bp away from the first DR, thus, it seems highly unlikely that the difference in the spacer acquisition rate is a consequence of the recognition of one array or the other by the integration complex.

However, the promoter activity displayed by the two CRISPR Arrays was completely different: CRISPR01 show a high activity, whereas CRISPR02 had no activity at all. Interestingly, the promoter activity was inversely correlated with the number of newly acquired spacers. Despite this result, it is important to highlight that the opposite effect has been reported in the type I-E CRISPR-Cas system from *E. coli* K-12 in which the most expressed CRISPR locus is also the one with more events of spacer acquisition (Pougach *et al.*, 2010; Datsenko *et al.*, 2012). Moreover, the small sequence variations previously reported between the CRISPR01 and CRISPR02 loci in *V. vulnificus* YJ016 could be the result of disruption of binding of other host factors involved in the adaptation stage (i.e., IHF or other structural proteins; Nuñez *et al.*, 2016). Thus, this putative association between array transcription levels and spacer acquisition efficiency merits to be further study in other systems to validate one of the proposed hypotheses.

The ability to acquire novel spacers by the integrase complex of VvYJ016 can be demonstrated using the canonical spacer acquisition assays based in a PCR amplification using the forward primer in the leader region and the reverse primer in the first spacer (Yosef *et al.*, 2012; Silas *et al.*, 2016). However, most of the functionally analyzed RT-containing type III CRISPR-Cas systems (13 out of 14) by Schmidt et., (2018) requires of a method known as "selective amplification of expanded CRISPR Arrays" (SENECA) to detect the acquisition of new spacers as result of the low acquisition rate. This would indicate that the RTCas1-Cas2A-Cas2B adaptation module of VvYJ016 acquire more efficiently in a heterologous host than most RT-CRISPR system. However, the use of novel methods of PCR amplification to maximize the detection limit of expanded arrays such as the proper SENECA (Schmidt *et al.*, 2018; Tanna *et al.*, 2020) or the use of degenerate or divergent primers (Mckenzie *et al.*, 2019) could allow the study of spacer integration events in conditions or systems with a low acquisition rate. For instance, these methods could be applied to increase the detection of novel spacers integrated in assays performed with the YAAA mutant of the RT domain.

Furthermore, the data shown in chapter 3 reveals that the RT activity plays an important function in the acquisition of novel spacers in the heterologous *E. coli* host, as the assays performed with the RT active site mutated (YAAA), which has been shown to completely abolish the RT activity (Figure R3.4), results in a great reduction (about >90%) in the number of newly integrated spacers. A similar drop in spacer acquisition has been observed in an analogous RT mutant in the RT-CRISPR system in *M. mediterranea* MMB-1 host but not when this system is used in *E. coli* (Silas *et al.*, 2016). In the experimental conditions assayed in the current study spacers originated from *rRNA* genes are the most abundant ones. This fact may reflect RNA abundance-dependent spacer acquisition, as *rRNA* genes are the most abundant RNA in the cell. Nevertheless, a bias towards highly transcribed regions is shown in the spacers acquired by the Cas6RTCas1-Cas2 complex in *M. mediterranea*, while in this system spacers were rarely acquired from rRNA (Silas *et al.*, 2016). The spacers acquired by the *F. saccharivorans* acquisition complex (FsRTCas1-Cas2) also show a bias towards highly expressed genes, even though these data are obtained from spacers acquisition assays carried out in an *E. coli* host. Although both spacer datasets, those from this study and those from Schmidt et al., 2018, were obtained from assays performed in *E. coli*, the comparison of the two datasets do not show any correlation. One possible explanation could be that in the present work a different *E. coli* strain has been used. However, after the analysis of the spacers acquired by VvYJ016 adaptive complex in an assay performed in *E. coli* BL21 (the strain used in Schmidt *et al.*, 2018) there is also no bias towards highly transcribed regions.

Thus, in contrast to the *M. mediterranea* and *F. saccharivorans* RT-CRISPR systems, here does not exist a correlation between the gene transcription level and the frequency of spacer acquisition. It is essential to point out that such correlation does not distinguish between whether a spacer is acquired from DNA or from RNA and only shows that the novel integrated spacers are preferentially originated from highly transcribed genes (Schmidt *et al.*, 2018). Taken together, the findings show in chapter 3 reflect mechanistically specific features underlying the process of acquisition of novel spacers carried out by VvRTCas1–Cas2A-Cas2B acquisition system that are worth further investigation. On the other hand, and consistent with the other two RT-CRISPR-Cas systems characterized (Silas *et al.*, 2016; Schmidt *et al.*, 2018; Mohr *et al.*, 2018), the spacer acquired by the VvYJ016 integration complex show a clear bias towards the antisense strand of coding sequences when the effector module of the CRISPR-Cas system is absent. This slight preference of integrating new spacers into the CRISPR Array which are originated from the antisense strand, leads to the production of functional crRNAs. Thus, the guides generated after the array processing are complementary to predicted transcripts and conduct to an effective interference stage performed by the effector modules of type III CRISPR-Cas systems (Pyenson and Marrafini, 2017).

On the other hand, a singular feature of the RT-containing CRISPR-Cas systems of the clade 6 of the phylogeny of RTs associated with CRISPR-Cas systems is the presence of two different Cas2 proteins (Toro et al., 2019a). Moreover, this property is also observed in other RT-CRISPR-Cas systems such as the clade 13 of this phylogeny. The results shown in the last chapter reveals that Cas2A and Cas2B are required for spacer acquisition *in vivo* confirming that the functional integration complex comprises a Cas2A-Cas2B heterodimeric unit as show in chapter 2 of the present work (section R.2.4.2). This singular heterodimer could facilitate the analysis of the role of Cas2 proteins in the adaptation step of the immunity mediated by CRISPR-Cas systems. Furthermore, the presence of two different Cas2 in the integrase complex might be related with the fact that spacers with a particular sequence are selected for their integration within the CRISPR Array.

The finding of specific bias in the nucleotide sequence of the newly acquired spacers were not reported for other characterized RT-containing CRISPR-Cas systems which contain only one *cas2* gene (Silas *et al.*, 2016, Schmidt *et al.*, 2018). The function of Cas2 dimer has been already described for type I CRISPR–Cas systems, where plays a structural role in the formation and stabilization of the adaptation complex, acting as a bridge between two Cas1 dimers and binding the central region of the prespacer (Wang *et al.*, 2015; Wan *et al.*, 2019). By analogy, in the system described in this study, the bias observed at different positions within the newly acquired spacers may indicate that, within the RTCas1-Cas2A-Cas2B complex, the RTCas1 protein preferentially binds to spacers with borders rich in ‘AT’ and flanked by a ‘G’ or a ‘C’ at the derived protospacer, whereas the Cas2A–Cas2B heterodimer could be responsible for the particular bias (towards ‘C’) observed at asymmetric positions (+14 and +15) within the spacer. To validate this hypothesis further studies on the spacer integration mechanism are required.

Thus, the VvYJ016 adaptation module is a new RT-CRISPR system with novel properties different from those of the systems previously studied (Silas *et al.*, 2016; Schmidt *et al.*, 2018). Indeed, this system represents a good model for further studies not only of the role of RTs in the acquisition of spacers, but also for elucidating the particular role of heterologous Cas2 complexes and the characteristics of the spacers acquired by type III CRISPR–Cas systems.

Finally, the use of the *F. saccharivorans* RT-Cas1-Cas2 system as an RNA-recording tool appears to result in skewing to AT-rich regions at the ends of the transcripts (Schmidt *et al.*, 2018), whereas the *V. vulnificus* YJ016 adaptation module can acquire spacers regardless of their 'GC' content and from any point in the coding sequence to overcome this limitation. These differences evidence the importance to analyze different CRISPR-Cas systems harboring RTs to optimize Record-seq technology.

Additionally, the results shown in chapter 3 also reveal that the type III-D CRISPR-Cas effector module from *V. vulnificus* YJ016 is functional and works similarly to other type III systems, as transcriptionally active target DNA appears to be a requisite for cleavage. Although the interference modules of type III CRISPR-Cas systems have been extensively characterized, particularly in archaea (Deng *et al.*, 2013; Elmore *et al.*, 2016; Kazlauskiene *et al.*, 2016), the use of these systems in genome editing applications, such as gene silencing (Peng *et al.*, 2014; Zebec *et al.*, 2014, Zink *et al.*, 2019), makes it worthwhile to characterize new type III effector modules.

Furthermore, the interest in *V. vulnificus* lies on this specie is a zoonotic pathogen which is also capable of causing disease in humans that could lead to sepsis and death (Hernández-Cabanyero and Amaro, 2020). Type III-D effector module natively encoded by *V. vulnificus* YJ016 could be reprogrammed to target host mRNA and be used as a gene knockdown technology. With this aim, this strain would be transforming with the CRISPR Array containing a synthetic spacer which would be complementary to specific host mRNAs. Then, the processed crRNA would guide the effector machinery to a specific region where the RNA would be degraded. Thus, this tool could be used to further investigate the role of uncharacterized genes encoded by YJ016 strains, such as those involved in virulence.

In summary, not only the adaptation operon but also the effector module of *V. vulnificus* YJ016 could be used to expand the CRISPR-Cas toolbox. On one side, the VvRTCas1-Cas2A-Cas2B integration complex could be used in Record-seq technologies. On the other side, the characterization of the interference stage carried out by the effector module could result in novel biotechnological applications that would allow a deeper understanding of important aspects of the biology of this pathogen.

Conclusions

1. **The different distribution of RTs types, , especially retrons and DGRs, between archaea and bacteria show evidence that some prokaryotic phyla/groups have recruited specific types of RTs to improve their responses to the ecological conditions in their natural environments.**

2. **RTs linked with the adaptive module of CRISPR-Cas systems are clustered in 15 phylogenetic clades which have at least three different origins, supporting a "multiple origins" model in which first the RT is recruited by a CRISPR-Cas system, the RT coevolved with the adjacent Cas1 forming in some cases a fusion protein. Finally, in at least two independent branches, a Cas6 domain was fused to the N-termini of RTCas1.**

3. **Most CRISPR-RTs are associated with most subtypes of type III systems, except those evolving from Abi-P2 that were linked to type I-C systems and some representatives from clade 7 and 12 that have been recruited by type VI-A CRISPR-Cas effector systems.**

4. **The RT alone from type III-B/C CRISPR-Cas system of *Scytonema hofmanni* PCC 7110 and the RTCas1 fusion protein from type III-D system of *Vibrio vulnificus* YJ016 present high exogenous RT activity *in vitro*.**

5. **Cas2A and Cas2B proteins from *V. vulnificus* YJ016 adaptive module, which present a 70% of identity, form a heterodimeric unit.**

6. **The RTCas1-Cas2A-Cas2B adaptive module from *V. vulnificus* YJ016 is able to acquire spacers in the heterologous *Escherichia coli* host, and this spacer acquisition is strongly impaired by the lack of the RT activity, and completely abolished by single mutation at the active site of the Cas1 domain and by deletions of either one or both Cas2A and Cas2B proteins.**

7. **Spacer acquisition rate of the two CRISPR Arrays present in *V. vulnificus* YJ016 are inversely correlated with the promoter activity of the leader region.**

8. **Spacers acquisition by the RT-associated adaptive complex displays a bias for the antisense strand in the absence of the effector unit suggesting a intrinsic preference of the adaptation complex to generate a crRNA complementary to predicted transcripts to conduct to an effective interference in Type III CRISPR–Cas systems.**

9. **The adaptive module from *V. vulnificus* YJ016 is able to acquire spacers directly from RNA molecules.**

10. **The type III-D effector module of *V. vulnificus* YJ016 requires a DNA target transcriptionally active to confer immunity as described for other type III CRISPR-Cas systems.**

Bibliography

Bibliography

Abudayyeh, O. O., Gootenberg, J. S., Konermann, S., Joung, J., Slaymaker, I. M., Cox, D. B., *et al.* (2016). C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* 353, aaf5573.

Al-Shayeb, B., Sachdeva, R., Chen, L. X., Ward, F., Munk, P., Devoto, A., *et al.* (2020). Clades of huge phages from across Earth's ecosystems. *Nature* 578, 425–431.

Alayyoubi, M., Guo, H., Dey, S., Golnazarian, T., Brooks, G. A., Rong, A., *et al.* (2013). Structure of the essential diversity-generating retroelement protein bAvd and its functionally important interaction with reverse transcriptase. *Structure* 21, 266–76.

Alkhnbashi, O. S., Costa, F., Shah, S. A., Garrett, R. A., Saunders, S. J., and Backofen, R. (2014). CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci. *Bioinformatics* 30, i489–96.

Almendros, C., Nobrega, F. L., McKenzie, R. E., and Brouns, S. J. J. (2019). Cas4-Cas1 fusions drive efficient PAM selection and control CRISPR adaptation. *Nucleic Acids Res* 47, 5223–5230.

Amaro, C., Aznar, R., Alcaide, E., and Lemos, M. L. (1990). Iron-binding compounds and related outer membrane proteins in *Vibrio cholerae* non-O1 strains from aquatic environments. *Appl Environ Microbiol* 56, 2410–6.

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

Anantharaman, V., Makarova, K. S., Burroughs, A. M., Koonin, E. V., and Aravind, L. (2013). Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing. *Biol Direct* 8, 15.

Anders, C., Niewoehner, O., Duerst, A., and Jinek, M. (2014). Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* 513, 569–73.

Anderson, D. G., and Kowalczykowski, S. C. (1997). The translocating RecBCD enzyme stimulates recombination by directing RecA protein onto ssDNA in a chi-regulated manner. *Cell* 90, 77–86.

Andersson, A. F., and Banfield, J. F. (2008). Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* 320, 1047–50.

Arambula, D., Wong, W., Medhekar, B. A., Guo, H., Gingery, M., Czornyj, E., *et al.* (2013). Surface display of a massively variable lipoprotein by a Legionella diversity-generating retroelement. *Proc Natl Acad Sci U S A* 110, 8212–7.

Artamonova, D., Karneyeva, K., Medvedeva, S., Klimuk, E., Kolesnik, M., Yasinskaya, A., *et al.* (2020). Spacer acquisition by Type III CRISPR-Cas system during bacteriophage infection of *Thermus thermophilus*. *Nucleic Acids Res* 48, 9787–9803.

Athukoralage, J. S., Rouillon, C., Graham, S., Grüschow, S., and White, M. F. (2018). Ring nucleases deactivate type III CRISPR ribonucleases by degrading cyclic oligoadenylate. *Nature* 562, 277–280.

Babu, M., Beloglazova, N., Flick, R., Graham, C., Skarina, T., Nocek, B., *et al.* (2011). A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair. *Mol Microbiol* 79, 484–502.

Baltimore, D. (1970). RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature* 226, 1209–11.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., *et al.* (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315, 1709–12.

Belfort, M., and Lambowitz, A. M. (2019). Group II Intron RNPs and Reverse Transcriptases: From Retroelements to Research Tools. *Cold Spring Harb Perspect Biol* 11.

Benler, S., Cobián-Güemes, A. G., McNair, K., Hung, S. H., Levi, K., Edwards, R., *et al.* (2018). A diversity-generating retroelement encoded by a globally ubiquitous Bacteroides phage. *Microbiome* 6, 191.

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

Bikard, D., Hatoum-Aslan, A., Mucida, D., and Marraffini, L. A. (2012). CRISPR interference can prevent natural transformation and virulence acquisition during *in vivo* bacterial infection. *Cell Host Microbe* 12, 177–86.

Biswas, A., Staals, R. H., Morales, S. E., Fineran, P. C., and Brown, C. M. (2016). CRISPRDetect: A flexible algorithm to define CRISPR arrays. *BMC Genomics* 17, 356.

Bland, C., Ramsey, T. L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N. C., *et al.* (2007). CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* 8, 209.

Blum, H., Beier, H., and Gross, H. J. (1987). Improved silver staining of plant proteins RNA and DNA in polyacrylamide gels. *Electrophoresis* 8, 93–99. doi:10.1002/elps.1150080203.

Bobonis, J., Mateus, A., Pfalz, B., Garcia-Santamarina, S., Galardini, M., Kobayashi, C., *et al.* (2020a). Bacterial retrons encode tripartite toxin/antitoxin systems. Preprint at bioRxiv. doi:10.1101/2020.06.22.160168.

Bobonis, J., Mitosch, K., Mateus, A., Kritikos, G., Elfenbein, J. R., Savitski, M. M., *et al.* (2020b). Phage proteins block and trigger retron toxin/antitoxin systems. Preprint at bioRxiv. doi:10.1101/2020.06.22.160242.

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

Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* 72, 248–54.

Brouns, S. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J., Snijders, A. P., *et al.* (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* 321, 960–4.

Burmistrz, M., Krakowski, K., and Krawczyk-Balska, A. (2020). RNA-Targeting CRISPR-Cas Systems and Their Applications. *Int J Mol Sci* 21.

Béguin, P., Charpin, N., Koonin, E. V., Forterre, P., and Krupovic, M. (2016). Casposon integration shows strong target site preference and recapitulates protospacer integration by CRISPR-Cas systems. *Nucleic Acids Res* 44, 10367–10376.

Cabanes, D., Boistard, P., and Batut, J. (2000). Identification of *Sinorhizobium meliloti* genes regulated during symbiosis. *J Bacteriol* 182, 3632–7.

Callanan, J., Stockdale, S. R., Shkoporov, A., Draper, L. A., Ross, R. P., and Hill, C. (2020). Expansion of known ssRNA phage genomes: From tens to over a thousand. *Sci Adv* 6, eaay5981.

Carte, J., Pfister, N. T., Compton, M. M., Terns, R. M., and Terns, M. P. (2010). Binding and cleavage of CRISPR RNA by Cas6. *RNA* 16, 2181–8.

Carte, J., Wang, R., Li, H., Terns, R. M., and Terns, M. P. (2008). Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev* 22, 3489–96.

Castelle, C. J., Brown, C. T., Anantharaman, K., Probst, A. J., Huang, R. H., and Banfield, J. F. (2018). Biosynthetic capacity, metabolic variety and unusual biology in the CPR and DPANN radiations. *Nat Rev Microbiol* 16, 629–645.

Castro, C., Smidansky, E. D., Arnold, J. J., Maksimchuk, K. R., Moustafa, I., Uchida, A., *et al.* (2009). Nucleic acid polymerases use a general acid for nucleotidyl transfer. *Nat Struct Mol Biol* 16, 212–8.

Charpentier, E., Richter, H., van der Oost, J., and White, M. F. (2015). Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. *FEMS Microbiol Rev* 39, 428–41.

Chen, C. Y., Wu, K. M., Chang, Y. C., Chang, C. H., Tsai, H. C., Liao, T. L., *et al.* (2003). Comparative genome analysis of *Vibrio vulnificus*, a marine pathogen. *Genome Res* 13, 2577–87.

Chillón, I., Martínez-Abarca, F., and Toro, N. (2011). Splicing of the *Sinorhizobium meliloti* RmInt1 group II intron provides evidence of retroelement behavior. *Nucleic Acids Res* 39, 1095–104.

Coros, C. J., Piazza, C. L., Chalamcharla, V. R., and Belfort, M. (2008). A mutant screen reveals RNase E as a silencer of group II intron retromobility in *Escherichia coli*. *RNA* 14, 2634–44.

Coros, C. J., Piazza, C. L., Chalamcharla, V. R., Smith, D., and Belfort, M. (2009). Global regulators orchestrate group II intron retromobility. *Mol Cell* 34, 250–6.

Cox, D. B. T., Gootenberg, J. S., Abudayyeh, O. O., Franklin, B., Kellner, M. J., Joung, J., *et al.* (2017). RNA editing with CRISPR-Cas13. *Science* 358, 1019–1027.

Crick, F. H. (1958). On protein synthesis. *Symp Soc Exp Biol* 12, 138–63.

Crowley, V. M., Catching, A., Taylor, H. N., Borges, A. L., Metcalf, J., Bondy-Denomy, J., *et al.* (2019). A Type IV-A CRISPR-Cas System in *Pseudomonas aeruginosa* mediates RNA-Guided Plasmid Interference *in vivo*. *CRISPR J* 2, 434–440.

Dai, L., and Zimmerly, S. (2002). Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Res* 30, 1091–102.

Dai, L., and Zimmerly, S. (2003). ORF-less and reverse-transcriptase-encoding group II introns in archaebacteria, with a pattern of homing into related group II intron ORFs. *RNA* 9, 14–9.

Datsenko, K. A., Pougach, K., Tikhonov, A., Wanner, B. L., Severinov, K., and Semenova, E. (2012). Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat Commun* 3, 945.

Deltcheva, E., Chylinski, K., Sharma, C. M., Gonzales, K., Chao, Y., Pirzada, Z. A., *et al*. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* 471, 602–7.

Deng, L., Garrett, R. A., Shah, S. A., Peng, X., and She, Q. (2013). A novel interference mechanism by a type IIIB CRISPR-Cmr module in *Sulfolobus*. *Mol Microbiol* 87, 1088–99.

Deveau, H., Garneau, J. E., and Moineau, S. (2010). CRISPR/Cas system and its role in phage-bacteria interactions. *Annu Rev Microbiol* 64, 475–93.

di Marzo-Veronese, F., Copeland, T., DeVico, A., Rahman, R., Oroszlan, S., Gallo, R., *et al*. (1986). Characterization of highly immunogenic p66/p51 as the reverse transcriptase of HTLV-III/LAV. *Science* 231, 1289–1291. doi:10.1126/science.2418504.

Díez-Villaseñor, C., Guzmán, N. M., Almendros, C., García-Martínez, J., and Mojica, F. J. (2013). CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of *Escherichia coli*. *RNA Biol* 10, 792–802.

Dillard, K. E., Brown, M. W., Johnson, N. V., Xiao, Y., Dolan, A., Hernandez, E., *et al*. (2018). Assembly and Translocation of a CRISPR-Cas Primed Acquisition Complex. *Cell* 175, 934–946.e15.

Dinsmore, P. K., and Klaenhammer, T. R. (1997). Molecular characterization of a genomic region in a *Lactococcus* bacteriophage that is involved in its sensitivity to the phage defense mechanism AbiA. *J Bacteriol* 179, 2949–57.

