UNIVERSIDAD DE GRANADA

CONSEJO SUPERIOR DE INVESTIGACIONES
CIENTÍFICAS

EL GENOMA DEL ENDOSIMBIONTE DIAZOTRÓFICO *Sinorhizobium meliloti* GR4: DISPERSIÓN NATURAL Y APLICACIÓN BIOTECNOLÓGICA DE NUEVOS INTRONES DEL GRUPO II

LAURA MARTÍNEZ RODRÍGUEZ

TESIS DOCTORAL
GRANADA 2013

Editor: Editorial de la Universidad de Granada
Autor: Laura Martínez Rodríguez
D.L.: GR 360-2014
ISBN: 978-84-9028-765-1

El genoma del endosimbionte diazotrófico *Sinorhizobium meliloti* GR4: dispersión natural y aplicación biotecnológica de nuevos intrones del grupo II

Memoria que presenta la licenciada en Biología
Dña. Laura Martínez Rodríguez
como aspirante al Grado de Doctor
por la Universidad de Granada

Fdo: **Laura Martínez Rodríguez**

Universidad de Granada
2013

Los Directores de la Tesis Doctoral,
Dr. Francisco Martínez-Abarca Pastor,
Investigador Científico del CSIC,
y Dr. Nicolás Toro García,
Profesor de Investigación del CSIC,

CERTIFICAN

Que los trabajos de investigación desarrollados en
la Memoria de Tesis Doctoral son aptos para ser presentados
por la Lda. Laura Martínez Rodríguez ante el Tribunal que
en su día se designe, y que, en la realización del trabajo,
se han respetado los derechos de otros autores a ser citados
cuando se han utilizado sus resultados o publicaciones.

Y para que así conste, en cumplimiento de las disposiciones vigentes,
extendemos el presente certificado a 16 de mayo de 2013

Fdo: Dr. Francisco Martínez-Abarca Pastor

Fdo: Dr. Nicolás Toro García

Este trabajo de Tesis Doctoral ha sido realizado en el Grupo de Ecología Genética de la Rizosfera del Departamento de Microbiología y Sistemas Simbióticos de la Estación Experimental del Zaidín (EEZ-CSIC), en Granada.

Para la realización de este trabajo, la Lda. Laura Martínez Rodríguez fue financiada mediante las siguientes fuentes:

- Beca predoctoral JAE-PREDOC, desde septiembre de 2008 hasta agosto de 2012.
- Beca de movilidad para estancias breves del CSIC, disfrutada en el Departamento de Biología de la Universidad de York (Reino Unido), bajo la dirección del Profesor J. Peter W. Young. Periodo: Julio-Octubre 2011.

Parte de los resultados presentados en esta Tesis Doctoral han sido publicados en revistas internacionales o están en preparación:

Martínez-Abarca, F., Martínez-Rodríguez, L., López-Contreras, J.A., Jiménez-Zurdo, J.I. & Toro, N. (2013). Complete genome sequence of the alfalfa symbiont *Sinorhizobium/Ensifer meliloti* strain GR4. *Genome Announc*. **1**(1).

Toro, N., Martínez-Rodríguez, L. & Martínez-Abarca, F. (2013). Insights from a bacterial group II intron fossil record in the nitrogen-fixing symbionts *Sinorhizobium meliloti* and *Sinorhizobium medicae* genomes. *Genome Biology and Evolution*. Enviado.

Martínez-Rodríguez, L., Martínez-Abarca, F. & Toro, N. RmInt1-related group II introns. A model of successful dispersion of retroelements. En preparación.

Martínez-Rodríguez, L. et al. Success in the genome closure of *Sinorhizobium meliloti* GR4: a strategy to solve the location and ambiguities (SNPs) of repeated elements based on NGS genomic data. En preparación.

A los que me quieren.

AGRADECIMIENTOS

Después de cinco años de trabajo, por fin se acaba esta etapa de mi vida, en la que he pasado momentos de alegría y momentos de estrés, pero siempre en buena compañía. En estas primeras líneas quisiera dar las gracias a todas esas personas que me han ayudado y apoyado durante el desarrollo de esta Tesis Doctoral.