Dong, D., Ren, K., Qiu, X., Zheng, J., Guo, M., Guan, X., *et al*. (2016). The crystal structure of Cpf1 in complex with CRISPR RNA. *Nature* 532, 522–6.

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

Doulatov, S., Hodes, A., Dai, L., Mandhana, N., Liu, M., Deora, R., et al. (2004). Tropism switching in *Bordetella* bacteriophage defines a family of diversity-generating retroelements. *Nature* 431, 476–81.

Dugar, G., Herbig, A., Förstner, K. U., Heidrich, N., Reinhardt, R., Nieselt, K., *et al.* (2013). High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter jejuni* isolates. *PLoS Genet* 9, e1003495.

Dunn, A. K., Millikan, D. S., Adin, D. M., Bose, J. L., and Stabb, E. V. (2006). New rfp- and pES213-derived tools for analyzing symbiotic *Vibrio fischeri* reveal patterns of infection and lux expression *in situ*. *Appl Environ Microbiol* 72, 802–10.

Durmaz, E., and Klaenhammer, T. R. (2007). Abortive phage resistance mechanism AbiZ speeds the lysis clock to cause premature lysis of phage-infected *Lactococcus lactis*. *J Bacteriol* 189, 1417–25.

East-Seletsky, A., O'Connell, M. R., Burstein, D., Knott, G. J., and Doudna, J. A. (2017). RNA Targeting by Functionally Orthogonal Type VI-A CRISPR-Cas Enzymes. *Mol Cell* 66, 373–383.e3.

East-Seletsky, A., O'Connell, M. R., Knight, S. C., Burstein, D., Cate, J. H., Tjian, R., *et al.* (2016). Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature* 538, 270–273.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5, 113.

Eickbush, T. H., and Jamburuthugoda, V. K. (2008). The diversity of retrotransposons and the properties of their reverse transcriptases. *Virus Res* 134, 221–34.

Elmore, J. R., Sheppard, N. F., Ramia, N., Deighan, T., Li, H., Terns, R. M., *et al.* (2016). Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR-Cas system. *Genes Dev* 30, 447–59.

Emond, E., Holler, B. J., Boucher, I., Vandenbergh, P. A., Vedamuthu, E. R., Kondo, J. K., et al. (1997). Phenotypic and genetic characterization of the bacteriophage abortive infection mechanism AbiK from *Lactococcus lactis*. *Appl Environ Microbiol* 63, 1274–83.

Engelke, D. R., Krikos, A., Bruck, M. E., and Ginsburg, D. (1990). Purification of *Thermus aquaticus* DNA polymerase expressed in *Escherichia coli*. *Analytical Biochemistry* 191, 396–400. doi:10.1016/0003-2697(90)90238-5.

Enyeart, P. J., Chirieleison, S. M., Dao, M. N., Perutka, J., Quandt, E. M., Yao, J., *et al.* (2013). Generalized bacterial genome editing using mobile group II introns and Cre-lox. *Mol Syst Biol* 9, 685.

Enyeart, P. J., Mohr, G., Ellington, A. D., and Lambowitz, A. M. (2014). Biotechnological applications of mobile group II introns and their reverse transcriptases: gene targeting, RNA-seq, and non-coding RNA analysis. *Mob DNA* 5, 2.

Estrella, M. A., Kuo, F. T., and Bailey, S. (2016). RNA-activated DNA cleavage by the Type III-B CRISPR-Cas effector complex. *Genes Dev* 30, 460–70.

Fagerlund, R. D., Wilkinson, M. E., Klykov, O., Barendregt, A., Pearce, F. G., Kieper, S. N., *et al.* (2017). Spacer capture and integration by a type I-F Cas1-Cas2-3 CRISPR adaptation complex. *Proc Natl Acad Sci U S A* 114, E5122–E5128.

Farzadfard, F., and Lu, T. K. (2014). Synthetic biology. Genomically encoded analog memory with precise *in vivo* DNA writing in living cell populations. *Science* 346, 1256272.

Ferat, J. L., and Michel, F. (1993). Group II self-splicing introns in bacteria. *Nature* 364, 358–61.

Fineran, P. C., and Charpentier, E. (2012). Memory of viral infections by CRISPR-Cas adaptive immune systems: acquisition of new information. *Virology* 434, 202–9.

Fineran, P. C., Gerritzen, M. J., Suárez-Diez, M., Künne, T., Boekhorst, J., van Houte. S. A., *et al.* (2014). Degenerate target sites mediate rapid primed CRISPR adaptation. *Proc Natl Acad Sci U S A* 111, E1629–38.

Finnegan, D. J. (2012). Retrotransposons. *Curr Biol* 22, R432–7.

Fonfara, I., Le, R. A., Chylinski, K., Makarova, K. S., Lécrivain, A. L., Bzdrenga, J., *et al.* (2014). Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic Acids Res* 42, 2577–90.

Fonfara, I., Richter, H., Bratovič, M., Le, R. A., and Charpentier, E. (2016). The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* 532, 517–21.

Fortier, L. C., Bouchard, J. D., and Moineau, S. (2005). Expression and site-directed mutagenesis of the lactococcal abortive phage infection protein AbiK. *J Bacteriol* 187, 3721–30.

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

García-Rodríguez, F. M., Hernández-Gutiérrez, T., Díaz-Prado, V., and Toro, N. (2014). Use of the computer-retargeted group II intron RmInt1 of *Sinorhizobium meliloti* for gene targeting. *RNA Biol* 11, 391–401.

García-Rodríguez, F. M., Neira, J. L., Marcia, M., Molina-Sánchez, M. D., and Toro, N. (2019). A group II intron-encoded protein interacts with the cellular replicative machinery through the β-sliding clamp. *Nucleic Acids Res* 47, 7605–7617.

Garneau, J. E., Dupuis, M. È., Villion, M., Romero, D. A., Barrangou, R., Boyaval, P., *et al.* (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature* 468, 67–71.

Garrett, R. A., Vestergaard, G., and Shah, S. A. (2011). Archaeal CRISPR-based immune systems: exchangeable functional modules. *Trends Microbiol* 19, 549–56.

Garside, E. L., Schellenberg, M. J., Gesner, E. M., Bonanno, J. B., Sauder, J. M., Burley, S. K., *et al.* (2012). Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. *RNA* 18, 2020–8.

Gesner, E. M., Schellenberg, M. J., Garside, E. L., George, M. M., and Macmillan, A. M. (2011). Recognition and maturation of effector RNAs in a CRISPR interference pathway. *Nat Struct Mol Biol* 18, 688–92.

Gong, B., Shin, M., Sun, J., Jung, C. H., Bolt, E. L., van der Oost, J., *et al.* (2014). Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. *Proc Natl Acad Sci U S A* 111, 16359–64.

Grainy, J., Garrett, S., Graveley, B. R., and Terns, M.P., (2019). CRISPR repeat sequences and relative spacing specify DNA integration by *Pyrococcus furiosus* Cas1 and Cas2. *Nucleic Acids Res* 47, 7518–7531.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007a). The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* 8, 172.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007b). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res* 35, W52–7.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* 59, 307–21.

Gulig, P. A., Tucker, M. S., Thiaville, P. C., Joseph, J. L., and Brown, R. N. (2009). USER friendly cloning coupled with chitin-based natural transformation enables rapid mutagenesis of *Vibrio vulnificus*. *Appl Environ Microbiol* 75, 4936–49.

Guo, H., Arambula, D., Ghosh, P., and Miller, J. F. (2014). Diversity-generating Retroelements in Phage and Bacterial Genomes. *Microbiol Spectr* 2.

Guo, H., Karberg, M., Long, M., Jones, J. P. 3rd, Sullenger, B., and Lambowitz, A. M. (2000). Group II introns designed to insert into therapeutically relevant DNA target sites in human cells. *Science* 289, 452–7.

Guo, H., Tse, L. V., Nieh, A. W., Czornyj, E., Williams, S., Oukil, S., *et al.* (2011). Target site recognition by a diversity-generating retroelement. *PLoS Genet* 7, e1002414.

Gwee, C. P., Khoo, C. H., Yeap, S. K., Tan, G. C., and Cheah, Y. K. (2019). 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* 7, e5989.

Haack, D. B., and Toor, N. (2020). Retroelement origins of pre-mRNA splicing. *Wiley Interdiscip Rev RNA* 11, e1589.

Hanada, S., Takaichi, S., Matsuura, K., and Nakamura, K. (2002*). Roseiflexus castenholzii* gen. nov. sp. nov., a thermophilic, filamentous, photosynthetic bacterium that lacks chlorosomes. *Int J Sys Evol Microbiol* 52, 187–193. doi:10.1099/00207713-52-1-187.

Handa, S., Jiang, Y., Tao, S., Foreman, R., Schinazi, R. F., Miller, J. F., *et al.* (2018). Template-assisted synthesis of adenine-mutagenized cDNA by a retroelement protein complex. *Nucleic Acids Res* 46, 9711–9725.

Handa, S., Paul, B. G., Miller, J. F., Valentine, D. L., and Ghosh, P. (2016). Conservation of the C-type lectin fold for accommodating massive sequence variation in archaeal diversity-generating retroelements. *BMC Struct Biol* 16, 13.

Handa, S., Shaw, K. L., and Ghosh, P. (2019). Crystal structure of a *Thermus aquaticus* diversity-generating retroelement variable protein. *PLoS One* 14, e0205618.

Harrington, L. B., Burstein, D., Chen, J. S., Paez-Espino, D., Ma, E., Witte, I. P., *et al.* (2018). Programmed DNA destruction by miniature CRISPR-Cas14 enzymes. *Science* 362, 839–842.

Hatoum-Aslan, A., Maniv, I., and Marraffini, L. A. (2011). Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site. *Proc Natl Acad Sci U S A* 108, 21218–22.

Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K., and Doudna, J. A. (2010). Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science* 329, 1355–8.

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

Hein, S., Scholz, I., Voß, B., and Hess, W. R. (2013). Adaptation and modification of three CRISPR loci in two closely related cyanobacteria. *RNA Biology* 10, 852–864. doi:10.4161/rna.24160.

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

Hernández-Cabanyero, C., and Amaro, C. (2020). Phylogeny and life cycle of the zoonotic pathogen *Vibrio vulnificus*. *Environ Microbiol* 22, 4133–4148.

Herrero, M., de, L. V., and Timmis, K. N. (1990). Transposon vectors containing non-antibiotic resistance selection markers for cloning and stable chromosomal insertion of foreign genes in gram-negative bacteria. *J Bacteriol* 172, 6557–67.

Hill, C., Miller, L. A., and Klaenhammer, T. R. (1990). Nucleotide sequence and distribution of the pTR2030 resistance determinant (hsp) which aborts bacteriophage infection in lactococci. *Appl Environ Microbiol* 56, 2255–8.

Hill, C., Pierce, K., and Klaenhammer, T. R. (1989). The conjugative plasmid pTR2030 encodes two bacteriophage defense mechanisms in lactococci, restriction modification (R+/M+) and abortive infection (Hsp+). *Appl Environ Microbiol* 55, 2416–9.

Hoikkala, V., Ravantti, J., Díez-Villaseñor, C., Tiirola, M., Conrad, R. A., McBride, M. J., *et al.* (2020). Cooperation between CRISPR-Cas types enables adaptation in an RNA-targeting system. Preprint at bioRxiv. doi:10.1101/2020.02.20.957498.

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

Imhoff, J. F. (2003). Phylogenetic taxonomy of the family Chlorobiaceae on the basis of 16S rRNA and fmo (Fenna-Matthews-Olson protein) gene sequences. *Int J Syst Evol Microbiol* 53, 941–951.

Inouye, S., Hsu, M. Y., Eagle, S., and Inouye, M. (1989). Reverse transcriptase associated with the biosynthesis of the branched RNA-linked msDNA in *Myxococcus xanthus*. *Cell* 56, 709–17.

Iyer, L. M. (2005). Origin and evolution of the archaeo-eukaryotic primase superfamily and related palm-domain proteins: structural insights and new members. *Nucleic Acids Research* 33, 3875–3896. doi:10.1093/nar/gki702.

Jackson, R. N., Golden, S. M., van Erp, P. B., Carter, J., Westra, E. R., Brouns, S. J., *et al.* (2014). Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from *Escherichia coli*. *Science* 345, 1473–9.

Jain, I., Minakhin, L., Mekler, V., Sitnik, V., Rubanova, N., Severinov, K., *et al.* (2019). Defining the seed sequence of the Cas12b CRISPR-Cas effector complex. *RNA Biol* 16, 413–422.

Jansen, R., Embden, J. D., Gaastra, W., and Schouls, L. M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* 43, 1565–75.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337, 816–21.

Jinek, M., Jiang, F., Taylor, D. W., Sternberg, S. H., Kaya, E., Ma, E., *et al.* (2014). Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* 343, 1247997.

Jore, M. M., Lundgren, M., van Duijn, E., Bultema, J. B., Westra, E. R., Waghmare, S. P., *et al.* (2011). Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nat Struct Mol Biol* 18, 529–36.

Karberg, M., Guo, H., Zhong, J., Coon, R., Perutka, J., and Lambowitz, A. M. (2001). Group II introns as controllable gene targeting vectors for genetic manipulation of bacteria. *Nat Biotechnol* 19, 1162–7.

Karvelis, T., Bigelyte, G., Young, J. K., Hou, Z., Zedaveinyte, R., Budre, K., *et al.* (2020). PAM recognition by miniature CRISPR-Cas12f nucleases triggers programmable double-stranded DNA target cleavage. *Nucleic Acids Res* 48, 5016–5023.

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30, 772–80.

Kazlauskas, D., Sezonov, G., Charpin, N., Česlovas Venclovas, Forterre, P., and Krupovic, M. (2018). Novel Families of Archaeo-Eukaryotic Primases Associated with Mobile Genetic Elements of Bacteria and Archaea. *J Mol Biol*430, 737–750. doi:10.1016/j.jmb.2017.11.014.

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

Kazlauskiene, M., Tamulaitis, G., Kostiuk, G., Venclovas, Č., and Siksnys, V. (2016). Spatiotemporal Control of Type III-A CRISPR-Cas Immunity: Coupling DNA Degradation with the Target RNA Recognition. *Mol Cell* 62, 295–306.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., *et al*. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28, 1647–9.

Kieper, S. N., Almendros, C., Behler, J., McKenzie, R. E., Nobrega, F. L., Haagsma, A. C., *et al*. (2018). Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation. *Cell Rep* 22, 3377–3384.

Kim, S., Loeff, L., Colombo, S., Jergic, S., Brouns, S. J. J., and Joo, C. (2020). Selective loading and processing of prespacers for precise CRISPR adaptation. *Nature* 579, 141–145.

Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S., and Sternberg, S. H. (2019). Transposon-encoded CRISPR-Cas systems direct RNA-guided DNA integration. *Nature* 571, 219–225.

Kojima, K. K., and Kanehisa, M. (2008). Systematic survey for novel types of prokaryotic retroelements based on gene neighborhood and protein architecture. *Mol Biol Evol* 25, 1395–404.

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

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

Krupovic, M., Makarova, K. S., Forterre, P., Prangishvili, D., and Koonin, E. V. (2014). Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. *BMC Biol* 12, 36.

Krupovic, M., Shmakov, S., Makarova, K. S., Forterre, P., and Koonin, E. V. (2016). Recent Mobility of Casposons, Self-Synthesizing Transposons at the Origin of the CRISPR-Cas Immunity. *Genome Biol Evol* 8, 375–86.

Kunin, V., Sorek, R., and Hugenholtz, P. (2007). Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* 8, R61.

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* 227, 680–5.

Lange, S. J., Alkhnbashi, O. S., Rose, D., Will, S., and Backofen, R. (2013). CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic Acids Res* 41, 8034–44.

LaRoche-Johnston, F., Bosan, R., and Cousineau, B. (2020). Group II introns generate functional chimeric relaxase enzymes with modified specificities through exon shuffling at both the RNA and DNA level. *Mol Biol Evol.*

Lambowitz, A. M., and Belfort, M. (2015). Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. *Microbiol Spectr* 3, MDNA3–0050-2014.

Lambowitz, A. M., and Zimmerly, S. (2011). Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb Perspect Biol* 3, a003616.

Lambowitz, A. M., and Zimmerly, S. (2004). Mobile group II introns. *Annu Rev Genet* 38, 1–35.

Lampson, B. C., Inouye, M., and Inouye, S. (1989). Reverse transcriptase with concomitant ribonuclease H activity in the cell-free synthesis of branched RNA-linked msDNA of *Myxococcus xanthus*. *Cell* 56, 701–7.

Lampson, B. C., Inouye, M., and Inouye, S. (2005). Retrons, msDNA, and the bacterial genome. *Cytogenet Genome Res* 110, 491–9.

Le Coq, J., and Ghosh, P. (2011). Conservation of the C-type lectin fold for massive sequence variation in a *Treponema* diversity-generating retroelement. *Proc Natl Acad Sci U S A* 108, 14649–53.

Leclercq, S., and Cordaux, R. (2012). Selection-driven extinction dynamics for group II introns in Enterobacteriales. *PLoS One* 7, e52268.

Lee, H., Dhingra, Y., and Sashital, D. G. (2019). The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. *Elife* 8.

Lee, H., Zhou, Y., Taylor, D. W., and Sashital, D. G. (2018). Cas4-Dependent Prespacer Processing Ensures High-Fidelity Programming of CRISPR Arrays. *Mol Cell* 70, 48–59.e5.

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

Lim, D., and Maas, W. K. (1989). Reverse transcriptase-dependent synthesis of a covalently linked, branched DNA-RNA compound in *E. coli* B. *Cell* 56, 891–904.

Lim, H., Jun, S., Park, M., Lim, J., Jeong, J., Lee, J. H., *et al.* (2020). Multiplex Generation, Tracking, and Functional Screening of Substitution Mutants Using a CRISPR/Retron System. *ACS Synth Biol* 9, 1003–1009.

Liu, M., Deora, R., Doulatov, S. R., Gingery, M., Eiserling, F. A., Preston, A., *et al.* (2002). Reverse transcriptase-mediated tropism switching in *Bordetella* bacteriophage. *Science* 295, 2091–4.

Liu, L., Li, X., Ma, J., Li, Z., You, L., Wang, J., *et al.* (2017). The Molecular Architecture for RNA-Guided RNA Cleavage by Cas13a. *Cell* 170, 714–726.e10.

Lopes, A., Amarir-Bouhram, J., Faure, G., Petit, M. A., and Guerois, R. (2010). Detection of novel recombinases in bacteriophage genomes unveils Rad52, Rad51 and Gp2.5 remote homologs. *Nucleic Acids Res* 38, 3952–62.

Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I., and Koonin, E. V. (2006). A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct* 1, 7.

Makarova, K. S., Haft, D. H., Barrangou, R., Brouns, S. J., Charpentier, E., Horvath, P., *et al.* (2011). Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol* 9, 467–77.

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

Makarova, K. S., Wolf, Y. I., Alkhnbashi, O. S., Costa, F., Shah, S. A., Saunders, S. J., *et al.* (2015). An updated evolutionary classification of CRISPR-Cas systems. *Nat Rev Microbiol* 13, 722–36.

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

Mao, J. R., Shimada, M., Inouye, S., and Inouye, M. (1995). Gene regulation by antisense DNA produced *in vivo*. *J Biol Chem* 270, 19684–7.

Marraffini, L. A., and Sontheimer, E. J. (2010). Self versus non-self discrimination during CRISPR RNA-directed immunity. *Nature* 463, 568–71.

Martin, W., and Koonin, E. V. (2006). Introns and the origin of nucleus-cytosol compartmentalization. *Nature* 440, 41–5.

Martín-Alonso, S., Frutos-Beltrán, E., and Menéndez-Arias, L. (2020). Reverse Transcriptase: From Transcriptomics to Genome Editing. *Trends Biotechnol.*

Martínez-Abarca, F., Barrientos-Durán, A., Fernández-López, M., and Toro, N. (2004). The RmInt1 group II intron has two different retrohoming pathways for mobility using predominantly the nascent lagging strand at DNA replication forks for priming. *Nucleic Acids Res* 32, 2880–8.

Matsuura, M., Saldanha, R., Ma, H., Wank, H., Yang, J., Mohr, G., *et al*. (1997). A bacterial group II intron encoding reverse transcriptase maturase, and DNA endonuclease activities: biochemical demonstration of maturase activity and insertion of new genetic information within the intron. *Genes & Development* 11, 2910–2924. doi:10.1101/gad.11.21.2910.

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

McGinn, J., and Marraffini, L. A. (2016). CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. *Mol Cell* 64, 616–623.

McKenzie, R. E., Almendros, C., Vink, J. N. A., and Brouns, S. J. J. (2019). Using CAPTURE to detect spacer acquisition in native CRISPR arrays. *Nat Protoc* 14, 976–990.

McMahon, S. A., Miller, J. L., Lawton, J. A., Kerkow, D. E., Hodes, A., Marti-Renom, M. A., et al. (2005). The C-type lectin fold as an evolutionary solution for massive sequence variation. *Nat Struct Mol Biol* 12, 886–92.

Medhekar, B., and Miller, J. F. (2007). Diversity-generating retroelements. *Curr Opin Microbiol* 10, 388–95.

Meeske, A. J., and Marraffini, L. A. (2018). RNA Guide Complementarity Prevents Self-Targeting in Type VI CRISPR Systems. *Mol Cell* 71, 791–801.e3.

Meeske, A. J., Nakandakari-Higa, S., and Marraffini, L. A. (2019). Cas13-induced cellular dormancy prevents the rise of CRISPR-resistant bacteriophage. *Nature* 570, 241–245.

Menéndez-Arias, L., Sebastián-Martín, A., and Álvarez, M. (2017). Viral reverse transcriptases. *Virus Res* 234, 153–176.

Mestre, M. R., González-Delgado, A., Gutiérrez-Rus, L. I., Martínez-Abarca, F., and Toro, N. (2020). Systematic prediction of genes functionally associated with bacterial retrons and classification of the encoded tripartite systems. *Nucleic Acids Res* 48, 12632–12647.

Michel, F., Costa, M., and Westhof, E. (2009). The ribozyme core of group II introns: a structure in want of partners. *Trends Biochem Sci* 34, 189–99.