A mis directores de Tesis, Francisco Martínez-Abarca y Nicolás Toro, por haberme dado la oportunidad de llevar a cabo este trabajo, por enseñarme a plantear y realizar experimentos y a discutir los resultados. En especial a Paco, por su dedicación durante estos años.

A José Ignacio y Manolo, por su disponibilidad y ayuda en los años que he pasado en el laboratorio. A Fernando, por estar siempre dispuesto a escuchar y discutir mis resultados. A Pablo, por su ayuda con los pedidos. A Tita, por esos favores a primera hora de la mañana.

A Peter Young, por su amabilidad y aportación científica en los meses que me dejó formar parte de su grupo. Nunca olvidaré lo bien que lo pasamos en la celebración de su 60º cumpleaños con bailes típicos irlandeses. Gracias a todos los que conocí en la Universidad de York y me amenizaron la estancia lejos de casa, especialmente a Anja, con la que pasé grandes momentos dentro y fuera del laboratorio.

A todos los miembros del tribunal, por aceptar formar parte de este gran día. En especial a Pepi, a quien conocí como profesora en la carrera, y con la que pasé, junto al resto de miembros del grupo de Saltamontes, mis cuatro primeros años como "científica". Al director de ese grupo, Juan Pedro, le debo mi inmersión en la ciencia: gracias por confiar en esa chiquilla de ter-

cero de carrera. Durante esos años en el Departamento de Genética conocí a grandes compañeros de profesión, y muchos se convirtieron en amigos. A todos ellos, gracias por ayudarme en el principio de este largo camino, por esas tardes en la becaria, por esas reuniones fuera del trabajo.

A mis compis de labo, por aguantarme y hacer más agradable el duro trabajo de laboratorio, por su amistad dentro y fuera del labo. A Isa, quien, además de compañera de taca en los últimos años, se convirtió poco a poco en una gran amiga. Por todo lo que me has dado profesional y personalmente. A Mercedes, por pensar en mi cuando venías con el peque. A José Antonio, por ese verano de calentamientos de cabeza con GR4 y la aportación de sus conocimientos informáticos que tanto me ayudaron. A Omar, por sus charlas a través del "euro-túnel" y la resolución de mis dudas cuando Isa se nos fue a las Américas. A Lola, que todo lo sabe, por su ayuda desde el principio hasta el final, incluso a través de e-mails. A Rafa, por las conversaciones sobre introns, tan necesarias en todo momento. A Hurry, por estar siempre dispuesto a ayudar, aunque sea desde la distancia. A Pepe, por aguantar mis insistentes preguntas y mi presencia en su ordenador. A Ali, por ese primer contacto con el labo y la organización del tan preciado volley-playa. A Gloria, por esos días en Alemania en los que conocí mejor a la chica "del otro labo". A Chema, por su compañía y favores desde que llegó.

A esos amigos con los que me gustaría pasar más tiempo pero que, por unas cosas u otras, no veo tan a menudo como quisiera. A Tere, por tantas risas y momentos inolvidables, por tus alegres visitas con Sergio (esperadme, que algún día nos veremos por Londres). A Sergio, por ese viaje a Los Caños donde casi salimos en las noticias. A María (wishi), por los helaítos y días de playa, por esa noche de la que siempre nos reiremos. A Mudarra, por esa amistad duradera. A Belén y Matías, por esas meriendas-cenas acompañadas de largas partidas de mús, por ser siempre tan atentos.

Agradecer también a toda mi familia (gaditana y murciana) su apoyo, no sólo en los últimos años, sino siempre que lo he necesitado. A mis abuelos,

que me lo han dado todo. A mis tíos y primos, por tratarme como una hija y hermana, ayudándome en los buenos y malos momentos. Por esas aventuras en barca, por esas visitas guiadas tanto en el extranjero como en España, por esas tartas, por esos viajes a Granada. A mi familia granaína, por todo el cariño que me han dado. A Araceli, Leticia y Oliver, por esos días de juego y distracción que tanto me gustan. A Jesús y Maribel, por hacerme sentir como en casa, por vuestro trato tan cercano. A mi hermana, porque, aunque estemos lejos, siempre nos tendremos la una a la otra. A mis padres, por todo lo que han hecho para que yo haya podido llegar hasta aquí, por lo orgullosos que están de mi. A mi padre, por su preocupación por mi bienestar a pesar de la distancia. A mi madre, por sus sacrificios para sacarme adelante, porque lo da todo por su niña. A ti, Jesús, por toda tu ayuda tanto en la Tesis como en el día a día, por soportar mis malas rachas, por lo bien que me haces sentir. Todo agradecimiento es poco, no tengo palabras para expresar lo importante que eres en mi vida.