Michel, F., and Ferat, J. L. (1995). Structure and activities of group II introns. *Annu Rev Biochem* 64, 435–61.

Miller J.H. (1972). Experiments in molecular genetics. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory.

Millman, A., Bernheim, A., Stokar-Avihail, A., Fedorenko, T., Voichek, M., Leavitt, A., *et al.* (2020). Bacterial retrons function in anti-phage defense. *Cell* 10;183(6):1551-1561.e12.

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

Mohr, G., Silas, S., Stamos, J. L., Makarova, K. S., Markham, L. M., Yao, J., *et al.* (2018). A Reverse Transcriptase-Cas1 Fusion Protein Contains a Cas6 Domain Required for Both CRISPR RNA Biogenesis and RNA Spacer Acquisition. *Mol Cell* 72, 700–714.e8.

Mohr, G., Smith, D., Belfort, M., and Lambowitz, A. M. (2000). Rules for DNA target-site recognition by a lactococcal group II intron enable retargeting of the intron to specific DNA sequences. *Genes Dev* 14, 559–73.

Mohr, S., Ghanem, E., Smith, W., Sheeter, D., Qin, Y., King, O., *et al.* (2013). Thermostable group II intron reverse transcriptase fusion proteins and their use in cDNA synthesis and next-generation RNA sequencing. *RNA* 19, 958–70.

Mojica, F. J., Díez-Villaseñor, C., García-Martínez, J., and Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* 60, 174–82.

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

Mojica, F. J., Juez, G., and Rodríguez-Valera, F. (1993). Transcription at different salinities of *Haloferax mediterranei* sequences adjacent to partially modified PstI sites. *Mol Microbiol* 9, 613–21.

Moran, J. V., Zimmerly, S., Eskes, R., Kennell, J. C., Lambowitz, A. M., Butow, R. A., *et al*. (1995). Mobile group II introns of yeast mitochondrial DNA are novel site-specific retroelements. *Molecular and Cellular Biology* 15, 2828–2838. doi:10.1128/mcb.15.5.2828.

Mulepati, S., Héroux, A., and Bailey, S. (2014). Structural biology. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. *Science* 345, 1479–84.

Müller, A. U., Leibundgut, M., Ban, N., and Weber-Ban, E. (2019). Structure and functional implications of WYL domain-containing bacterial DNA damage response regulator PafBC. *Nature Communications* 10. doi:10.1038/s41467-019-12567-x.

Nam, K. H., Haitjema, C., Liu, X., Ding, F., Wang, H., DeLisa, M. P., *et al.* (2012). Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. *Structure* 20, 1574–84.

Naorem, S. S., Han, J., Wang, S., Lee, W. R., Heng, X., Miller, J. F., *et al.* (2017). DGR mutagenic transposition occurs via hypermutagenic reverse transcription primed by nicked template RNA. *Proc Natl Acad Sci U S A* 114, E10187–E10195.

Nguyen, L. T., Schmidt, H. A., von, H. A., and Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* 32, 268–74.

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

Nimkulrat, S., Lee, H., Doak, T. G., and Ye, Y. (2016). Genomic and Metagenomic Analysis of Diversity-Generating Retroelements Associated with *Treponema denticola*. *Front Microbiol* 7, 852.

Nisa-Martínez, R., Jiménez-Zurdo, J. I., Martínez-Abarca, F., Muñoz-Adelantado, E., and Toro, N. (2007). Dispersion of the RmInt1 group II intron in the *Sinorhizobium meliloti* genome upon acquisition by conjugative transfer. *Nucleic Acids Res* 35, 214–22.

Nisa-Martínez, R., Molina-Sánchez, M. D., and Toro, N. (2016). Host Factors Influencing the Retrohoming Pathway of Group II Intron RmInt1, Which Has an Intron-Encoded Protein Naturally Devoid of Endonuclease Activity. *PLoS One* 11, e0162275.

Nishimasu, H., Ran, F. A., Hsu, P. D., Konermann, S., Shehata, S. I., Dohmae, N., *et al*. (2014). Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* 156, 935–49.

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

Novikova, O., and Belfort, M. (2017). Mobile Group II Introns as Ancestral Eukaryotic Elements. *Trends Genet* 33, 773–783.

Novikova, O., Smith, D., Hahn, I., Beauregard, A., and Belfort, M. (2014). Interaction between conjugative and retrotransposable elements in horizontal gene transfer. *PLoS Genet* 10, e1004853.

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

Nuñez, J. K., Bai, L., Harrington, L. B., Hinder, T. L., and Doudna, J. A. (2016). CRISPR Immunological Memory Requires a Host Factor for Specificity. *Mol Cell* 62, 824–833.

Nuñez, J. K., Harrington, L. B., Kranzusch, P. J., Engelman, A. N., and Doudna, J. A. (2015b). Foreign DNA capture during CRISPR-Cas adaptive immunity. *Nature* 527, 535–8.

Nuñez, J. K., Kranzusch, P. J., Noeske, J., Wright, A. V., Davies, C. W., and Doudna, J. A. (2014). Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat Struct Mol Biol* 21, 528–34.

Nuñez, J. K., Lee, A. S., Engelman, A., and Doudna, J. A. (2015a). Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. *Nature* 519, 193–8.

O'Connell, M. R. (2019). Molecular Mechanisms of RNA Targeting by Cas13-containing Type VI CRISPR-Cas Systems. *J Mol Biol* 431, 66–87.

Odegrip, R., Nilsson, A. S., and Haggård-Ljungquist, E. (2006). Identification of a gene encoding a functional reverse transcriptase within a highly variable locus in the P2-like coliphages. *J Bacteriol* 188, 1643–7.

Oude-Elferink. S. J., Akkermans-van Vliet, W. M., Bogte, J. J., and Stams, A. J. (1999). *Desulfobacca acetoxidans* gen. nov., sp. nov., a novel acetate-degrading sulfate reducer isolated from sulfidogenic granular sludge. *Int J Syst Bacteriol* 49 Pt 2, 345–50.

Özcan, A., Pausch, P., Linden, A., Wulf, A., Schühle, K., Heider, J., *et al*. (2019). Type IV CRISPR RNA processing and effector complex formation in *Aromatoleum aromaticum. Nat Microbiol* 4, 89–96.

Park, J., Zhang, Y., Buboltz, A. M., Zhang, X., Schuster, S. C., Ahuja, U., *et al*. (2012). Comparative genomics of the classical *Bordetella* subspecies: the evolution and exchange of virulence-associated diversity amongst closely related pathogens. *BMC Genomics* 13, 545.

Paul, B. G., Burstein, D., Castelle, C. J., Handa, S., Arambula, D., Czornyj, E., *et al*. (2017). Retroelement-guided protein diversification abounds in vast lineages of Bacteria and Archaea. *Nat Microbiol* 2, 17045.

Peng, W., Feng, M., Feng, X., Liang, Y. X., and She, Q. (2015). An archaeal CRISPR type III-B system exhibiting distinctive RNA targeting features and mediating dual RNA and DNA interference. *Nucleic Acids Res* 43, 406–17.

Pfeffer, C., and Oliver, J. D. (2003). A comparison of thiosulphate-citrate-bile salts-sucrose (TCBS) agar and thiosulphate-chloride-iodide (TCI) agar for the isolation of *Vibrio* species from estuarine environments. *Lett Appl Microbiol* 36, 150–1.

Pinilla-Redondo, R., Mayo-Muñoz, D., Russel, J., Garrett, R. A., Randau, L., Sørensen, S. J., *et al*. (2020). Type IV CRISPR-Cas systems are highly diverse and involved in competition between plasmids. *Nucleic Acids Res* 48, 2000–2012.

Ploquin, M., Bransi, A., Paquet, E. R., Stasiak, A. Z., Stasiak, A., Yu, X., *et al*. (2008). Functional and structural basis for a bacteriophage homolog of human RAD52. *Curr Biol* 18, 1142–6.

Pougach, K., Semenova, E., Bogdanova, E., Datsenko, K. A., Djordjevic, M., Wanner, B. L., *et al*. (2010). Transcription, processing and function of CRISPR cassettes in *Escherichia coli*. *Mol Microbiol* 77, 1367–79.

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

Pyenson, N. C., and Marraffini, L. A. (2017). Type III CRISPR-Cas systems: when DNA cleavage just isn't enough. *Curr Opin Microbiol* 37, 150–154.

Qu, G., Piazza, C. L., Smith, D., and Belfort, M. (2018). Group II intron inhibits conjugative relaxase expression in bacteria by mRNA targeting. *Elife* 7.

Redding, S., Sternberg, S. H., Marshall, M., Gibb, B., Bhat, P., Guegler, C. K., *et al.* (2015). Surveillance and Processing of Foreign DNA by the *Escherichia coli* CRISPR-Cas System. *Cell* 163, 854–65.

Reeks, J., Sokolowski, R. D., Graham, S., Liu, H., Naismith, J. H., and White, M. F. (2013). Structure of a dimeric crenarchaeal Cas6 enzyme with an atypical active site for CRISPR RNA processing. *Biochem J* 452, 223–30.

Reinoso-Colacio, M., García-Rodríguez, F. M., García-Cañadas, M., Amador-Cubero, S., García, P. J. L., and Toro, N. (2015). Localization of a bacterial group II intron-encoded protein in human cells. *Sci Rep* 5, 12716.

Rippka, R., Stanier, R. Y., Deruelles, J., Herdman, M., and Waterbury, J. B. (1979). Generic Assignments Strain Histories and Properties of Pure Cultures of Cyanobacteria. *Microbiology* 111, 1–61. doi:10.1099/00221287-111-1-1.

Rollie, C., Graham, S., Rouillon, C., and White, M. F. (2018). Prespacer processing and specific integration in a Type I-A CRISPR system. *Nucleic Acids Res* 46, 1007–1020.

Rollie, C., Schneider, S., Brinkmann, A. S., Bolt, E. L., and White, M. F. (2015). Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. *Elife* 4.

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

Roux, S., Paul, B. G., Bagby, S. C., Allen, M. A., Attwood, G., Cavicchioli, R., et al. (2020). Ecology and molecular targets of hypermutation in the global microbiome. Preprint at bioRxiv. doi:10.1101/2020.04.01.020958.

San Filippo, J., and Lambowitz, A. M. (2002). Characterization of the C-terminal DNA-binding/DNA endonuclease region of a group II intron-encoded protein. *J Mol Biol* 324, 933–51.

Samai, P., Pyenson, N., Jiang, W., Goldberg, G. W., Hatoum-Aslan, A., and Marraffini, L. A. (2015). Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. *Cell* 161, 1164–1174.

Sambrook, J., Fritsch E.F., Maniatis T. (1989). Molecular cloning: a laboratory manual. 2nd ed. Edn. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory.

Sashital, D. G., Jinek, M., and Doudna, J. A. (2011). An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. *Nat Struct Mol Biol* 18, 680–7.

Scaltriti, E., Launay, H., Genois, M. M., Bron, P., Rivetti, C., Grolli, S., *et al.* (2011). Lactococcal phage p2 ORF35-Sak3 is an ATPase involved in DNA recombination and AbiK mechanism. *Mol Microbiol* 80, 102–16.

Scaltriti, E., Moineau, S., Launay, H., Masson, J. Y., Rivetti, C., Ramoni, R., *et al.* (2010). Deciphering the function of lactococcal phage ul36 Sak domains. *J Struct Biol* 170, 462–9.

Schillinger, T., Lisfi, M., Chi, J., Cullum, J., and Zingler, N. (2012). Analysis of a comprehensive dataset of diversity generating retroelements generated by the program DiGReF. *BMC Genomics* 13, 430.

Schillinger, T., and Zingler, N. (2012). The low incidence of diversity-generating retroelements in sequenced genomes. *Mob Genet Elements* 2, 287–291.

Schmidt, F., Cherepkova, M. Y., and Platt, R. J. (2018). Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* 562, 380–385.

Schubert, M. G., Goodman, D. B., Wannier, T. M., Kaur, D., Farzadfard, F., Lu, T. K., *et al.* (2020). High throughput functional variant screens via in-vivo production of single-stranded DNA. Preprint at bioRxiv. doi:10.1101/2020.03.05.975441.

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

Shao, Y., and Li, H. (2013). Recognition and cleavage of a nonstructured CRISPR RNA by its processing endoribonuclease Cas6. *Structure* 21, 385–93.

Sharifi, F., and Ye, Y. (2019). MyDGR: a server for identification and characterization of diversity-generating retroelements. *Nucleic Acids Res* 47, W289–W294.

Sharon, E., Chen, S. A., Khosla, N. M., Smith, J. D., Pritchard, J. K., and Fraser, H. B. (2018). Functional Genetic Variants Revealed by Massively Parallel Precise Genome Editing. *Cell* 175, 544–557.e16.

Shih, P. M., Wu, D., Latifi, A., Axen, S. D., Fewer, D. P., Talla, E., *et al.* (2013). Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing. *Proc Natl Acad Sci U S A* 110, 1053–8.

Shiimori, M., Garrett, S. C., Graveley, B. R., and Terns, M. P. (2018). Cas4 Nucleases Define the PAM, Length, and Orientation of DNA Fragments Integrated at CRISPR Loci. *Mol Cell* 70, 814–824.e6.

Shimamoto, T., Inouye, M., and Inouye, S. (1995). 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* 270, 581–8.

Shipman, S. L., Nivala, J., Macklis, J. D., and Church, G. M. (2017). CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature* 547, 345–349.

Shmakov, S., Abudayyeh, O. O., Makarova, K. S., Wolf, Y. I., Gootenberg, J. S., Semenova, E., *et al.* (2015). Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. *Mol Cell* 60, 385–97.

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

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

Silas, S., Makarova, K. S., Shmakov, S., Páez-Espino, D., Mohr, G., Liu, Y., *et al.* (2017a). On the Origin of Reverse Transcriptase-Using CRISPR-Cas Systems and Their Hyperdiverse, Enigmatic Spacer Repertoires. *mBio* 8.

Silas, S., Mohr, G., Sidote, D. J., Markham, L. M., Sanchez-Amat, A., Bhaya, D., *et al.* (2016). Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. *Science* 351, aad4234.

Simon, A. J., Ellington, A. D., and Finkelstein, I. J. (2019). Retrons and their applications in genome engineering. *Nucleic Acids Res* 47, 11007–11019.

Simon, A. J., Morrow, B. R., and Ellington, A. D. (2018). Retroelement-Based Genome Editing and Evolution. *ACS Synth Biol* 7, 2600–2611.

Simon, D. M., and Zimmerly, S. (2008). A diversity of uncharacterized reverse transcriptases in bacteria. *Nucleic Acids Res* 36, 7219–29.

Sinkunas, T., Gasiunas, G., Fremaux, C., Barrangou, R., Horvath, P., and Siksnys, V. (2011). Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. *EMBO J* 30, 1335–42.

Smargon, A. A., Cox, D. B. T., Pyzocha, N. K., Zheng, K., Slaymaker, I. M., Gootenberg, J. S., *et al*. (2017). Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28. *Mol Cell* 65, 618–630.e7.

Spaink, H. P., Okker, R. J., Wijffelman, C. A., Pees, E., and Lugtenberg, B. J. (1987). Promoters in the nodulation region of the *Rhizobium leguminosarum* Sym plasmid pRL1JI. *Plant Mol Biol* 9, 27–39.

Staals, R. H., Jackson, S. A., Biswas, A., Brouns, S. J., Brown, C. M., and Fineran, P. C. (2016). Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR-Cas system. *Nat Commun* 7, 12853.

Staals, R. H., Zhu, Y., Taylor, D. W., Kornfeld, J. E., Sharma, K., Barendregt, A., *et al*. (2014). RNA targeting by the type III-A CRISPR-Cas Csm complex of *Thermus thermophilus*. *Mol Cell* 56, 518–30.

Stern, A., Keren, L., Wurtzel, O., Amitai, G., and Sorek, R. (2010). Self-targeting by CRISPR: gene regulation or autoimmunity? *Trends Genet* 26, 335–40.

Strecker, J., Jones, S., Koopal, B., Schmid-Burgk, J., Zetsche, B., Gao, L., *et al*. (2019). Engineering of CRISPR-Cas12b for human genome editing. *Nat Commun* 10, 212.

Strecker, J., Ladha, A., Gardner, Z., Schmid-Burgk, J. L., Makarova, K. S., Koonin, E. V., *et al*. (2019). RNA-guided DNA insertion with CRISPR-associated transposases. *Science* 365, 48–53.

Studier, F. W. (1991). Use of bacteriophage T7 lysozyme to improve an inducible T7 expression system. *J Mol Biol* 219, 37–44.

Swarts, D. C., and Jinek, M. (2019). Mechanistic Insights into the cis- and trans-Acting DNase Activities of Cas12a. *Mol Cell* 73, 589–600.e4.

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

Swarts, D. C., van der Oost, J., and Jinek, M. (2017). Structural Basis for Guide RNA Processing and Seed-Dependent DNA Targeting by CRISPR-Cas12a. *Mol Cell* 66, 221–233.e4.

Tamulaitis, G., Kazlauskiene, M., Manakova, E., Venclovas, Č., Nwokeoji, A. O., Dickman, M. J., *et al.* (2014). Programmable RNA shredding by the type III-A CRISPR-Cas system of *Streptococcus thermophilus*. *Mol Cell* 56, 506–17.

Tangney, M., and Fitzgerald, G. F. (2002). Effectiveness of the lactococcal abortive infection systems AbiA, AbiE, AbiF and AbiG against P335 type phages. *FEMS Microbiol Lett* 210, 67–72.

Tangney, M., and Fitzgerald, G. F. (2002). AbiA, a lactococcal abortive infection mechanism functioning in *Streptococcus thermophilus*. *Appl Environ Microbiol* 68, 6388–91.

Tanna, T., Ramachanderan, R., and Platt, R. J. (2020). Engineered bacteria to report gut function: technologies and implementation. *Curr Opin Microbiol* 59, 24–33.

Tanna, T., Schmidt, F., Cherepkova, M. Y., Okoniewski, M., and Platt, R. J. (2020). Recording transcriptional histories using Record-seq. *Nat Protoc* 15, 513–539.

Taylor, D. W., Zhu, Y., Staals, R. H., Kornfeld, J. E., Shinkai, A., van der Oost, J., *et al.* (2015). Structural biology. Structures of the CRISPR-Cmr complex reveal mode of RNA target positioning. *Science* 348, 581–5.

Temin, H. M., and Mizutani, S. (1970). RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature* 226, 1211–3.

Toro, N. (2003). Bacteria and Archaea Group II introns: additional mobile genetic elements in the environment. *Environmental Microbiology* 5, 143–151. doi:10.1046/j.1462-2920.2003.00398.x.

Toro, N., Martínez-Abarca, F., Mestre, M. R., and González-Delgado, A. (2019a). Multiple origins of reverse transcriptases linked to CRISPR-Cas systems. *RNA Biol* 16, 1486–1493.

Toro, N., Mestre, M. R., Martínez-Abarca, F., and González-Delgado, A. (2019b). Recruitment of Reverse Transcriptase-Cas1 Fusion Proteins by Type VI-A CRISPR-Cas Systems. *Front Microbiol* 10, 2160.

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

Vallota-Eastman, A., Arrington, E. C., Meeken, S., Roux, S., Dasari, K., Rosen, S., *et al.* (2020). Role of diversity-generating retroelements for regulatory pathway tuning in cyanobacteria. *BMC Genomics* 21, 664.

van der Oost. J., Westra, E. R., Jackson, R. N., and Wiedenheft, B. (2014). Unravelling the structural and mechanistic basis of CRISPR-Cas systems. *Nat Rev Microbiol* 12, 479–92.

Velázquez, E., Lorenzo, V., and Al-Ramahi, Y. (2019). Recombination-Independent Genome Editing through CRISPR/Cas9-Enhanced TargeTron Delivery. *ACS Synth Biol* 8, 2186–2193.

Waite, D. W., Vanwonterghem, I., Rinke, C., Parks, D. H., Zhang, Y., Takai, K., *et al.* (2017). Comparative Genomic Analysis of the Class Epsilonproteobacteria and Proposed Reclassification to Epsilonbacteraeota (phyl. nov.). *Frontiers in Microbiology* 8. doi:10.3389/fmicb.2017.00682.

Waldern, J., Schiraldi, N. J., Belfort, M., and Novikova, O. (2020). Bacterial Group II Intron Genomic Neighborhoods Reflect Survival Strategies: Hiding and Hijacking. *Mol Biol Evol* 37, 1942–1948.

Wan, H., Li, J., Chang, S., Lin, S., Tian, Y., Tian, X., *et al.* (2019). Probing the Behaviour of Cas1-Cas2 upon Protospacer Binding in CRISPR-Cas Systems using Molecular Dynamics Simulations. *Sci Rep* 9, 3188.

Wang, C., Villion, M., Semper, C., Coros, C., Moineau, S., and Zimmerly, S. (2011). A reverse transcriptase-related protein mediates phage resistance and polymerizes untemplated DNA *in vitro*. *Nucleic Acids Res* 39, 7620–9.

Wang, G. H., and Seeger, C. (1992). The reverse transcriptase of hepatitis B virus acts as a protein primer for viral DNA synthesis. *Cell* 71, 663–70.

Wang, J., Li, J., Zhao, H., Sheng, G., Wang, M., Yin, M., *et al.* (2015). Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. *Cell* 163, 840–53.

Wang, J. Y., Hoel, C. M., Al-Shayeb, B., Banfield, J. F., Brohawn, S. G., and Doudna, J. A. (2020). Structural coordination between active sites of a Cas6-reverse transcriptase-Cas1Cas2 CRISPR integrase complex. Preprint at bioRxiv. doi:10.1101/2020.10.18.344481.

Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L., Driscoll, T., Gabbard, J. L., *et al.* (2014). PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res* 42, D581–91.

Wei, Y., Terns, R. M., and Terns, M. P. (2015). Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. *Genes Dev* 29, 356–61.

Wen, Z., Lu, M., Ledesma-Amaro, R., Li, Q., Jin, M., and Yang, S. (2020). TargeTron Technology Applicable in Solventogenic Clostridia: Revisiting 12 Years' Advances. *Biotechnol J* 15, e1900284.

Westra, E. R., Nilges, B., van Erp, P. B., van der Oost, J., Dame, R. T., and Brouns, S. J. (2012). Cascade-mediated binding and bending of negatively supercoiled DNA. *RNA Biol* 9, 1134–8.

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

Wilkinson, M., Drabavicius, G., Silanskas, A., Gasiunas, G., Siksnys, V., and Wigley, D. B. (2019). Structure of the DNA-Bound Spacer Capture Complex of a Type II CRISPR-Cas System. *Mol Cell* 75, 90–101.e5.

Williams, T. C., Blackman, E. R., Morrison, S. S., Gibas, C. J., and Oliver, J. D. (2014). Transcriptome sequencing reveals the virulence and environmental genetic programs of *Vibrio vulnificus* exposed to host and estuarine conditions. *PLoS One* 9, e114376.

Wolf, Y. I., Silas, S., Wang, Y., Wu, S., Bocek, M., Kazlauskas, D., *et al.* (2020). Doubling of the known set of RNA viruses by metagenomic analysis of an aquatic virome. *Nature Microbiology* 5, 1262–1270. doi:10.1038/s41564-020-0755-4.

Wright, A. V., and Doudna, J. A. (2016). Protecting genome integrity during CRISPR immune adaptation. *Nat Struct Mol Biol* 23, 876–883.

Wright, A. V., Liu, J. J., Knott, G. J., Doxzen, K. W., Nogales, E., and Doudna, J. A. (2017). Structures of the CRISPR genome integration complex. *Science* 357, 1113–1118.

Wright, A. V., Wang, J. Y., Burstein, D., Harrington, L. B., Paez-Espino, D., Kyrpides, N. C., *et al.* (2019). A Functional Mini-Integrase in a Two-Protein-type V-C CRISPR System. *Mol Cell* 73, 727–737.e3.

Wu, L., Gingery, M., Abebe, M., Arambula, D., Czornyj, E., Handa, S., *et al.* (2018). Diversity-generating retroelements: natural variation, classification and evolution inferred from a large-scale genomic survey. *Nucleic Acids Res* 46, 11–24.

Xiao, Y., Luo, M., Hayes, R. P., Kim, J., Ng, S., Ding, F., *et al.* (2017). Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System. *Cell* 170, 48–60.e11.

Xiao, Y., Ng, S., Nam, K. H., and Ke, A. (2017). How type II CRISPR-Cas establish immunity through Cas1-Cas2-mediated spacer integration. *Nature* 550, 137–141.

Xiong, Y., and Eickbush, T. H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO J* 9, 3353–62.

Xue, C., Whitis, N. R., and Sashital, D. G. (2016). Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity. *Mol Cell* 64, 826–834.

Yan, F., Yu, X., Duan, Z., Lu, J., Jia, B., Qiao, Y., *et al.* (2019). Discovery and characterization of the evolution, variation and functions of diversity-generating retroelements using thousands of genomes and metagenomes. *BMC Genomics* 20, 595.

Yan, W. X., Chong, S., Zhang, H., Makarova, K. S., Koonin, E. V., Cheng, D. R., *et al.* (2018). Cas13d Is a Compact RNA-Targeting Type VI CRISPR Effector Positively Modulated by a WYL-Domain-Containing Accessory Protein. *Mol Cell* 70, 327–339.e5.

Yan, W. X., Hunnewell, P., Alfonse, L. E., Carte, J. M., Keston-Smith, E., Sothiselvam, S., *et al.* (2019). Functionally diverse type V CRISPR-Cas systems. *Science* 363, 88–91.

Yang, H., Gao, P., Rajashankar, K. R., and Patel, D. J. (2016). PAM-Dependent Target DNA Recognition and Cleavage by C2c1 CRISPR-Cas Endonuclease. *Cell* 167, 1814–1828.e12.

Ye, Y. (2014). Identification of diversity-generating retroelements in human microbiomes. *Int J Mol Sci* 15, 14234–46.

Yee, T., Furuichi, T., Inouye, S., and Inouye, M. (1984). Multicopy single-stranded DNA isolated from a gram-negative bacterium, *Myxococcus xanthus*. *Cell* 38, 203–9.

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

Yosef, I., Goren, M. G., and Qimron, U. (2012). Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res* 40, 5569–76.

Yuan, T. Z., Overstreet, C. M., Moody, I. S., and Weiss, G. A. (2013). Protein engineering with biosynthesized libraries from *Bordetella bronchiseptica* bacteriophage. *PLoS One* 8, e55617.

Zebec, Z., Manica, A., Zhang, J., White, M. F., and Schleper, C. (2014). CRISPR-mediated targeted mRNA degradation in the archaeon *Sulfolobus solfataricus*. *Nucleic Acids Research* 42, 5280–5288. doi:10.1093/nar/gku161.

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

Zhao, C., Liu, F., and Pyle, A. M. (2018). An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA* 24, 183–195.

Zhuang, F., Karberg, M., Perutka, J., and Lambowitz, A. M. (2009). EcI5, a group IIB intron with high retrohoming frequency: DNA target site recognition and use in gene targeting. *RNA* 15, 432–49.

Zimmerly, S., Guo, H., Perlman, P. S., and Lambowitz, A. M. (1995). Group II intron mobility occurs by target DNA-primed reverse transcription. *Cell* 82, 545–54.

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

Zink, I. A., Pfeifer, K., Wimmer, E., Sleytr, U. B., Schuster, B., and Schleper, C. (2019). CRISPR-mediated gene silencing reveals involvement of the archaeal S-layer in cell division and virus infection. *Nat Commun* 10, 4797.

Appendix

Appendix A1. List of 280 RTs associated with CRISPR-Cas systems. RTs associated to CRISPR-Cas system section R1.3.1 are indicated in blue background. RT-CRISPR added from Silas *et al.*, 2017a analysis are indicated in pin RT-CRISPR added to the dataset in section R1.3.4 are indicated in white background.

RTCRISPR-Clade	CRISPR Loci	Accesion	Type of RT	Domain	Phylum	Class	Species/strain
1	III-A	fig|1434102.4.peg.2173	RT	Archaea	Euryarcheota	Methanomicrobia	*Methanosarcina sp. WH1*
1	III-A	fig|1434100.4.peg.3809	RT	Archaea	Euryarcheota	Methanomicrobia	*Methanosarcina sp. MTP4*
1	III-D	fig|1434110.4.peg.2496	RT	Archaea	Euryarcheota	Methanomicrobia	*Methanosarcina horonobensis HB-1 = JCM 15518*
1	III-C	fig|867904.9.peg.1159	RT	Archaea	Euryarcheota	Methanomicrobia	*Methanomethylovorans hollandica DSM 15978*
Silas_BR10	I-E	fig|66370.5.peg.118	RT	Bacteria	Actinobacteria	Actinobacteria	*Streptomyces flaveus strain NRRL ISP-5371*
Silas_BR10	I-E	WP_050506660.1	RT	Bacteria	Actinobacteria	Actinobacteria	*Streptomyces griseoflavus*
Silas_BR10	I-E	WP_046929267.1	RT	Bacteria	Actinobacteria	Actinobacteria	*Streptomyces lydicus*
Silas_BR10	I-E	WP_046085841.1	RT	Bacteria	Actinobacteria	Actinobacteria	*Streptomyces antioxidans*
Silas_BR10	I-E	fig|1156841.3.peg.4348	RT	Bacteria	Actinobacteria	Actinobacteria	*Streptomyces sp. ScaeMP-e10*
Silas_BR10	I-E	WP_003956872.1	RT	Bacteria	Actinobacteria	Actinobacteria	*Streptomyces clavuligerus ATCC 27064*
UN	Partial	WP_078692845.1	RT	Bacteria	Fusobacteria	Fusobacteriia	*Cetobacterium ceti*
UN	NA	PKN05312.1	RT,cas1	Bacteria	Proteobacteria	Deltaproteobacteria	*Deltaproteobacteria bacterium HGW-Deltaproted*
UN	NA	PJB64043.1	RT,cas1	Bacteria	Chloroflexi	Anaerolineae	*Anaerolineae bacterium CG_4_9_14_3_um_filter_*
5	III-B	fig|1174528.4.peg.770	RT	Bacteria	Cyanobacteria		*Fischerella sp. PCC 9339*
5	Partial III-B	fig|32057.3.peg.9081	RT	Bacteria	Cyanobacteria		*Calothrix sp. PCC 7103*
5	III-B	fig|128403.3.peg.8170	RT	Bacteria	Cyanobacteria		*Scytonema hofmanni PCC 7110*
5	Partial III-B	fig|1469607.3.peg.8359	RT	Bacteria	Cyanobacteria		*[Scytonema hofmanni] UTEX 2349*
5	Partial	fig|373994.3.peg.5259	RT	Bacteria	Cyanobacteria		*Rivularia sp. PCC 7116*
5	NA	fig|2014531.3.peg.5957	RT	Bacteria	Cyanobacteria		*Nostoc sp. 'Peltigera membranacea cyanobiont' 2*
5	Partial III-B	fig|1594576.4.peg.6931	RT	Bacteria	Cyanobacteria		*Mastigocladus laminosus UU774*
5	Partial III-B	fig|1729650.4.peg.4513	RT	Bacteria	Cyanobacteria		*Planktothrix sp. PCC 11201 strain BBR_PRJEB1099*
5	Partial III-B	fig|1173022.3.peg.668	RT	Bacteria	Cyanobacteria		*Crinalium epipsammum PCC 9333*
5	Partial	fig|102232.3.peg.1787	RT	Bacteria	Cyanobacteria		*Gloeocapsa sp. PCC 73106*
5	Partial	fig|65393.13.peg.5986	RT	Bacteria	Cyanobacteria		*Cyanothece sp. PCC 7424*
5	Partial	fig|1160286.3.peg.5620	RT	Bacteria	Cyanobacteria		*Microcystis aeruginosa PCC 9717*
5	NA	fig|721123.3.peg.3453	RT	Bacteria	Cyanobacteria		*Microcystis aeruginosa PCC 9701*
5	III-D	WP_075890713.1	RT,cas1	Bacteria	Cyanobacteria		*Limnothrix rosea*
5	III-D	fig|490193.3.peg.3185	RT,cas1	Bacteria	Cyanobacteria		*Synechococcus sp. NKBG042902*
5	III-B	WP_006515493.1	RT,cas1	Bacteria	Cyanobacteria		*Leptolyngbya sp. PCC 7375*
5	Partial	fig|1809277.3.peg.3395	RT,cas1	Bacteria	Cyanobacteria		*Pseudophormidium sp. E1*
5	III-B	WP_075600180.1	RT,cas1	Bacteria	Cyanobacteria		*Leptolyngbya sp. 'hensonii'*
5	III-B	WP_017302244.1	RT,cas1	Bacteria	Cyanobacteria		*Nodosilinea nodulosa*
5	Partial	WP_009625648.1	RT,cas1	Bacteria	Cyanobacteria		*Pseudanabaena biceps*
6	III-D	WP_099078969.1	RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Vibrio sp. PID17_43*
6	III-D	WP_011152750.1	RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Vibrio vulnificus YJ016*
6	III-D	fig|1004326.3.peg.1525	RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Vibrio sp. CAIM 1540*
6	III-D	WP_038884984.1	RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Vibrio rotiferianus*
6	NA	fig|1947768.3.peg.896	RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Vibrio sp. UBA2437*
6	NA	WP_046007427.1	RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Pseudoalteromonas rubra*
4	III-B	fig|1798424.3.peg.2009	RT	Bacteria	Ignavibacteriae	Ignavibacteria	*Ignavibacteria bacterium GWB2_35_6b*
4	Partial III-C	fig|1197129.4.peg.349	RT	Bacteria	Planctomycetes	Planctomycetia	*Candidatus Brocadia sinica JPN1*
4	III-A	CAJ74578.1	RT	Bacteria	Planctomycetes	Planctomycetia	*Candidatus Kuenenia stuttgartiensis*
4	III-B	fig|1775672.3.peg.443	RT	Bacteria	Proteobacteria	Deltaproteobacteria	*delta proteobacterium ML8_D strain ML8_D*

4	NA	fig|1973958.3.peg.664	RT	Bacteria	Proteobacteria	Deltaproteobacteria	*Deltaproteobacteria bacterium CG_4_10_14_0_8_*
4	III-A	fig|1284222.4.peg.3686	RT	Bacteria	Planctomycetes	Planctomycetia	*Candidatus Scalindua sp. husup-a2*
4	Partial	fig|671143.13.peg.836	RT	Bacteria	candidate division NC10		*Candidatus Methylomirabilis oxyfera strain Ru_en*
4	Partial	fig|880072.3.peg.1406	RT	Bacteria	Proteobacteria	Deltaproteobacteria	*Desulfobacca acetoxidans DSM 11109*
4	III-A	fig|274537.5.peg.1048	RT,cas1	Bacteria	Chlorobi	Chlorobia	*Chlorobaculum limnaeum strain DSM 1677*
4	III-A	WP_011745868.1	RT,cas1	Bacteria	Chlorobi	Chlorobia	*Chlorobium phaeobacteroides DSM 266*
4	III-A	WP_012509117.1	RT,cas1	Bacteria	Chlorobi	Chlorobia	*Pelodictyon phaeoclathratiforme BU-1*
4	NA	PIW70627.1	RT,cas1	Bacteria	Ignavibacteriae	Ignavibacteria	*Ignavibacteriales bacterium CG12_big_fil_rev_8_2*
4	NA	fig|1975524.3.peg.606	cas6,RT,cas1	Bacteria	Candidatus Marinimicrobia		*Candidatus Marinimicrobia bacterium CG08_land_*
4	NA	PIV67096.1	RT,cas1	Bacteria	Nitrospirae		*Nitrospirae bacterium CG01_land_8_20_14_3_00_*
4	III-B	PCI66716.1	RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Thiotrichaceae bacterium*
4	III-D	WP_027150711.1	RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Methylobacter tundripaludum*
4	Partial	PIE57287.1	RT,cas1	Bacteria	Proteobacteria	Deltaproteobacteria	*Desulfobulbus propionicus*
4	III-B	PPD32300.1	RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Methylomonas sp.*
4	Partial	fig|917.4.peg.179	RT,cas1	Bacteria	Proteobacteria	Betaproteobacteria	*Nitrosomonas marina strain Nm71*
4	Partial III-B	fig|33059.15.peg.2636	RT,cas1	Bacteria	Proteobacteria	Acidithiobacillia	*Acidithiobacillus caldus strain DX*
4	Partial	fig|1798306.3.peg.2661	RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Gammaproteobacteria bacterium RIFOXYA12_FUL*
4	III-D	ETX03376.1	RT,cas1	Bacteria	Candidatus Tectomicrobia		*Candidatus Entotheonella gemina*
7	III-D	WP_075783338.1	RT,cas1	Bacteria	Proteobacteria	Alphaproteobacteria	*Rhodovulum sulfidophilum*
7	III-D	OIQ35231.1	RT,cas1	Bacteria	Proteobacteria	Alphaproteobacteria	*Roseobacter sp. MedPE-SWchi*
7	partial	PIE06493.1	RT,cas1	Bacteria	Proteobacteria	Alphaproteobacteria	*Rhodobacterales bacterium*
7	VI-A	WP_080615428.1	RT,cas1	Bacteria	Proteobacteria	Alphaproteobacteria	*Rhodovulum sp. MB263*
7	III-D	BAQ71286.1	RT,cas1	Bacteria	Proteobacteria	Alphaproteobacteria	*Rhodovulum sulfidophilum*
7	III-B	WP_014188713.1	RT,cas1	Bacteria	Proteobacteria	Alphaproteobacteria	*Azospirillum lipoferum*
7	III-D	WP_103010808.1	RT,cas1	Bacteria	Proteobacteria	Alphaproteobacteria	*Rhodopseudomonas palustris*
7	III-D	WP_008391842.1	RT,cas1	Bacteria	Proteobacteria	Alphaproteobacteria	*Rhodovulum sp. PH10*
7	III-D	fig|2003584.3.peg.1215	RT,cas1	Bacteria	Proteobacteria	Alphaproteobacteria	*Stappia sp. TSB10GB4*
7	III-D	WP_014747450.1	RT,cas1	Bacteria	Proteobacteria	Alphaproteobacteria	*Tistrella mobilis*
7	NA	WP_062763150.1	cas1*	Bacteria	Proteobacteria	Alphaproteobacteria	*Tistrella mobilis*
7	NA	EJW09481.1	RT,cas1	Bacteria	Proteobacteria	Alphaproteobacteria	*Rhodovulum sp. PH10*
7	III-B	WP_019956891.1	RT,cas1	Bacteria	Proteobacteria	Alphaproteobacteria	*Yoonia vestfoldensis*
7	III-D	WP_012973664.1	RT,cas1	Bacteria	Proteobacteria	Alphaproteobacteria	*Azospirillum lipoferum*
7	III-D	WP_013258512.1	RT,cas1	Bacteria	Proteobacteria	Alphaproteobacteria	*Desulfarculus baarsii*
7	III-B	WP_019960649.1	RT,cas1	Bacteria	Proteobacteria	Alphaproteobacteria	*Woodsholea maritima*
7	III-D	KQB14189.1	RT,cas1	Bacteria	Proteobacteria	Alphaproteobacteria	*Rhodobacter capsulatus*
7	III-D	WP_092621548.1	RT,cas1	Bacteria	Proteobacteria	Alphaproteobacteria	*Roseospirillum parvum*
7	III-D	WP_096173198.1	RT,cas1	Bacteria	Proteobacteria	Alphaproteobacteria	*Cohaesibacter sp. ES.047*
7	III-D	SNR44634.1	RT	Bacteria	Proteobacteria	Alphaproteobacteria	*Puniceibacterium sediminis*
7	partial	fig|1121381.3.peg.338	RT	Bacteria	Deinococcus-Thermus	Deinococci	*Deinococcus marmoris DSM 12784*
7	partial	fig|1510458.3.peg.3133	RT	Bacteria	Proteobacteria	Alphaproteobacteria	*Pseudoruegeria sabulilitoris strain GJMS-35*
7	III-D	fig|648757.4.peg.1971	RT	Bacteria	Proteobacteria	Alphaproteobacteria	*Rhodomicrobium vannielii ATCC 17100*
7	NA	fig|905052.3.peg.1368	RT,cas1	Bacteria	Proteobacteria	Betaproteobacteria	*Thauera selenatis AX*
7	III-D	OZB62245.1	RT,cas1	Bacteria	Proteobacteria	Betaproteobacteria	*Thiomonas sp. 13-66-29*
7	III-D	EGJ09042.1	RT,cas1	Bacteria	Proteobacteria	Betaproteobacteria	*Rubrivivax benzoatilyticus*

7	III-B	WP_089416921.1	RT,cas1	Bacteria	Proteobacteria	Betaproteobacteria	*Vitreoscilla filiformis*
7	partial	fig|1797572.3.peg.345	RT,cas1	Bacteria	Proteobacteria	Betaproteobacteria	*Burkholderiales bacterium RIFOXYC12_FULL_65_23*
7	III-B	KFB76584.1	RT,cas1	Bacteria	Proteobacteria	Betaproteobacteria	*Candidatus Accumulibacter sp. SK-02*
7	partial	fig|1047063.3.peg.107	RT	Bacteria			*candidate division WS1 bacterium JGI 0000059-K21*
10	Partial	WP_079139548.1	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Streptomyces sp. AVP053U2*
10	Partial	WP_086793258.1	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Streptomyces thermovulgaris*
10	III-B	WP_084012720.1	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Thermobifida halotolerans*
10	III-B	fig|1236902.3.peg.4955	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Nocardiopsis baichengensis*
10	Partial III-A	fig|366584.3.peg.4449	RT	Bacteria	Actinobacteria	Actinobacteria	*Pseudonocardia oroxyli strain CGMCC 4.3143*
10	Partial	WP_052914180.1	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Frankia*
10	III-D	WP_092521862.1	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Actinopolyspora saharensis*
10	III-B	WP_083978991.1	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Micromonospora rosaria*
10	Partial III-D	KGM14440.1	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Cellulomonas bogoriensis*
10	III-D	WP_099698861.1	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Rhodococcus enclensis*
10	III-D	EKX89922.1	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Corynebacterium durum*
10	Partial	WP_073191102.1	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Tessaracoccus bendigoensis*
10	III-D	WP_084635283.1	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Propionicicella superfundia*
10	III-B	WP_052396493.1	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Kutzneria sp. 744*
10	III-D	BAK34153.1	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Microlunatus phosphovorus*
10	Partial	PIE27059.1	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Micrococcales bacterium*
10	III-D	fig|1223536.4.peg.39	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Skermania piniformis NBRC 15059*
10	III-D	WP_070740732.1	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Corynebacterium sp. HMSC073D01*
10	III-D	WP_103064264.1	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Actinomyces sp. 553*
10	III-D	WP_053587381.1	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Actinomyces sp. oral taxon 414*
10	III-D	WP_081379745.1	RT	Bacteria	Actinobacteria	Actinobacteria	*Actinomyces naeslundii*
10	III-D	WP_083663296.1	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Actinomyces mediterranea*
10	III-D	WP_005962375.1	RT,cas1	Bacteria	Actinobacteria	Actinobacteria	*Actinomyces cardiffensis F0333*
13	III-B	PID64632.1	RT	Bacteria	Proteobacteria	Gammaproteobacteria	*Gammaproteobacteria bacterium*
13	Partial	fig|765912.4.peg.325	RT	Bacteria	Proteobacteria	Gammaproteobacteria	*Thioflavicoccus mobilis 8321*
13	III-B	WP_028490526.1	RT	Bacteria	Proteobacteria	Gammaproteobacteria	*Thiothrix lacustris*
13	III-B	fig|1704499.3.peg.2235	RT	Bacteria	Proteobacteria	Gammaproteobacteria	*Methylovulum psychrotolerans strain HV10_M2*
13	III-B	WP_082674220.1	RT	Bacteria	Proteobacteria	Gammaproteobacteria	*Thiohalocapsa sp. ML1*
13	NA	ESQ15779.1	RT	Bacteria	Proteobacteria	Gammaproteobacteria	*uncultured Thiohalocapsa sp. PB-PSB1*
13	Partial	fig|1662188.4.peg.2603	RT	Bacteria			*Candidatus Achromatium palustre*
13	NA	fig|2026735.21.peg.187	RT	Bacteria	Proteobacteria	Deltaproteobacteria	*Deltaproteobacteria bacterium strain ER2bin7*
13	Partial	KPA15875.1	RT	Bacteria	Proteobacteria	Deltaproteobacteria	*Candidatus Magnetomorum sp. HK-1*
13	Partial	fig|1121400.3.peg.3691	RT	Bacteria	Proteobacteria	Deltaproteobacteria	*Desulfobacterium vacuolatum DSM 3385 strain DS*
13	III-D	WP_035075942.1	RT	Bacteria	Proteobacteria	Deltaproteobacteria	*Desulfovibrio zosterae*
13	III-D	WP_018158292.1	RT	Bacteria	Proteobacteria	Deltaproteobacteria	*Thioalkalivibrio sp. ALE14*
13	III-D	WP_018870803.1	RT	Bacteria	Proteobacteria	Gammaproteobacteria	*Thioalkalivibrio sp. ALgr3*
13	Partial	fig|160660.9.peg.1539	RT	Bacteria	Proteobacteria	Gammaproteobacteria	*Acidihalobacter prosperus strain V6*
13	NA	WP_044413451.1	RT	Bacteria	Proteobacteria	Gammaproteobacteria	*Thiomicrospira microaerophila*
13	III-D	fig|1977087.18.peg.3009	RT	Bacteria	Proteobacteria		*Proteobacteria bacterium strain DOLZORAL124_50*
13	NA	WP_006787899.1	RT	Bacteria	Proteobacteria	Gammaproteobacteria	*Thiorhodospira sibirica ATCC 700588*