A todos los citados, y a los que se me olvidan, gracias de todo corazón.

SUMMARY

Next-generation sequencing technologies developed in recent years have let an increase in the number of bacterial genome sequencing projects. Currently, Proteobacteria is the phylum which presents the larger number of bacterial genome sequencing projects, and Alphaproteobacteria contains more than 600 completely sequenced genomes. Within this class, rhizobia is a well studied soil bacterial group that establishes symbiotic interaction with leguminous plants. *Sinorhizobium* bacteria are the microsymbionts of *Medicago* (e.g., *M. sativa* and *M. truncatula*), *Melilotus* and *Trigonella* legume species. *S. meliloti* 1021 was the first sequenced strain within this genera, and it is considered one of the rhizobia reference genomes.

In recent years several isolates belonging to *S. meliloti* species have been completely sequenced, such as SM11, AK83, BL225C, Rm41 and 2011. In chapter 1 we present the complete genome sequencing project of *S. meliloti* GR4, a widely studied strain isolated from Granada agronomic soils 30 years ago. It contains a 7.14 Mb size genome composed of five replicons: chromosome (3.6 Mb), two cryptic plasmids, pRmeGR4a (176 kb) and pRmeGR4b (226 kb), and two symbiotic plasmids, pRmeGR4c (1.4 Mb) and pRmeGR4d (1.7 Mb). In order to perform genome resolution, a new parameter called Fold Coverage Index (FCI) has been described. FCI estimates the copy number of any genome sequence, and it has been useful to assign any sequence to an specific replicon.

An important constituent of bacterial genomes is the mobilome, which comprises all mobile genetic elements (MGEs) of a genome. One of these MGEs are group II introns. They are catalytic RNAs and mobile retroelements initially found in organelles of plants and lower eukaryotes, and later

identified in bacteria and archaea. One of the best studied group II introns is RmInt1, discovered in *S. meliloti* GR4. We have characterized two RmInt1 phylogenetically related group II introns: RmInt2 (a novel intron found in *S. meliloti* GR4) and SmedInt1 (discovered in *S. medicae* WSM419 and named Sr.md.I1 in the group II intron database). Their distribution as well as the dispersion of their DNA target (different IS elements) are analyzed *in silico* and compared with RmInt1 spread (chapter 2).

Group II introns consist of a highly structured RNA (ribozyme) in six domains (DI to DVI) and an internally encoded protein (known as IEP) within DIV. Both components, ribozyme and IEP, are necessary for excision and mobility processes, and several studies have demonstrated that host factors are also required. In chapter 2 we analyze the host chaperone GroEL implication in RmInt1 excision. In chapter 3 we demonstrate the excision and mobility of RmInt2 and SmedInt1. By constructing chimeric introns that combine different IEPs and ribozymes from studied introns (RmInt1, RmInt2 and SmedInt1) we empirically corroborate the suggested coevolution between these two components (chapter 3). In addition, RmInt2 can increase the group II intron collection available to be used as a gene-targeting vector for genetic enginering since it shows mobility in heterologous hosts as *Escherichia coli* (chapter 3).