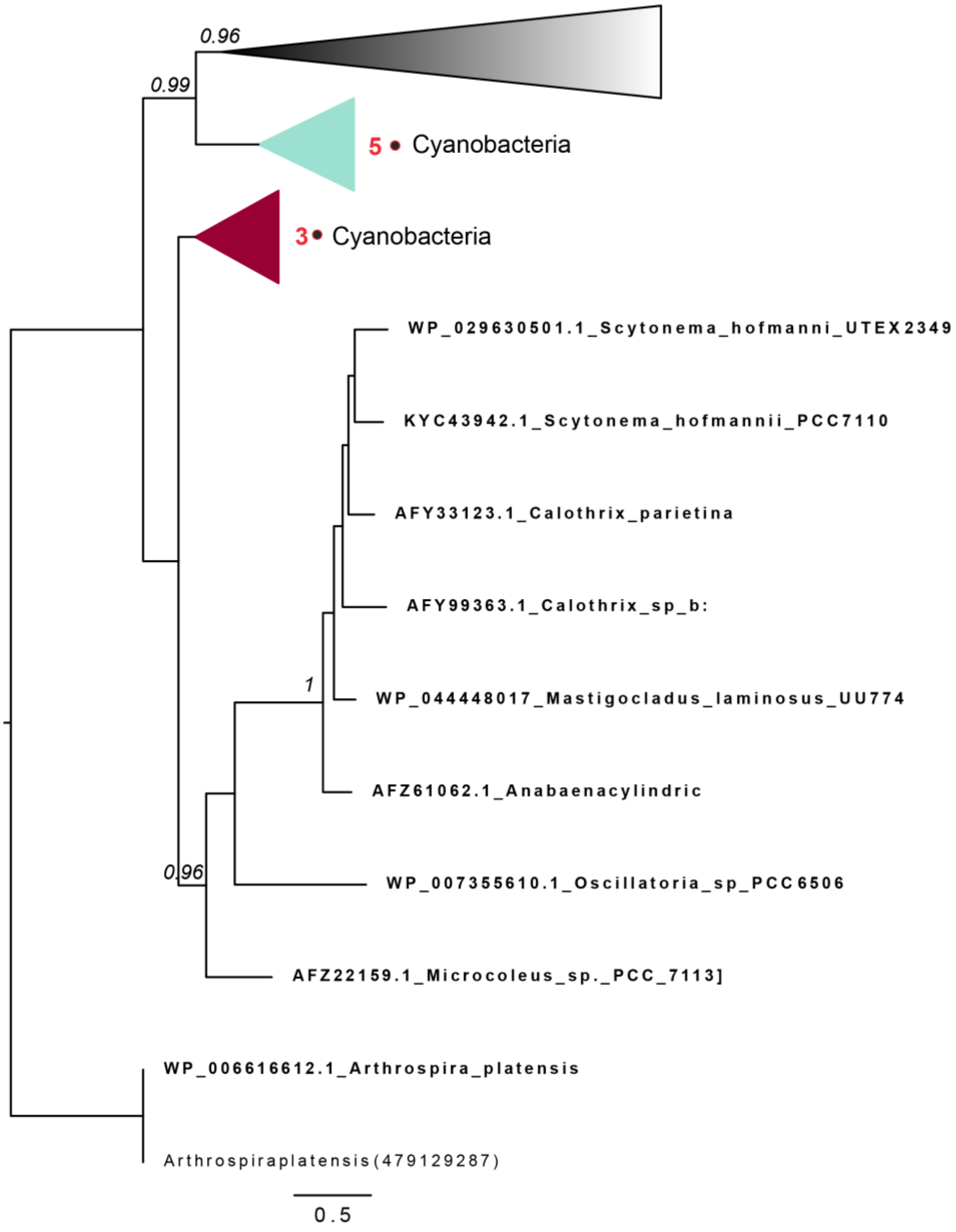
13	NA	SFQ11412.1	RT	Bacteria	Proteobacteria	Gammaproteobacteria	*Nitrosomonas cryotolerans*
13	Partial III-D	WP_007039880.1	RT	Bacteria	Proteobacteria	Gammaproteobacteria	*Thiorhodococcus drewsii AZ1*
13	III-C	fig|1953093.3.peg.367	RT	Bacteria	Candidatus Bipolaricaulota		*Acetothermia bacterium UBA3560*
13	Partial	fig|1950206.3.peg.4687	RT	Bacteria	Chloroflexi		*Anaerolineales bacterium UBA3905*
13	Partial	fig|2026724.112.peg.165	RT	Bacteria	Chloroflexi		*Chloroflexi bacterium strain JP1_8*
9	NA	fig|1104566.3.peg.2332	RT	Bacteria	Candidatus Poribacteria		*Poribacteria bacterium WGA-3G*
9	Partial	WP_011957721.1	RT	Bacteria	Chloroflexi	Chloroflexia	*Roseiflexus sp. RS-1*
9	Partial	WP_012121172.1	RT	Bacteria	Chloroflexi	Chloroflexia	*Roseiflexus castenholzii DSM 13941*
9	Partial	fig|1707952.4.peg.3320	RT	Bacteria	Chloroflexi	Chloroflexia	*Chloroflexus sp. isl-2*
9	Partial	WP_012259222.1	RT	Bacteria	Chloroflexi	Chloroflexia	*Chloroflexus aurantiacus J-10-fl*
9	Partial	ABX05025.1	RT	Bacteria	Chloroflexi	Chloroflexia	*Herpetosiphon aurantiacus ATCC 23779*
9	NA	fig|1935099.3.peg.2126	RT	Bacteria	Chloroflexi	Anaerolineae	*Anaerolineales bacterium JdFR-64 strain JdFR-64*
9	NA	fig|1973907.3.peg.502	RT,cas1	Bacteria	Chloroflexi	Anaerolineae	*Anaerolineae bacterium CG17_big_fil_post_rev_8_*
13	III-D	fig|1801687.3.peg.3424	RT,primpol	Bacteria	Nitrospinae/Tectomicr	Nitrospinae	*Nitrospinae bacterium RIFCSPLOWO2_12_FULL_45_*
13	NA	fig|1805005.3.peg.1300	RT	Bacteria	Chloroflexi	Anaerolineae	*Anaerolineae bacterium CG2_30_64_16*
13	Partial	fig|1986204.3.peg.1051	RT	Bacteria	Chloroflexi	Anaerolineae	*Anaerolineaceae bacterium CAMBI-1 strain CAMBI-*
13	Partial	fig|1134406.4.peg.2720	RT	Bacteria	Chloroflexi	Anaerolineae	*Ornatilinea apprima strain P3M-1*
8A	III-D	ESQ08042.1	cas6,RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*uncultured Thiohalocapsa sp. PB-PSB1*
8A	III-D	WP_093186185.1	cas6,RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Thiocapsa sp. KS1*
8A	III-B	WP_077732696.1	cas6,RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Methylocaldum sp. 14B*
8A	NA	PIE37212.1	cas6,RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Gammaproteobacteria bacterium*
8A	NA	ESQ17084.1	cas6,RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*uncultured Thiohalocapsa sp. PB-PSB1*
8A	NA	PID60451.1	cas6,RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Gammaproteobacteria bacterium*
8A	III-D	WP_019606016.1	cas6,RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Teredinibacter turnerae*
8A	III-B	PIQ31291.1	cas6,RT,cas1	Bacteria	Proteobacteria	Zetaproteobacteria	*Zetaproteobacteria bacterium CG17_big_fil_post_re*
8A	III-D	fig|1974074.3.peg.1401	cas6,RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Piscirickettsiaceae bacterium CG_4_10_14_0_8_um_*
8A	III-D	WP_038137810.1	cas6,RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Hydrogenovibrio sp. Milos-T1*
8A	III-B	WP_015817555.1	cas6,RT	Bacteria	Proteobacteria	Gammaproteobacteria	*Teredinibacter turnerae*
8A	III-B	fig|687.13.peg.3591	cas6,RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Vibrio gazogenes strain type strain: CECT 5068*
8A	III-D	WP_038188758.1	cas6,RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Vibrio sinaloensis*
8A	III-D	WP_055043549.1	cas6,RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Vibrio metoecus*
8A	III-D	WP_047875592.1	cas6,RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Photobacterium aphoticum*
8A	III-D	WP_072955141.1	cas6,RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Vibrio gazogenes DSM 21264*
8A	III-D	fig|1913989.202.peg.530	cas6,RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Gammaproteobacteria bacterium strain DOLJORAL7*
8A	III-B	WP_013659858.1	cas6,RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Marinomonas mediterranea*
8A	III-B	WP_028883449.1	cas6,RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Teredinibacter turnerae*
8A	III-D	WP_028302067.1	cas6,RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Oceanospirillum beijerinckii*
8A	III-D	KUI97421.1	cas6,RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Vibrio sp. MEBiC08052*
8A	Partial III-A/D	fig|43662.8.peg.1008	cas6,RT	Bacteria	Proteobacteria	Gammaproteobacteria	*Pseudoalteromonas piscicida strain S2040*
8A	III-D	WP_007469744.1	cas6,RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Photobacterium marinum*
8B	NA	KKO19838.1	RT,cas1	Bacteria	Planctomycetes	Planctomycetia	*Candidatus Brocadia fulgida*
8B	Partial III-B	WP_052565451.1	RT,cas1	Bacteria	Planctomycetes	Planctomycetia	*Candidatus Brocadia sinica*
8B	NA	KFZ44108.1	RT,cas1	Bacteria	Proteobacteria	Deltaproteobacteria	*Smithella sp. D17*
8B	Partial III-D	KFB71594.1	cas6,RT,cas1	Bacteria	Proteobacteria	Betaproteobacteria	*Candidatus Accumulibacter sp. BA-91*

8B	III-D	fig|1871110.4.peg.1170	RT,cas1	Bacteria	Nitrospirae	Nitrospira	*Thermodesulfovibrio sp. N1*
8B	III-D	fig|1948271.3.peg.4471	RT,cas1	Bacteria	Verrucomicrobia		*Verrucomicrobia bacterium UBA6176*
8B	III-B	WP_019672870.1	RT,cas1	Bacteria	Proteobacteria	Gammaproteobacteria	*Psychrobacter lutiphocae*
8B	Partial	WP_085101365.1	cas6,RT,cas1	Bacteria	Proteobacteria	Deltaproteobacteria	*Desulfovibrio sp. K3S*
8B	III-B	WP_015334627.1	cas6,RT,cas1	Bacteria	Proteobacteria	Deltaproteobacteria	*Desulfovibrio hydrothermalis*
8B	III-B	WP_027180402.1	cas6,RT,cas1	Bacteria	Proteobacteria	Deltaproteobacteria	*Desulfovibrio bastinii*
8B	III-B	fig|1961408.3.peg.659	cas6,RT,cas1	Bacteria	Proteobacteria	Deltaproteobacteria	*Desulfovibrio sp. UBA6079*
8B	NA	OQX07113.1	cas6,RT,cas1	Bacteria	Proteobacteria	Deltaproteobacteria	*Desulfobulbaceae bacterium A2*
8	III-B	fig|1947377.3.peg.1556	RT,cas1	Bacteria	Proteobacteria	Betaproteobacteria	*Rhodoferax sp. UBA4127*
8	III-A/D	KJR40057.1	cas6,RT,cas1	Bacteria	Nitrospirae	Nitrospira	*Candidatus Magnetoovum chiemensis*
8	III-A/D	fig|1961537.3.peg.1464	cas6,RT,cas1	Bacteria	Proteobacteria	Deltaproteobacteria	*Desulfobacteraceae bacterium UBA5605*
8	III-B	fig|1947870.3.peg.515	cas6,RT	Bacteria	Candidatus Cloacimonetes		*Cloacimonetes bacterium UBA6081*
8	III-D	WP_012910084.1	cas6,RT,cas1	Bacteria	Planctomycetes	Planctomycetia	*Pirellula staleyi*
3	III-D	WP_010995638.1	RT,cas1	Bacteria	Cyanobacteria		*Nostocaceae*
3	III-B	WP_015196021.1	RT,cas1	Bacteria	Cyanobacteria		*Calothrix parietina*
3	III-D	WP_029630506.1	RT,cas1	Bacteria	Cyanobacteria		*[Scytonema hofmanni] UTEX 2349*
3	III-D	WP_033334699.1	RT,cas1	Bacteria	Cyanobacteria		*Scytonema hofmannii*
3	III-B	WP_044448019.1	RT,cas1	Bacteria	Cyanobacteria		*Mastigocladus laminosus*
3	III-B	AFZ61061.1	RT,cas1	Bacteria	Cyanobacteria		*Anabaena cylindrica*
3	III-B	WP_073595699.1	RT,cas1	Bacteria	Cyanobacteria		*Phormidium ambiguum*
3	Partial	fig|671072.4.peg.5865	RT,cas1	Bacteria	Cyanobacteria		*Planktothrix tepida PCC 9214*
3	III-B	WP_007355619.1	RT,cas1	Bacteria	Cyanobacteria		*Kamptonema*
3	III-D	WP_015186217.1	RT,cas1	Bacteria	Cyanobacteria		*Microcoleus sp. PCC 7113*
3	III-B	WP_087711850.1	RT,cas1	Bacteria	Cyanobacteria		*Phormidium sp. HE10JO*
3	Partial	WP_088428978.1	RT,cas1	Bacteria	Cyanobacteria		*Halomicronema hongdechloris*
3	Partial	WP_008312855.1	RT,cas1	Bacteria	Cyanobacteria		*Leptolyngbya sp. PCC 6406*
3	III-D	KPQ33062.1	RT,cas1	Bacteria	Cyanobacteria		*Phormidesmis priestleyi Ana*
3	NA	WP_024971209.1	RT,cas1	Bacteria	Cyanobacteria		*Microcystis aeruginosa*
3	III-B	WP_014275551.1	RT,cas1	Bacteria	Cyanobacteria		*Arthrospira platensis*
3	III-B	WP_013334746.1	RT,cas1	Bacteria	Cyanobacteria		*Cyanothece sp. PCC 7822*
3	III-B	WP_072619174.1	RT,cas1	Bacteria	Cyanobacteria		*Spirulina major*
UN	III-D	fig|561720.4.peg.2507	RT	Bacteria	Synergistetes	Synergistia	*Dethiosulfovibrio salsuginis strain USBA 82*
UN	III-B	CBL28738.1	RT	Bacteria	Synergistetes	Synergistia	*Fretibacterium fastidiosum*
UN	Partial	fig|1908690.5.peg.5617	RT	Bacteria	Planctomycetes	Planctomycetia	*Fimbriiglobus ruber strain SP5*
2B	III-B	fig|1950581.3.peg.79	RT,cas1	Bacteria	Bacteroidetes	Bacteroidia	*Bacteroidales bacterium UBA4639*
2B	Partial III-D	WP_081798861.1	RT	Bacteria	Bacteroidetes	Bacteroidia	*Bacteroides barnesiae*
2B	III-B	fig|1841857.3.peg.2215	RT,cas1	Bacteria	Bacteroidetes	Bacteroidia	*Odoribacter sp. Marseille-P2698*
2B	III-D	WP_007481073.1	RT,cas1	Bacteria	Bacteroidetes	Bacteroidia	*Bacteroides salyersiae*
2B	III-B	WP_032556864.1	RT,cas1	Bacteria	Bacteroidetes	Bacteroidia	*Bacteroides fragilis*
2B	III-B	fig|1952870.3.peg.3771	RT,cas1	Bacteria	Bacteroidetes	Saprospiria	*Saprospiraceae bacterium UBA2365*
2B	III-D	OJX90108.1	RT,cas1	Bacteria	Bacteroidetes	Bacteroidia	*Paludibacter sp. 47-17*
2B	NA	KHE91657.1	RT,cas1	Bacteria	Bacteroidetes	Bacteroidia	*Candidatus Scalindua brodae*
2B	Partial III-A/D	WP_007220853.1	RT,cas1	Bacteria	Planctomycetes	Planctomycetia	*Candidatus Jettenia caeni*
2B	III-B	PIE72267.1	RT,cas1	Bacteria	Proteobacteria	Deltaproteobacteria	*Deltaproteobacteria bacterium*

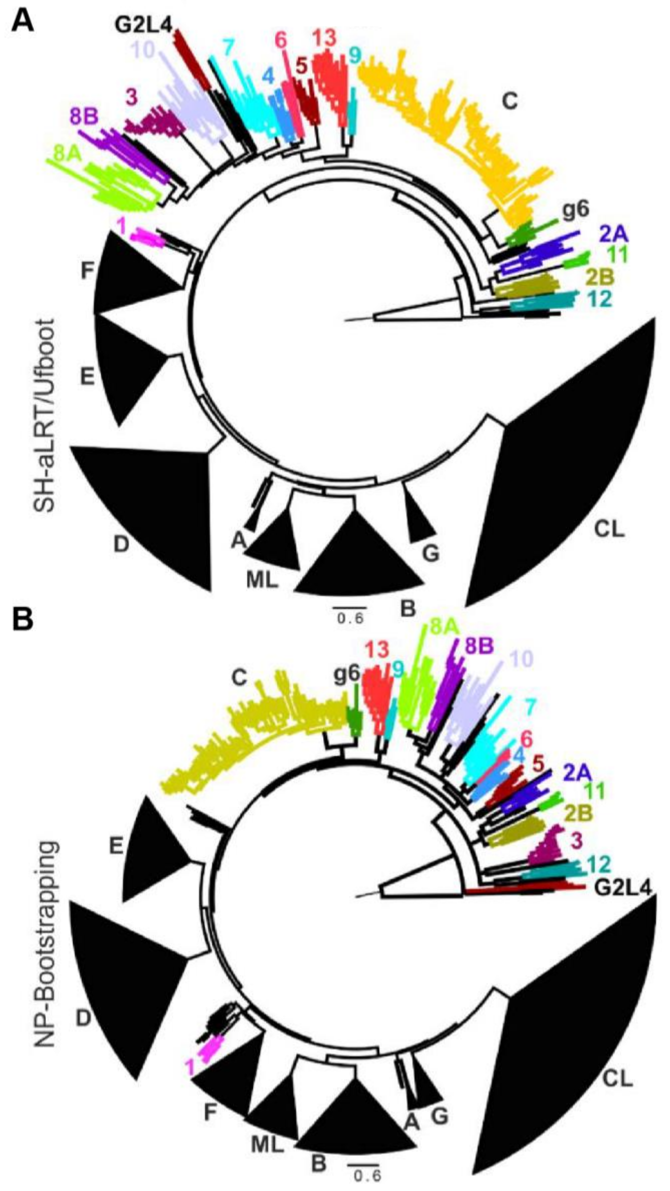
2B	NA	fig|1940691.3.peg.1221	RT,cas1	Bacteria	Proteobacteria	Deltaproteobacteria	*Desulfobacca sp. 4484_104 strain 4484_104*
2B	III-D	WP_013707702.1	RT,cas1	Bacteria	Proteobacteria	Deltaproteobacteria	*Desulfobacca acetoxidans*
2B	NA	fig|1961106.3.peg.1959	RT	Bacteria	Ignavibacteriae	Ignavibacteria	*Ignavibacteriales bacterium UTCHB3 strain UTC*
2B	III-B	fig|1950646.3.peg.3604	RT,cas1	Bacteria	Bacteroidetes	Bacteroidia	*Bacteroidales bacterium UBA5537*
2B	NA	KKO17867.1	RT,cas1	Bacteria	Planctomycetes	Planctomycetia	*Candidatus Brocadia fulgida*
2B	III-A	PID26398.1	RT,cas1	Bacteria	Candidatus Cloacimonetes		*Candidatus Cloacimonetes bacterium*
2B	NA	fig|1956175.3.peg.1464	RT,cas1	Bacteria			*candidate division KSB1 bacterium 4484_188 st*
11	III-B	WP_039443024.1	cas6,RT,cas1	Bacteria	Bacteroidetes	Bacteroidia	*Porphyromonas gulae*
11	III-B	WP_013815267.1	cas6,RT,cas1	Bacteria	Bacteroidetes	Bacteroidia	*Porphyromonas gingivalis*
11	NA	WP_036885018.1	cas6,RT,cas1	Bacteria	Bacteroidetes	Bacteroidia	*Porphyromonas gingivicanis*
11	Partial	PID94761.1	cas6,RT,cas1	Bacteria	Bacteroidetes		*Bacteroidetes bacterium*
2/11	NA	KPA10619.1	RT,cas1	Bacteria	Proteobacteria	Deltaproteobacteria	*Candidatus Magnetomorum sp. HK-1*
2/11	NA	ETR69258.1	RT,cas1	Bacteria	Proteobacteria	Deltaproteobacteria	*Candidatus Magnetoglobus multicellularis str.*
2A	NA	fig|1802251.3.peg.1081	RT,cas1	Bacteria	Proteobacteria	Epsilonproteobacteria [b]	*Sulfurimonas sp. RIFCSPHIGHO2_12_FULL_36_9*
2A	NA	fig|1802262.3.peg.175	RT,cas1	Bacteria	Proteobacteria	Epsilonproteobacteria [b]	*Sulfurimonas sp. RIFOXYD2_FULL_37_8*
2A	NA	fig|1947599.3.peg.823	RT,cas1	Bacteria	Proteobacteria	Epsilonproteobacteria [b]	*Sulfurospirillum sp. UBA6791*
2A	III-B	WP_046996094.1	RT,cas1	Bacteria	Proteobacteria	Epsilonproteobacteria [b]	*Arcobacter butzleri*
2A	III-D	fig|1802256.3.peg.1041	RT,cas1	Bacteria	Proteobacteria	Epsilonproteobacteria [b]	*Sulfurimonas sp. RIFOXYB12_FULL_35_9*
2A	III-D	fig|1032240.3.peg.1266	RT,cas1	Bacteria	Proteobacteria	Epsilonproteobacteria [b]	*Arcobacter thereius LMG 24486*
2A	III-A	WP_087578415.1	RT,cas1	Bacteria	Proteobacteria	Epsilonproteobacteria [b]	*Campylobacter concisus*
2A	III-D	WP_021087740.1	RT,cas1	Bacteria	Proteobacteria	Epsilonproteobacteria [b]	*Campylobacter concisus*
2A	Partial	WP_005873073.1	RT,cas1	Bacteria	Proteobacteria	Epsilonproteobacteria [b]	*Campylobacter gracilis*
2A	III-D	WP_075539949.1	RT,cas1	Bacteria	Proteobacteria	Epsilonproteobacteria [b]	*Campylobacter geochelonis*
2A	Partial III-D	WP_087700548.1	RT,cas1	Bacteria	Proteobacteria	Epsilonproteobacteria [b]	*Campylobacter jejuni*
2A	III-A	WP_007475276.1	RT	Bacteria	Proteobacteria	Epsilonproteobacteria [b]	*Caminibacter mediatlanticus TB-2*
2A	III-B	WP_025270209.1	RT,cas1	Bacteria	Proteobacteria	Deltaproteobacteria [b]	*Hippea sp. KM1*
2A	III-A	WP_084275429.1	RT,cas1	Bacteria	Proteobacteria	Epsilonproteobacteria [b]	*Nitratiruptor tergarcus*
12	NA	WP_082424645.1	RT,cas1	Bacteria	Firmicutes	Clostridia	*Roseburia inulinivorans*
12	III-D	WP_015568484.1	RT	Bacteria	Firmicutes	Clostridia	*[Eubacterium] rectale*
12	III-D	OLA06495.1	RT,cas1	Bacteria	Firmicutes	Clostridia	*Eubacterium sp. 38_16*
12	VI-A	WP_090127495.1	RT,cas1	Bacteria	Firmicutes	Clostridia	*Eubacteriaceae bacterium CHKCI004*
12	III-D	WP_008751399.1	RT,cas1	Bacteria	Firmicutes	Clostridia	*Lachnoanaerobaculum saburreum*
12	III-D	WP_060932241.1	RT,cas1	Bacteria	Firmicutes	Clostridia	*Lachnoanaerobaculum saburreum*
12	III-D	WP_090022708.1	RT,cas1	Bacteria	Firmicutes	Clostridia	*Lachnospiraceae bacterium C10*
12	III-B	WP_051592781.1	RT,cas1	Bacteria	Firmicutes	Erysipelotrichia	*[Clostridium] saccharogumia DSM 17460*
12	III-D	WP_092966766.1	RT,cas1	Bacteria	Firmicutes	Erysipelotrichia	*Ruminococcaceae bacterium P7*
12	NA	fig|1951970.3.peg.1857	RT,primpol	Bacteria	Firmicutes	Clostridia	*Lachnospiraceae bacterium UBA1066*
12	Partial	fig|1952090.3.peg.547	RT,primpol	Bacteria	Firmicutes	Clostridia	*Lachnospiraceae bacterium UBA4364*
12	III-D	CDE54369.1	RT	Bacteria	Firmicutes	Clostridia	*Roseburia sp. CAG:303*
12	III-D	WP_042734856.1	RT	Bacteria	Firmicutes	Clostridia	*Lachnospiraceae bacterium TWA4*
12	III-D	fig|169283.5.peg.3238	RT,primpol	Bacteria	Firmicutes	Bacilli	*Geobacillus lituanicus strain N-3*
12	III-D	fig|404937.3.peg.1451	RT,primpol	Bacteria	Firmicutes	Bacilli	*Anoxybacillus thermarum strain AF/04*
12	III-D	fig|1904753.3.peg.2820	RT,primpol	Bacteria	Firmicutes	Bacilli	*Bacillus sp. SA5d-4*
12	III-B	ETR73831.1	RT	Bacteria	Proteobacteria	Deltaproteobacteria	*Candidatus Magnetoglobus multicellularis str.*

12	Partial	WP_069875436.1	RT	Bacteria	Firmicutes	Clostridia	*Fusibacter sp. 3D3*
12	NA	ETR68993.1	RT	Bacteria	Proteobacteria	Deltaproteobacteria	*Candidatus Magnetoglobus multicellularis str*
12	NA	fig|1948077.3.peg.1392	RT,primpol	Bacteria	Firmicutes		*Firmicutes bacterium UBA5625*
12	NA	fig|1953133.3.peg.161	RT	Bacteria	Armatimonadetes		*Armatimonadetes bacterium UBA5377*
14	III-B	SFS79440.1	RT	Bacteria	Bacteroidetes	Flavobacteriia	*Lutibacter maritimus*
14	Partial III-D	fig|1622118.3.peg.525	RT	Bacteria	Bacteroidetes	Flavobacteriia	*Lutibacter sp. LP1*
14	III-B	fig|1947390.3.peg.2730	RT	Bacteria	Bacteroidetes	Cytophagia	*Roseivirga sp. UBA6061*
14	III-B	WP_092178911.1	RT	Bacteria	Bacteroidetes	Cytophagia	*Cyclobacterium halophilum*
14	III-B	fig|478744.3.peg.2482	RT	Bacteria	Bacteroidetes	Saprospiria	*Lewinella agarilytica strain DSM 24740*
14	Partial	fig|760192.3.peg.6439	RT	Bacteria	Bacteroidetes	Saprospiria	*Haliscomenobacter hydrossis DSM 1100*
14	Partial	fig|1524460.3.peg.4312	RT	Bacteria	Bacteroidetes	Saprospiria	*Phaeodactylibacter xiamenensis KD52*
14	III-D	WP_096194726.1	RT	Bacteria	Bacteroidetes	Cytophagia	*Cytophagales bacterium TFI 002*
14	III-B	WP_013769055.1	RT	Bacteria	Bacteroidetes	Saprospiria	*Haliscomenobacter hydrossis*
14	III-B	fig|1952866.3.peg.4727	RT	Bacteria	Bacteroidetes	Saprospiria	*Saprospiraceae bacterium UBA2329*
UN	Partial	fig|857293.11.peg.1034	RT,primpol	Bacteria	Firmicutes	Clostridia	*Caloramator australicus RC3*
15	I-C	WP_1000052051.1	RT	Bacteria	Proteobacteria	Gammaproteobacteria	*Basfia succiniciproducens*
15	I-C	WP_005646965.1	RT	Bacteria	Proteobacteria	Gammaproteobacteria	*Haemophilus haemolyticus HK386*

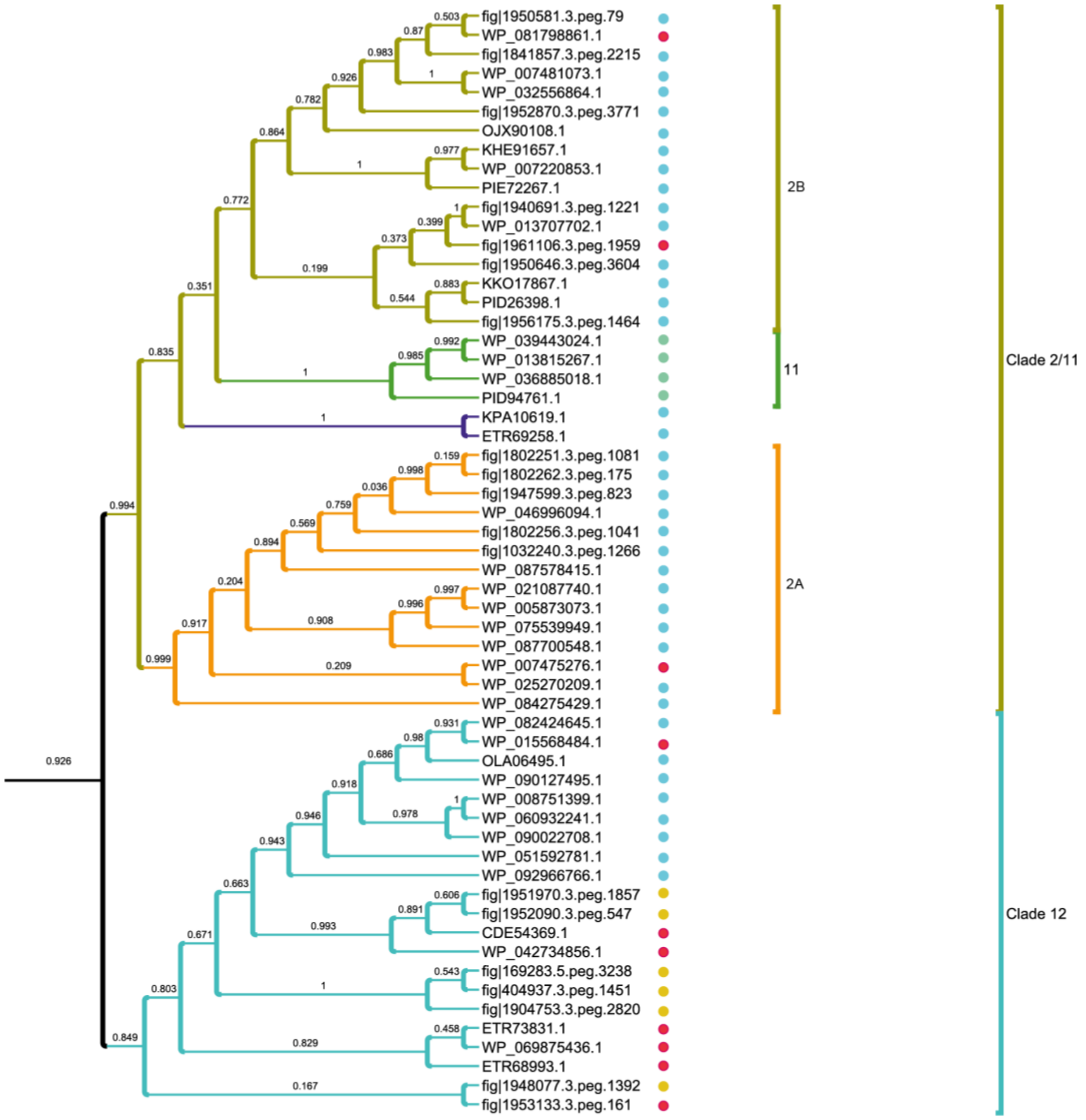
Appendix A2. Cas1 phylogenetic tree inferred by addition of 9 Cas1 sequences present separately from the RTCas1 fusion in the CRISPR-Cas modules from the clade 3 to the alignment used to infer the Cas1 phylogenetic tree shown in Figure R1.5. FastTree support value ≥ 0.92 are indicated at the nodes. For the sake of simplicity, Cas1 clades were collapsed.


0.96
0.99
5 ● Cyanobacteria
3● Cyanobacteria
WP_029630501.1_Scytonema_hofmanni_UTEX2349
KYC43942.1_Scytonema_hofmannii_PCC7110
AFY33123.1_Calothrix_parietina
AFY99363.1_Calothrix_sp_b:
1
WP_044448017_Mastigocladus_laminosus_UU774
AFZ61062.1_Anabaenacylindric
0.96
WP_007355610.1_Oscillatoria_sp_PCC6506
AFZ22159.1_Microcoleus_sp._PCC_7113]
WP_006616612.1_Arthrospira_platensis
Arthrospiraplatensis(479129287)
0.5

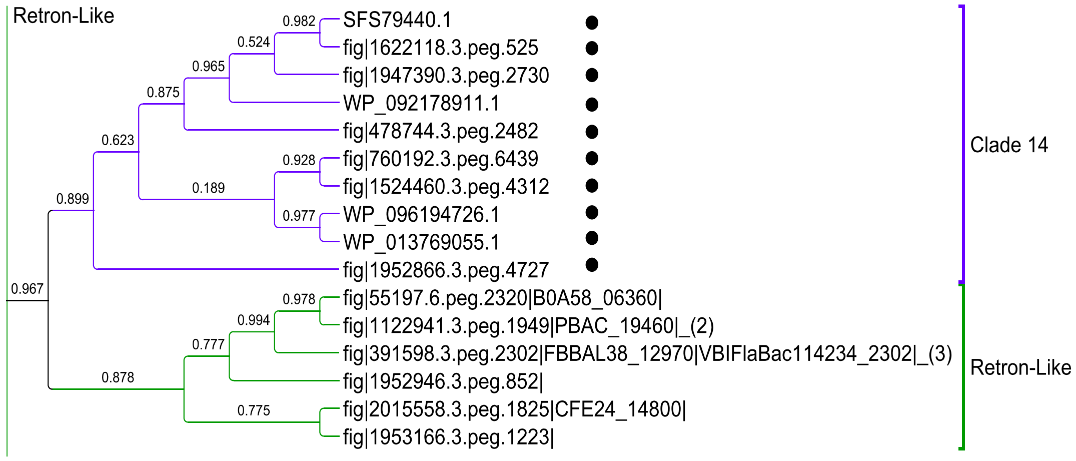
Appendix A3. The phylogenetic trees inferred with IQ-Tree using SH-aLRT/Ufboot or non-parametric bootstrapping (**A** and **B**, respectively) equivalent to the phylogenetic tree obtained in Figure R1.6 are represented. With the exception of group II introns of class C and the variety *g6* the other intron classes were collapsed for the sake of simplicity. The 13 identified clades of RTs associated to type III CRISPR-Cas systems are indicated in color. G2L4 is a group of RTs that lack the intron RNA structure and are not associated to CRISPR-Cas systems**.**



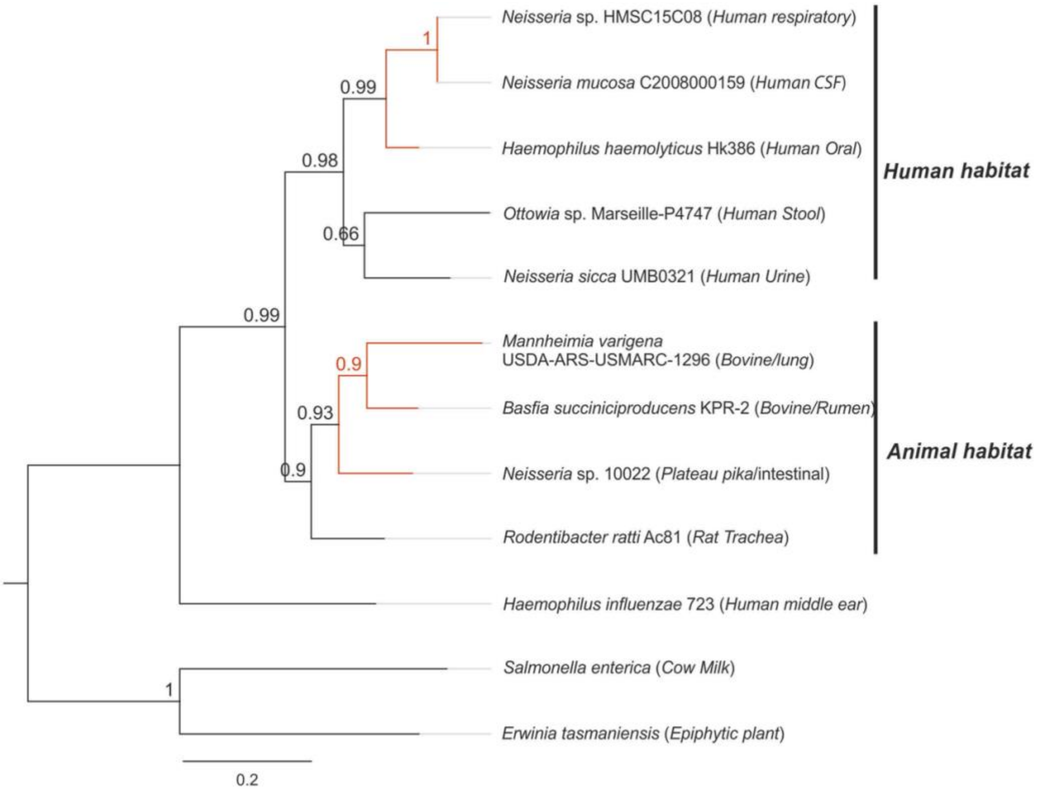
Appendix A4. A cladogram showing the CRISPR-Cas RTs of clades 2, 11 and 12. The protein ID from PATRIC or NCBI is indicated. The dots indicate the identified RT domains for each sequence within the CRISPR-Cas clades: RT alone (red), RTCas1 fusions (blue), Cas6RTCas1 fusions (Green), RTPrim_S fusion (Yellow).



Appendix A5. A cladogram showing the CRISPR-Cas RTs of clade 14 and the closest retron/retron-like RT sequences. The red dots indicate that the RTs are adjacent but not fused to Cas1.



Appendix A6. Phylogeny of RTs linked to CRISPR-Cas systems type I-C. The tree was constructed with FastTree, from an alignment including the Abi-P2 RTs from *Basfia succiniciproducens* and *Haemophilus haemolyticus* strain HK386. Other close relatives identified by Blast searches of the NCBI database also associated with type I-C systems were included. The RTs linked to type I-C systems correspond to the red branches. The host and environment of the isolates are indicated.


1
0.99
0.98
0.66
0.99
0.9
0.93
0.9
1
Neisseria sp. HMSC15C08 (Human respiratory)
Neisseria mucosa C2008000159 (Human CSF)
Haemophilus haemolyticus Hk386 (Human Oral)
Ottowia sp. Marseille-P4747 (Human Stool)
Neisseria sicca UMB0321 (Human Urine)
Human habitat
Mannheimia varigena USDA-ARS-USMARC-1296 (Bovine/lung)
Basfia succiniciproducens KPR-2 (Bovine/Rumen)
Neisseria sp. 10022 (Plateau pika/intestinal)
Rodentibacter ratti Ac81 (Rat Trachea)
Animal habitat
Haemophilus influenzae 723 (Human middle ear)
Salmonella enterica (Cow Milk)
Erwinia tasmaniensis (Epiphytic plant)
0.2

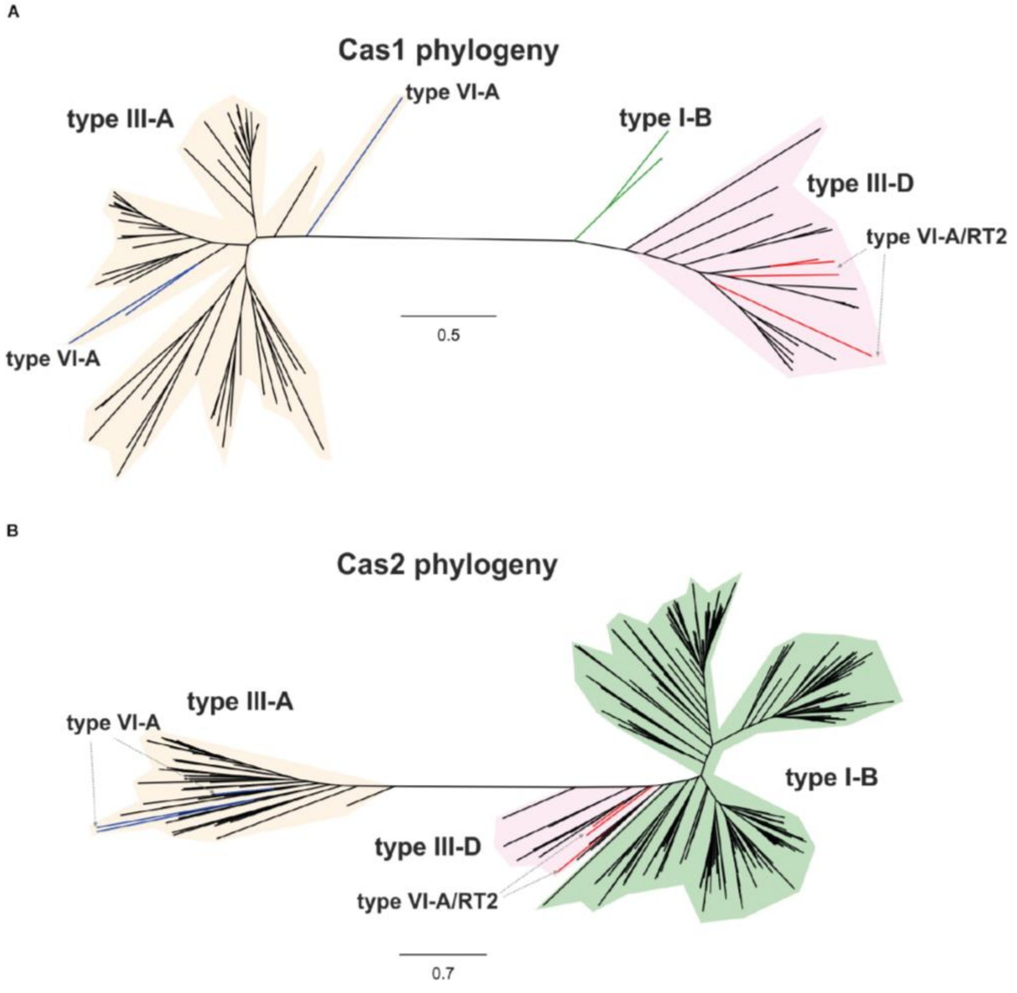
Appendix A7. List of Cas13 proteins used in figure R1.11.