ÍNDICE GENERAL

INTRODUCCIÓN

I.1 SECUENCIACIÓN MASIVA

La secuenciación del DNA ha sido ampliamente utilizada en biología desde hace más de 30 años, y hoy en día es una herramienta indispensable para la investigación en cualquier campo. A mediados de la década de los 70 se desarrollaron métodos rápidos de secuenciación del DNA que sustituyeron a los anteriores métodos usados, como eran el método "de punto corrido" de Gilbert & Maxam (1973) o el método de "más y menos" de Sanger & Coulson (1975). El primero de estos métodos rápidos publicado fue el propuesto por Maxam & Gilbert (1977), en el cual se modifica químicamente el DNA y seguidamente se corta en bases específicas. Sin embargo, el método de terminación de cadena desarrollado por Sanger *et al* (1977), basado en el uso de dideoxinucleótidos trifosfato (ddNTPs; nucleótidos a los que les faltan dos grupos trifosfato, impidiendo así la unión de otro nucleótido), se acabó imponiendo por su sencillez y precisión. Con este método se llevó a cabo la secuenciación del primer genoma, el bacteriófago ϕX174, de 5.386 nt (Sanger *et al*, 1978). La secuenciación mediante el método Sanger se ha ido mejorando con el uso de etiquetas fluorescentes en vez de radiactivas, para detectar la marca de terminación, y con el uso de electroforesis capilar en vez de geles de poliacrilamida, para discriminar las bases. A pesar de que se han explorado varias aproximaciones para reemplazarlo, como la secuenciación por hibridación, por microscopía de fuerza atómica, por espectómetro de masas o por síntesis, el método Sanger ha sido la tecnología dominante durante cerca de 20 años (Rothberg & Leamon, 2008).

Los proyectos de secuenciación a gran escala normalmente han necesitado la clonación de fragmentos de DNA en vectores bacterianos, amplificación y purificación de las muestras individuales, y una secuenciación por el método Sanger (figura I.1A). Éste se ha considerado una tecnología de secuenciación de primera generación, que finalmente ha sido reemplazado por nuevos métodos de secuenciación llamados de siguiente generación (NGS, *Next Generation Sequencing*) que llevan a cabo una secuenciación más versátil (Lee & Tang, 2012). No obstante, el rápido avance en la metodología ha llevado a la diferenciación de técnicas de secuenciación de segunda gene-