Clade	Node	Source	Nucleotide ID	Cas13a Protein ID	RTCas1[a]	Cas1	Cas2	Surrounding Arrays nº(repeats)	Trans CRISPR-Cas systems	Tra
		1 Herbinix hemicellulosilytica	NZ_QNRW01000010.1	**WP_103203632.1**				1(6)	2x acquisition modules	
		2 Leptotrichia wadei F0279	NZ_KI271395.1	**WP_036059678.1**		WP_021746002.1	WP_021746001.1	1(3)		
		3 Bacteroides ihuae	NZ_FNVX01000005.1	**WP_071146234.1**				1(13)		
		4 Paludibacter propionicigenes	NC_014734.1	**WP_013443710.1**				1(27)		
		5 Carnobacterium gallinarum	NZ_JQLU01000005.1	**WP_034560163.1**				2(7/10)	I-B / Acquisition module	
		6 Carnobacterium gallinarum	NZ_JQLU01000005.1	**WP_034563842.1**				2(7/8)		
		7 Listeria newyorkensis	NZ_UAWS01000035.1	**WP_036091002.1**				1(5)		
		8 Listeria weihenstephanensis	NZ_CP011102.1	**WP_118907415.1**				2(4/9)	I-B	
		9 Listeria costaricensis	NZ_FXUT01000037.1	**WP_099225408.1**				1(4)		
		10 Listeria seeligeri	NC_013891.1	**WP_012985477.1**				1(6)	I-B/A	
1		11 Rhizobium sp. SPY-1	NZ_SMTL01000009.1	**WP_133318297.1**	WP_133318296.1		WP_133318295.1	1(6)		
1		12 Rhizobium sp. FKY42	NZ_STGA01000016.1	**WP_137134457.1**	WP_137134456		WP_137134455.1	2(5/3)	I-E	
1		13 Rhodovulum sp. MB263	NZ_CP020384.1	**WP_080615427.1**	WP_080615428.1		WP_080615428.1	1(3)	Acquisition module / I-E	
1		14 Rhodovulum kholense	NZ_QAYC01000027.1	**WP_108028905.1**	WP_108028906.1		WP_108028907.1	2(4/6)		
1		15 Rhodobacter capsulatus	NZ_AYQC01000019.1	**WP_023911507.1**				1(5)	I-C	
1		16 Ruegeria sp. 318-1	NZ_SMUV01000032.1	**WP_133357912.1**						
1		17 Spirochaeta sp. LUC14_002_19_P3	MUIB01000034.1	**OQX30025.1**					V-B	
1		18 Insolitispirillum peregrinum	NZ_FTOA01000001.1	**WP_076398593.1**				1(4)	I-E / II-C / I-C	
1		19 Rhodovulum viride	NZ_MUAV01000038.1	**WP_112317339.1**				1(9)		
1		20 Thalassospira sp. TSL5-1	NZ_KV880638.1	**WP_073955355.1**				1(6)		
1		21 Thalassospira profundimaris	NZ_JPWH01000001.1	**WP_114086813.1**				1(20)		
1		22 Rhodovulum steppense	NZ_SLVM01000007.1	**WP_132694182.1**				1(9)	I-E / I-E	
1		23 Ferrovibrio sp.	PEKV01000005.1	**PJI41863.1**				1(3)		
1		24 Bradyrhizobium sp. TSA1	NZ_LFJC01000003.1	**WP_100176879.1**						
		25 Maledivibacter halophilus	NZ_FUZT01000022.1	**WP_079495749.1**		WP_079495748.1	WP_079495747.1			
		26 Leptotrichia massiliensis	NZ_FNVZ01000004.1	**WP_071124126.1**		WP_071124127.1	WP_071124128.1	1(7)		
		27 Leptotrichia wadei F0279	KI271424.1	**ERK47820.1**		ERK47819.1	ERK47818.1		Acquisition module	
		28 Leptotrichia wadei	NZ_KI271421.1	**WP_021746774.1**		WP_021746773.1	WP_021746772.1	2(5/4)		
		29 Leptotrichia buccalis	NC_013192.1	**WP_015770004.1**					I-B / III-D	
		30 Leptotrichia massiliensis	NZ_FNVZ01000005.1	**WP_071125398.1**				1(6)		
		31 Leptotrichia sp. oral taxon 225	NZ_KI272904.1	**WP_021768357.1**						
		32 Leptotrichia shahii	NZ_KB890278.1	**WP_018451595.1**		WP_083917144.1	WP_018451593.1	1(4)	I-B / III-A/D	
		33 Leptotrichia sp. oral taxon 879	NZ_KI271320.1	**WP_021744063.1**		WP_021744062.1	WP_021744061.1		III-A	
2		34 Ruminococcus sp. AM40-10AC	NZ_QUIR01000043.1	**WP_118572797.1**				1(7)	II-A / I-B / I-C / I-E	
2		35 [Eubacterium] rectale AF19-3AC	NZ_QRWK01000025.1	**WP_117998314.1**	WP_117998320.1[b]	WP_117998323.1[b]	WP_117998317.1	1(8)		
2		36 **Ruminococcus sp. TF11-2AC**	NZ_QUKI01000013.1	**WP_118614261.1**				1(5)	III-D / I-B	
2		37 **Blautia sp. Marseille-P2398**	NZ_LT546010.1	**WP_062808098.1**		WP_062808097.1	WP_062808329.1	1(16)	III-A/D / I-B	
2		38 *Lachnospiraceae bacterium NK4A179*	NZ_ATWC01000054.1	**WP_022785443.1**				2(4/7)	II-A	
2		39 **Lachnospiraceae bacterium NE2001**	NZ_FOEK01000016.1	**WP_089928016.1**		WP_089928019.1	WP_089928054.1	1(10)		
2		40 Butyrivibrio sp. YAB3001	NZ_FOKR01000002.1	**WP_092321585.1**		WP_092321588.1	WP_092321591.1	1(13)		
2		41 *Lachnospiraceae bacterium MA2020*	NZ_JQKK01000015.1	**WP_044921188.1**		WP_081903028.1	WP_081903027.1	3(8/11/3)	II-A	
2		42 **Pseudobutyrivibrio sp. OR37**	NZ_FOQF01000039.1	**WP_090551759.1**						
2		43 *Lachnospiraceae bacterium NK4A144*	NZ_AUJT01000030.1	**WP_027114339.1**		WP_027114338.1	WP_044982681.1	2(7/4)	I-C / I-E	
2		44 **[Clostridium] aminophilum**	NZ_FOZC01000010.1	**WP_031473346.1**		WP_081844215.1	WP_038288458.1	1(12)	III A/D	
2		45 *Drancourtella sp. An57*	NZ_NFHY01000010.1	**WP_087253216.1**	WP_087253220.1[b]/WP_087253222.1[b]		WP_087253218.1	1(6)		
2		46 **Eubacteriaceae bacterium CHKCI004**	NZ_FCNR01000048.1	**WP_090127496.1**	WP_090127495.1		WP_090127494.1	2(4/4)	III	
2		47 [Eubacterium] rectale T1-815	NZ_CVRQ01000008.1	**WP_055061018.1**				2(5/8)		
2		48 [Eubacterium] rectale TM10-3	NZ_QSOB01000012.1	**WP_117482613.1**			WP_117482614.1	1(6)		
2		49 [Eubacterium] rectale AF25-15	NZ_QRUJ01000006.1	**WP_118003838.1**	WP_118003845.1		WP_117482614.1	1(5/6)		

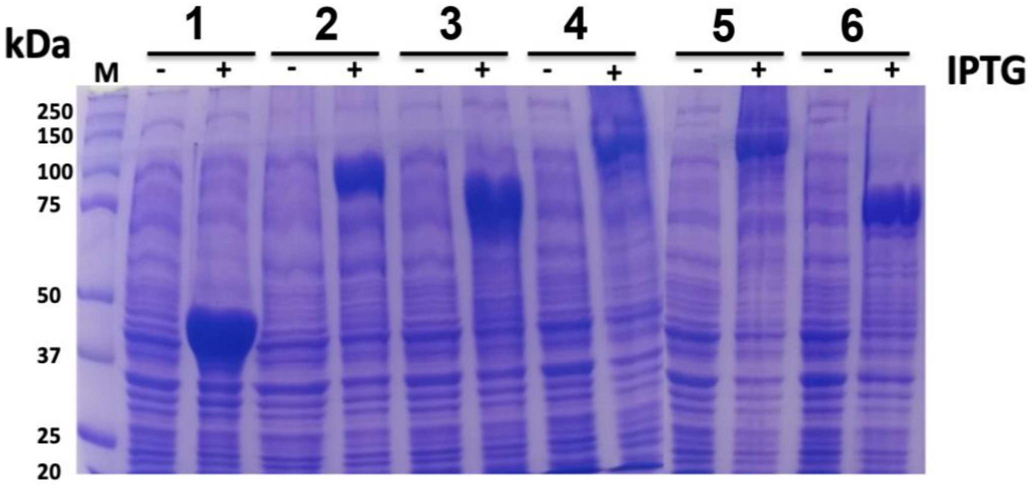
a) RT-Cas1 fusions are highlighted in red

b) A frameshift split RT and cas1 domains

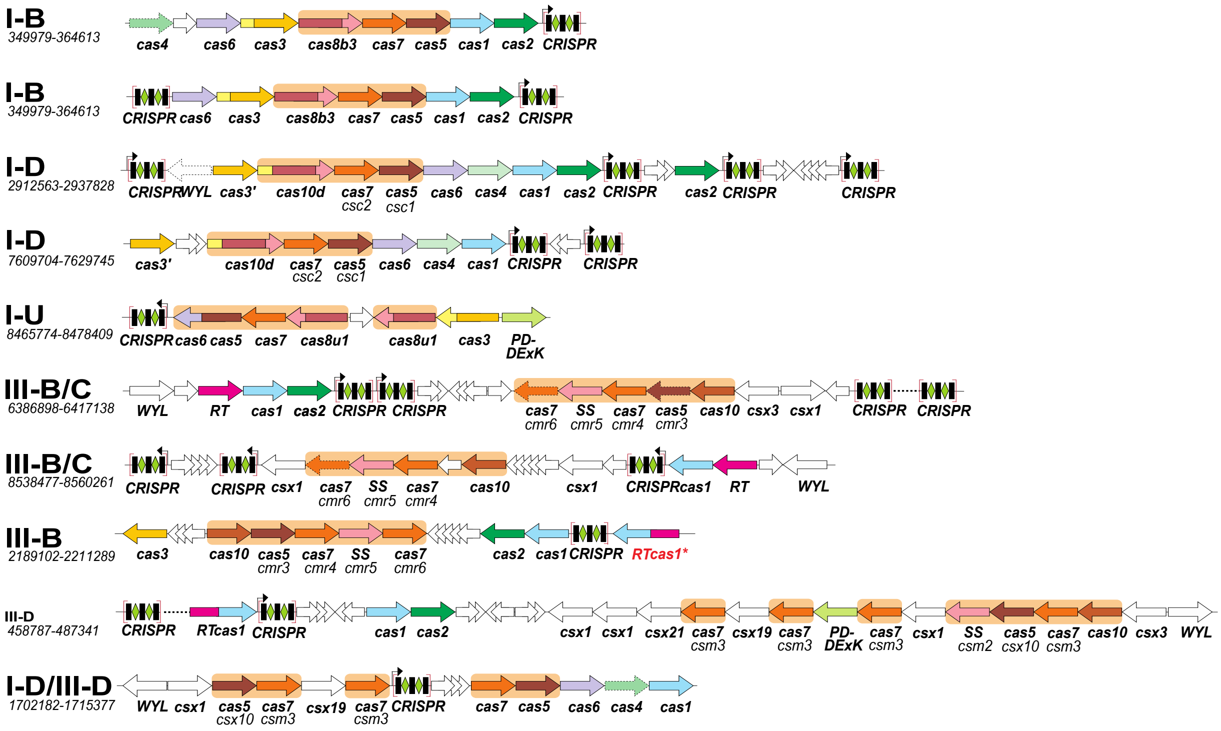
Appendix A8 Phylogeny of RTCas1/Cas1 and Cas2 proteins associated with clade 2 Cas13a sequences. The unrooted trees were constructed with FastTree from alignments of 99 closely related Cas1 sequences **(A)**, and 537 closely related Cas2 sequences **(B)** mostly harbored by bacteria from phylum Firmicutes. The Cas1 and Cas2 sequences and their associated CRISPR-*cas* loci are provided in Toro *et al.*, 2019a (Supplementary Tables S3, S4, respectively). Branches in red indicate the RTCas1 and adjacent Cas2 sequences associated with Cas13a proteins in type VI-A/RT2 systems, and branches in blue show Cas1 sequences lacking an RT domain and Cas2-associated sequences within clade 2. Other identified CRISPR-Cas systems, classified according to the interference module, are also specified.



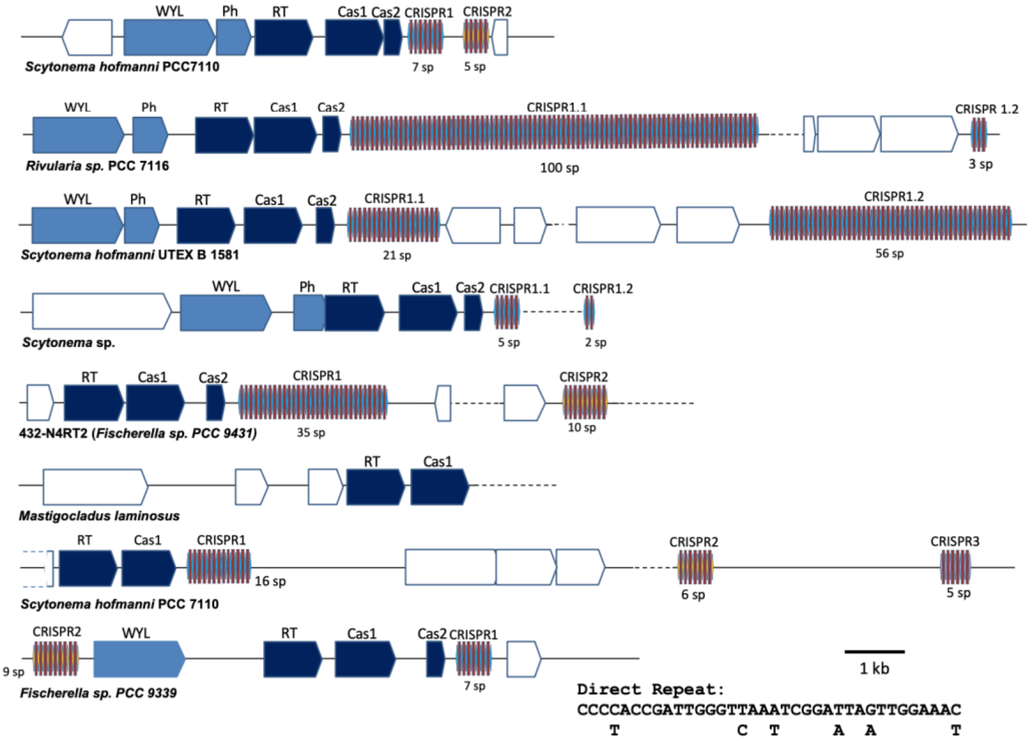
Appendix B1. Expression of pMal-Flag-RT or RTCas1 construct vectors. SDS-PAGE gel (10%) showing the level of expression of the different RT or RTCas1 proteins. (-) previous and (+) after induction with IPTG. The recombinant proteins are the following: MBP-Flag (used as control) (1), MBP-RT from *Roxeiflexus castenholzii* DSM 13941 (2), MBP-RT from *Scytonema hofmanni* PCC 7110 (3), MBP-RTCas1 from *Vibrio vulnificus* YJ016 (3), MBP-RTCas1 from *Chlorobium limicola* DSM 245 (5) and the MBP-RT from *Desulfobacca acetoxidans* DSM 11109 (6). (M, Molecular Weight Marker in kDa).



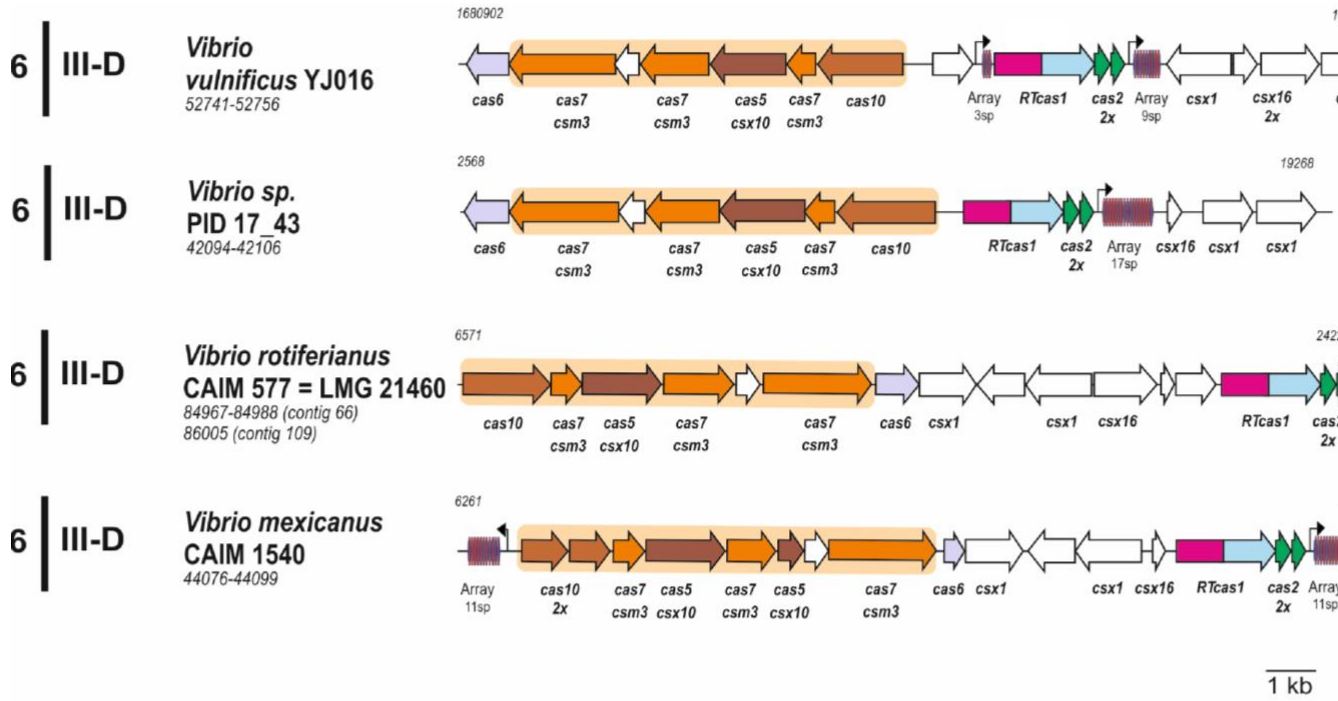
Appendix B2. Architecture of CRISPR-*cas* loci harbor by *Scytonema hofmanni* strain PCC 7110.
modules are indicated in beige background. Cas3 helicase/nuclease is shown in yellow. The gene encoding Ca
in purple. RT and Cas1 domains are indicated in fuchsia and blue, respectively, whereas *cas2* gene are colo
CRISPR Arrays are shown in brackets and are not to scale, as the rest of the loci. Black arrows indicate the lead
Ancillary and unknown genes are not color-coded. The CRISPR-Cas type and the genomic coordinates are
left.


I-B
349979-364613
cas4 cas6 cas3 cas8b3 cas7 cas5 cas1 cas2 CRISPR
I-B
349979-364613
CRISPR cas6 cas3 cas8b3 cas7 cas5 cas1 cas2 CRISPR
I-D
2912563-2937828
CRISPR WYL cas3' cas10d cas7 csc2 cas5 csc1 cas6 cas4 cas1 cas2 CRISPR cas2 CRISPR CRISPR
I-D
7609704-7629745
cas3' cas10d cas7 csc2 cas5 csc1 cas6 cas4 cas1 CRISPR CRISPR
I-U
8465774-8478409
CRISPR cas6 cas5 cas7 cas8u1 cas8u1 cas3 PD-DExK
III-B/C
6386898-6417138
WYL RT cas1 cas2 CRISPR CRISPR cas7 cmr6 SS cmr5 cas7 cmr4 cas5 cmr3 cas10 csx3 csx1 CRISPR CRISPR
III-B/C
8538477-8560261
CRISPR CRISPR csx1 cas7 cmr6 SS cmr5 cas7 cmr4 cas10 csx1 CRISPR cas1 RT WYL
III-B
2189102-2211289
cas3 cas10 cas5 cmr3 cas7 cmr4 SS cmr5 cas7 cmr6 cas2 cas1 CRISPR RTcas1*
III-D
458787-487341
CRISPR RTcas1 CRISPR cas1 cas2 csx1 csx1 csx21 cas7 csm3 csx19 cas7 csm3 PD-DExK cas7 csm3 csx1 SS csm2 cas5 csx10 cas7 csm3 cas10 csx3 WYL
I-D/III-D
1702182-1715377
WYL csx1 cas5 csx10 cas7 csm3 csx19 cas7 csm3 CRISPR cas7 cas5 cas6 cas4 cas1

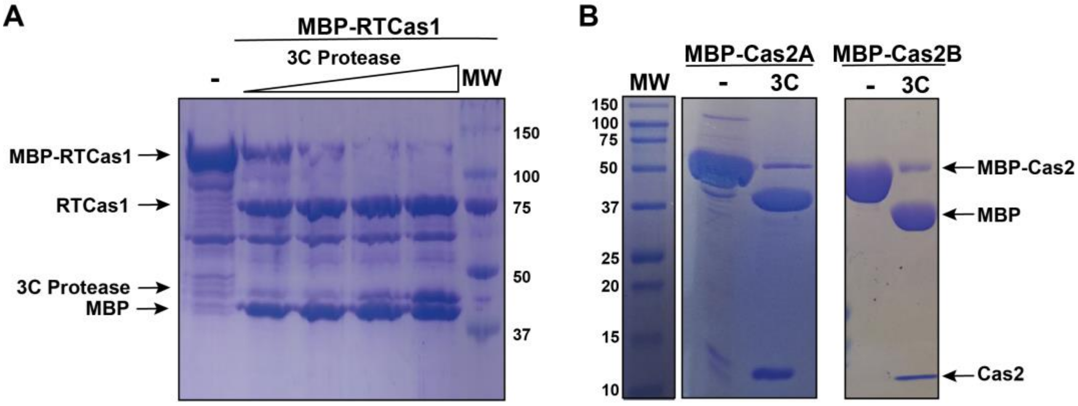
Appendix B3. Architecture of CRISPR-Cas adaptation modules phylogenetically related to the RT alone of *Scytonema hofmanni* PCC 7110. Adaptive genes (*RT, cas1 and cas2*) are indicated in dark blue. A WYL-domain containing protein and a putative phosphohydrolase (Ph) that usually appears just upstream of the adaptive operon are shown in light blue. Unknown genes are non-colored coded. The CRISPR Arrays are indicated with red and blue rectangles. The number (1, 2 or 3) of the Arrays indicates different sequences of the Direct Repeat. The consensus sequence of the most extended array (CRISPR1) Direct repeat is indicated below. The name of the strain containing is shown just below of each of each CRISPR-*cas* locus



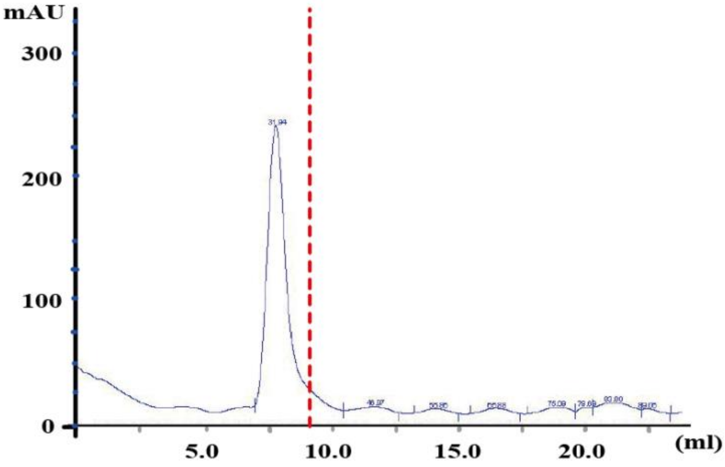
Appendix B4. Architecture of type III-D CRISPR-Cas systems closely related to that from *V. vulnificus* representative operon shown in this figure belongs to Vibrio species clustered in clade 6 of RT-CRISPR phy effector modules are indicated in beige background. The gene encoding Cas6 is colored in purple. RT and C are indicated in fuchsia and blue, respectively, whereas *cas2* gene are colored in green. The CRISPR Arrays of species are indicated. The black arrows indicate the putative promoter sequence Ancillary and unknown g color-coded. For each locus, the clade of the RT-CRISPR phylogeny, the type of CRISPR-Cas system, the organism, the corresponding gene locus tag (final digits) and the genomic coordinates are indicated.



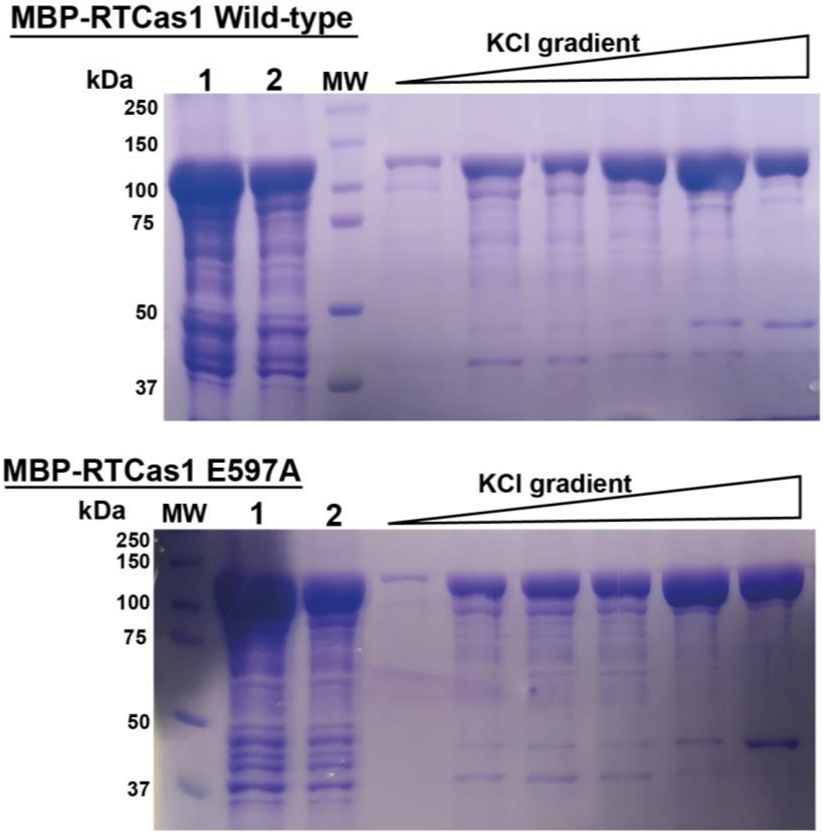
Appendix B5. Cleavage of RTCas1, Cas2A and Cas2B of the VvYJ016 adaptive module with the 3C protease. (A) Cleavage of MBP-RTCas1 with 3C protease. 10 µg of MBP-RTCas1 are added in each point. A control (-) and a decreasing gradient of 3C (2, 0.4, 0.1 and 0,05 units) protease are shown in a 10% SDS-PAGE gel stained with Coomassie blue. **(B)** Cleavage of MBP-Cas2A and MBP-Cas2B with 3C protease. 20 µg of protein are shown before (-) and after (3C) cleavage with 1 unit of 3C protease (4ºC overnight).



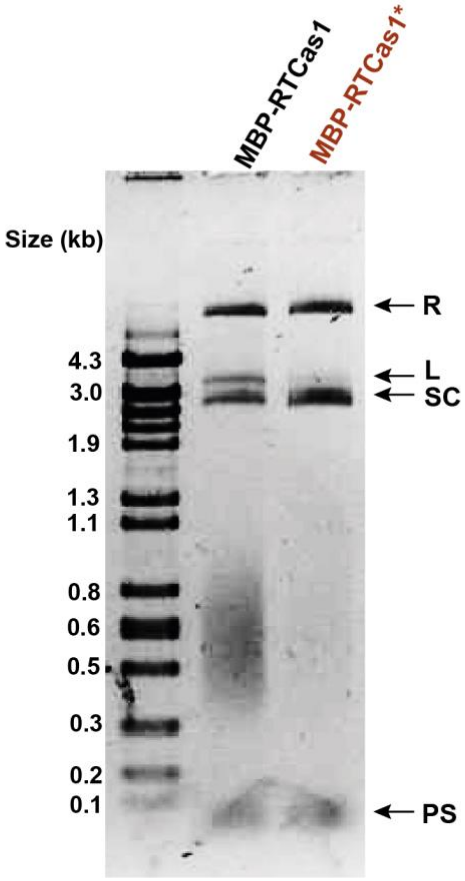
Appendix B6. Oligomeric state of MBP-RTCas1 of *V. vulnificus* YJ016. 300 µg of purified MBP-RTCas1 (Figure R2.10) was loaded into Superdex 200 increase 10/300 GL. A single peak eluted before the void volume (dashed red line; 8 ml), suggest that the protein mainly form a soluble aggregate (>600 kDa).



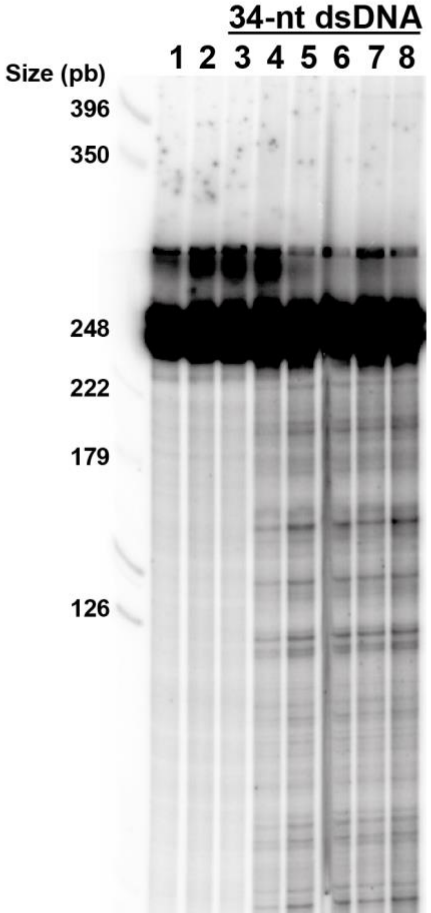
Appendix B7. Purification process of MBP-RTCas1 wild-type (above) and the E597A mutant (below). Both proteins were purified using amylose and heparin columns. The eluted fractions after elution with maltose (1), the flow-through of the heparin (2) and the eluted fractions after a KCl gradient (0.25 to 1.5 M) are shown.



Appendix B8. Integration assay using a MBP-RTCas1 sample with and without nucleic acid contamination. MBP-RTCas1 was purified following the amylose beads protocol (section M.11.2). MBP-RTCas1* was purified using an additional step in which polyethylenimine (PEI) was added to remove nucleic acids present in the sample. 450 nM of each protein were used over pCRISPR-439. A smear is detected in the MBP-RTCas1 sample (left). Products are separated by agarose gel electrophoresis (1%). Relaxed (R), linear (L), and supercoiled (SD) pCRISPR-439, as well as protospacer (PS) are indicated.



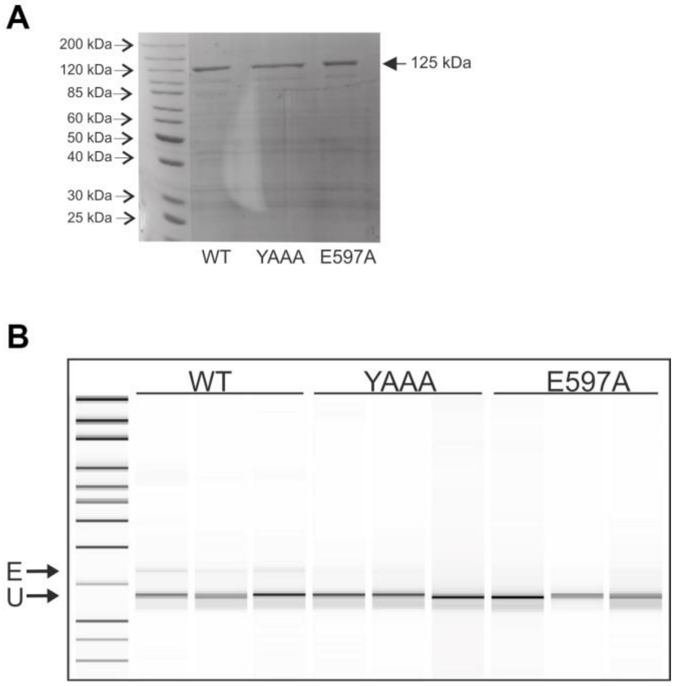
Appendix B9. Cleavage/ligation assay with the proteins of VvYJ016 adaptation module. Internally labeled CRISPR DNA of 248 nt containing part of the leader sequence, the first two direct repeats, and the first two spacers. This substrate was incubated with several negative controls: $\mathbf{H_2O}$ (lane 1); reaction buffer (lane 2); 34-nt dsDNA and the reaction buffer (lane 3). The substrate and the 34-nt dsDNA in the reaction buffer were incubated with: MBP-RTCas1 (5µM) (lane 4); MBP-RTCas1(5µM), MBP-Cas2A (2.5µM) and MBP-Cas2B (2.5µM) (lane 5); MBP-RTCas1(5µM), MBP-Cas2A (7.5µM) and MBP-Cas2B (7.5µM) (lane 6); MBP-RTCas1(5µM), MBP-Cas2A (15 µM) (lane 7); MBP-RTCas1(5µM), MBP-Cas2B (15 µM) (lane 8).



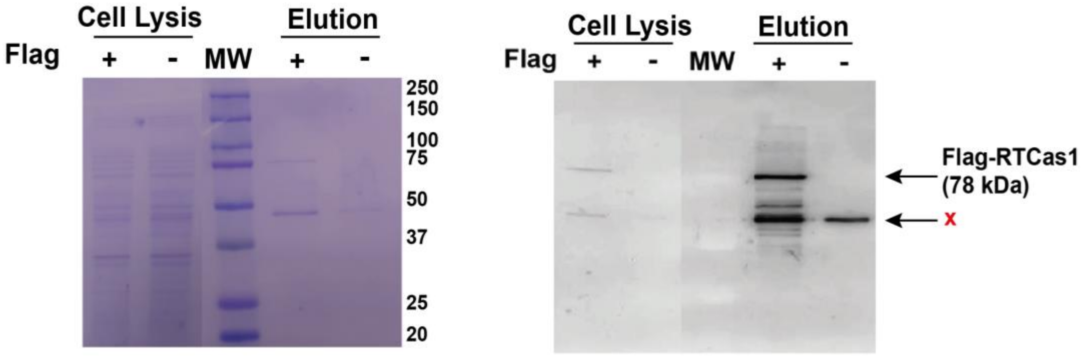
Appendix C1. Total number of genome- and plasmid-derived spacer acquisition events in all experiments with wild-type RTCas1 in this thesis.

	Genomic	Plasmids
Number of spacers	177,876	8,531

Appendix C2. Spacer acquisition of wild-type RTCas1 and mutants. (A) SDS-Page gel electrophoresis of 0.5 µg of purified MBP fusion proteins WT, YAAA and E597A mutants used in the *in vitro* RT assays. Similar yields of MBP fusion protein of 125 kDa can be observed in all samples. **(B)** Bioanalyzer results for purified sliced bands after two rounds of PCR in three replicates of the spacer acquisition assay performed with the wild-type (WT), the RT mutant (YAAA) and the mutant with an inactivated Cas1 domain (E597A). In the three replicates of the WT assay, expanded arrays (E) were detected, whereas, in the YAAA and E597A assays, only the unexpanded array (U) was visible.



Appendix C3. Pull-down assays of Flag-RTCas1 in *Vibrio vulnificus* YJ016. Detection of Flag-RTCas1 (+) using anti-Flag antibodies in samples from *V. vulnificus* YJ016 growth until exponential phase (M.12.2.2). RTCas1 without Flag-Tag was used as a negative control (-). 1 µl of total protein after cell lysis and 2 µL of the eluted protein after the co-inmunoprecipitation step with 3XFLAG (M.12.2.2) were subjected to SDS-PAGE and stained with Coomassie Blue (left gel) and a western blot was performed with antibodies against Flag (right gel). The position of the Flag-RTCas1 protein is indicated as well as inespecific cross-reacting antiflag proteins (x).



Appendix C4. High-throughput sequencing data presented in this thesis.

accession	study	Bioproject accession	Biosample accession	sample_name	library_ID	title	design_description
SRR8962134	SRP193951	PRJNA539885	SAMN11514629	pAGDt-439_pCA2s-1DR-Rep1	1_S1_wt_Array 2	Spacer acquisition VvRT-Cas1-Array2-1	RT-Cas1, Cas2A+B expre (DE3) E. coli host over Ar independent cultures - replic
SRR8962133	SRP193951	PRJNA539885	SAMN11514630	pAGDt-439_pCA2s-1DR-Rep2	1_S2_wt_Array 2	Spacer acquisition VvRT-Cas1-Array2-2	RT-Cas1, Cas2A+B expre (DE3) E. coli host over A independent cultures - replic
SRR8962132	SRP193951	PRJNA539885	SAMN11514631	pAGDt-439_pCA2s-1DR-Rep3	1_S3_wt_Array 2	Spacer acquisition VvRT-Cas1-Array2-3	RT-Cas1, Cas2A+B expre (DE3) E. coli host over Ar independent cultures - replic
SRR8962131	SRP193951	PRJNA539885	SAMN11514632	pAGDt-439_pCA2s-1DR-Rep4	1_S4_wt_Array 2	Spacer acquisition VvRT-Cas1-Array2-4	RT-Cas1, Cas2A+B expre (DE3) E. coli host over Ar independent cultures - replic
SRR8962138	SRP193951	PRJNA539885	SAMN11514633	pAGDt-439_pCA2s-1DR-Rep5	1_S5_wt_Array 2	Spacer acquisition VvRT-Cas1-Array2-5	RT-Cas1, Cas2A+B expre (DE3) E. coli host over Ar independent cultures - replic
SRR8962137	SRP193951	PRJNA539885	SAMN11514634	pAGDt-439_pCA2s-1DR-Rep6	1_S6_wt_Array 2	Spacer acquisition VvRT-Cas1-Array2-6	RT-Cas1, Cas2A+B expre (DE3) E. coli host over Ar independent cultures - replic
SRR8962136	SRP193951	PRJNA539885	SAMN11514635	pAGDt-439_pCA1s-1DR-Rep1	1_S7_wt_Array 1	Spacer acquisition VvRT-Cas1-Array1-1	RT-Cas1, Cas2A+B expre (DE3) E. coli host over Ar independent cultures - replic
SRR8962135	SRP193951	PRJNA539885	SAMN11514636	pAGDt-439_pCA1s-1DR-Rep2	1_S8_wt_Array 1	Spacer acquisition VvRT-Cas1-Array1-2	RT-Cas1, Cas2A+B expre (DE3) E. coli host over Ar independent cultures - replic
SRR8962140	SRP193951	PRJNA539885	SAMN11514637	pAGDt-439-YAAA_pCA2s-1DR-Rep1	1_S9_YAAA_Array 2	Spacer acquisition YAAA mutant-1	RT-Cas1 (YAAA-mutant), in HMS 174 (DE3) E. coli ho of three independent cultures
SRR8962139	SRP193951	PRJNA539885	SAMN11514638	pAGDt-439-YAAA_pCA2s-1DR-Rep2	1_S10_YAAA_Array 2	Spacer acquisition YAAA mutant-2	RT-Cas1 (YAAA-mutant), in HMS 174 (DE3) E. coli ho of three independent cultures
SRR8962120	SRP193951	PRJNA539885	SAMN11514639	pAGDt-439-YAAA_pCA2s-1DR-Rep3	1_S11_YAAA_Array 2	Spacer acquisition YAAA mutant-3	RT-Cas1 (YAAA-mutant), in HMS 174 (DE3) E. coli ho of three independent cultures

SRR8962119	SRP193951	PRJNA539885	SAMN11514640	pAGDt-439-E517A_pCA2s-1DR	1_S12_E517A_Array 2	Spacer acquisition E517A mutant	RT-Cas1 (E517A-mutant), in HMS 174 (DE3) E. coli h of three independent cultures
SRR8962122	SRP193951	PRJNA539885	SAMN11514641	pAGDt-439-E597A_pCA2s-1DR	1_S13_E597A_Array 2	Spacer acquisition E597A mutant	RT-Cas1 (E597A-mutant), in HMS 174 (DE3) E. coli h of three independent cultures
SRR8962121	SRP193951	PRJNA539885	SAMN11514642	pAGDt-439-Cas2A_pCA2s-1DR-Rep1	1_S14_DCas2A_Array2	Spacer acquisition DCas2A mutant-1	RT-Cas1, Cas2B expressed i coli host over Array 2. Mix cultures - replicate 1
SRR8962124	SRP193951	PRJNA539885	SAMN11514643	pAGDt-439-Cas2A_pCA2s-1DR-Rep2	1_S15_DCas2A_Array2	Spacer acquisition DCas2A mutant-2	RT-Cas1, Cas2B expressed i coli host over Array 2. Mix cultures - replicate 2
SRR8962123	SRP193951	PRJNA539885	SAMN11514644	pAGDt-439-Cas2A_pCA2s-1DR-Rep3	1_S16_DCas2A_Array2	Spacer acquisition DCas2A mutant-3	RT-Cas1, Cas2B expressed i coli host over Array 2. Mix cultures - replicate 3
SRR8962126	SRP193951	PRJNA539885	SAMN11514645	pAGDt-439-Cas2B_pCA2s-1DR-Rep1	1_S17_DCas2B_Array2	Spacer acquisition DCas2B mutant-1	RT-Cas1, Cas2A expressedi coli host over Array 2. Mix cultures - replicate 1
SRR8962125	SRP193951	PRJNA539885	SAMN11514646	pAGDt-439-Cas2B_pCA2s-1DR-Rep2	1_S18_DCas2B_Array2	Spacer acquisition DCas2B mutant-2	RT-Cas1, Cas2A expressedi coli host over Array 2. Mix cultures - replicate 2
SRR8962128	SRP193951	PRJNA539885	SAMN11514647	pAGDt-439-Cas2B_pCA2s-1DR-Rep3	1_S19_DCas2B_Array2	Spacer acquisition DCas2B mutant-3	RT-Cas1, Cas2A expressed i coli host over Array 2. Mix cultures - replicate 3
SRR8962127	SRP193951	PRJNA539885	SAMN11514648	pAGDt-439-Cas2AB_pCA2s-1DR	1_S20_DCas2AB_Array2	Spacer acquisition DCas2AB mutant	RT-Cas1 expressed in HM host over Array 2. Mix c cultures
SRR8962129	SRP193951	PRJNA539885	SAMN11514649	pAGDt-439_pCA2s-1DR-80c	L1_wt_Array2	Spacer acquisition (80) VvRT-Cas1-Array2	RT-Cas1, Cas2A+B expre (DE3) E. coli host over A independent cultures
SRR8962130	SRP193951	PRJNA539885	SAMN11514650	pAGDt-439-tdI_pCA2s-1DR	L_NGS_wt-tdI-Array2	Spacer acquisition (80_tdI) VvRT-Cas1-Array2	Targeted DNA sequencing t td intron RNA is present as (DE3) E. coli host over Arr RT-Cas1, Cas2A+Cas2B - M cultures