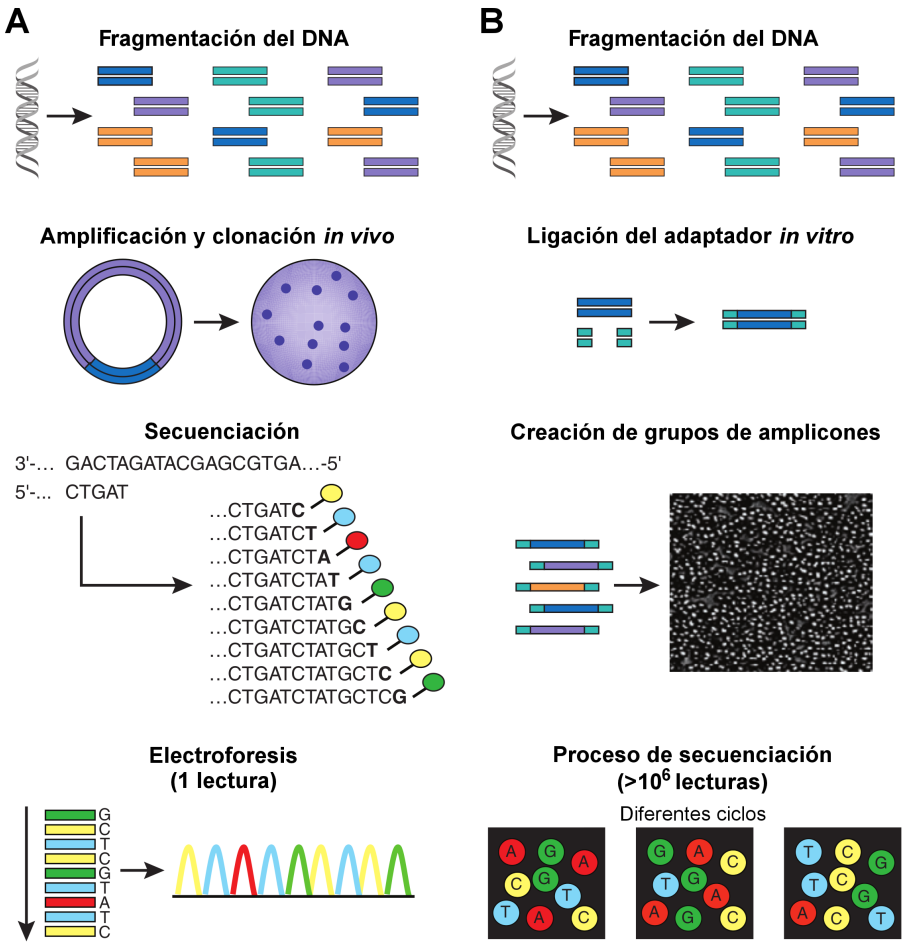


Figura I.1: **Comparación del flujo de trabajo de la secuenciación de primera y segunda generación.** (A) La secuenciación mediante el método Sanger conlleva una etapa de clonación *in vivo* y, tras el único ciclo de secuenciación que se realiza, se obtiene una lectura por cada electroforesis. (B) Las técnicas de SGS generan grupos de amplicones que llevan un adaptador con el que se realiza posteriormente la secuenciación. El proceso de secuenciación consta de varios ciclos, tras los que se consiguen más de 10^6 lecturas. Adaptada de Shendure & Ji, 2008.

ración (SGS), basadas en reacciones de PCR para obtener la muestra que se secuenciará, y de tercera generación (TGS), donde se secuencian moléculas únicas de DNA (Schadt *et al*, 2010).

La tecnología NGS tiene gran diversidad de aplicaciones, como ChIP-Seq, RNA-seq, transcriptómica, metagenómica, así como secuenciación de nuevos genomas y resecuenciación de otros. Sin embargo, presenta principalmente dos inconvenientes: la longitud de las lecturas (menor que la secuenciación convencional) y la precisión de la técnica. Por tanto, es necesario evaluar las ventajas e inconvenientes de cada técnica para así determinar cuál de ellas se adapta mejor a los distintos estudios que se pueden llevar a cabo (Metzker, 2010).

I.1.1 Técnicas de SGS

Hoy en día existen una serie de plataformas de SGS basadas en diversas aproximaciones, aunque todas ellas siguen un flujo de trabajo conceptualmente similar (figura I.1B). Primero se genera una genoteca mediante fragmentación aleatoria del DNA, en la que se produce la ligación, *in vitro*, de una secuencia llamada adaptador. Y después se lleva a cabo la creación de grupos de amplicones clonados sobre un área, que se secuencian mediante un proceso que consiste en sucesivos ciclos donde se alternan la reacción de extensión y la adquisición de imágenes de ese área (Shendure & Ji, 2008). Las principales plataformas de SGS se describen a continuación.

I.1.1.1 *Plataforma 454/Roche*

Ésta fue la primera plataforma de secuenciación masiva disponible en el mercado, y se fundamenta en el método de pirosecuenciación descrito por Ronaghi *et al* (1996), el cual no necesita electroforesis ni cebadores marcados, sino que se basa en la detección de los pirofosfatos liberados en la reacción de polimerización del DNA. La integración de ese proceso de pirosecuenciación con un método de PCR de emulsión (Dressman *et al*, 2003), llevados a cabo en placas de secuenciación con pequeños pocillos, dio lugar a la tecnología 454 (Margulies *et al*, 2005).

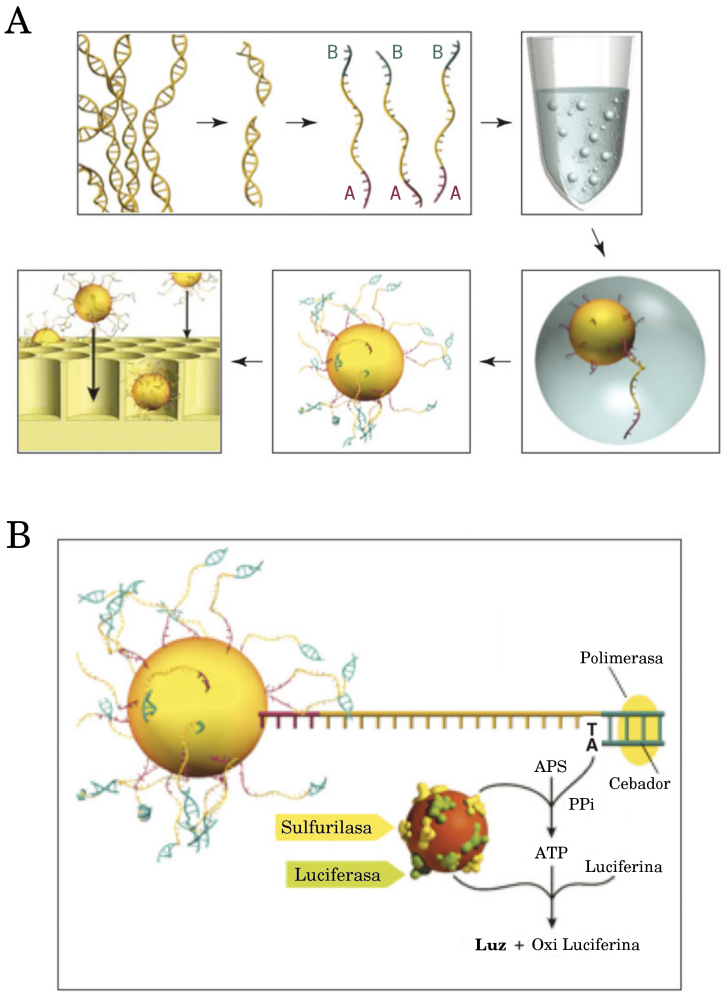


Figura I.2: **Proceso de secuenciación llevado a cabo por la tecnología 454 GS-FLX de Roche.** (A) Primero se fragmenta el DNA, se ligan adaptadores a los fragmentos y éstos se unen a unas bolitas en las que tiene lugar una PCR de emulsión. Cada bolita con los grupos de amplicones en su superficie se deposita en uno de los pocillos de la placa de secuenciación. (B) Reacción de pirosecuenciación que ocurre cuando la polimerasa incorpora nucleótidos, la cual termina con la emisión de una señal lumínica que se mide con una cámara acoplada al sistema. Adaptada de Mardis, 2008.

La primera etapa, común a las técnicas de SGS, consiste en la creación de una genoteca mediante fragmentación del DNA y ligación de dos adaptadores diferentes en ambos extremos (figura I.2A). Cada fragmento resultante se une a la superficie de una bolita de agarosa de 28 μm de diámetro en condiciones que favorecen la unión de un fragmento por bolita. En ese momento se lleva a cabo la PCR de emulsión con un cebador de secuencia complementaria a uno de los adaptores (denominado A en la figura I.2), generando así subgrupos de amplicones que son clones. Tras la emulsión, el DNA se desnaturaliza y las bolitas que portan los clones se depositan en cada uno de los cerca de dos millones de pocillos que contiene la placa de secuenciación. Posteriormente, se añaden bolitas de látex y magnéticas de 1 μm de diámetro que contienen las enzimas necesarias para la pirosecuenciación. En el proceso de pirosecuenciación (figura I.2B), la polimerasa va incorporando nucleótidos complementarios al amplicón que sirve de molde a partir de un cebador que hibrida con el otro adaptador ligado durante la creación de la genoteca (denominado B en la figura I.2). Tras la incorporación del nucleótido, éste libera un pirofosfato, que es convertido en ATP por una ATP sulfurilasa, y gracias a la luciferasa presente en la reacción, se emite una señal lumínica que es medida por una cámara de alta resolución acoplada a la placa. Después de cada medición, la placa se lava y comienza el ciclo con la incorporación del siguiente nucleótido. La luz emitida es directamente proporcional a la cantidad de un nucleótido determinado incorporado. Por este motivo, la mayor limitación de la tecnología 454 reside en los homopolímeros, puesto que la señal puede exceder el nivel de saturación del detector (Margulies *et al*, 2005; Rothberg & Leamon, 2008; Shendure & Ji, 2008; Kircher & Kelso, 2010).

El instrumento más utilizado para llevar a cabo la tecnología 454 es el secuenciador GS-FLX Titanium, que proporciona lecturas de unos 400 nt (Lee & Tang, 2012). No obstante, la técnica se está mejorando, y el nuevo instrumento en desarrollo permitirá producir lecturas de una longitud similar a la secuenciación Sanger (Forde & O'Toole, 2013). La principal ventaja de esta tecnología es la elevada longitud de las lecturas frente a las lecturas gene-

radas por otras plataformas, lo que facilita el ensamblaje de genomas completos (tema que trataremos más adelante, en el apartado I.1.3). Además, es una de las tecnologías que menor tasa de error comete. La mayor parte de estos errores son pequeños InDels (inserciones o deleciones), causados por la presencia de homopolímeros. La intensidad de la señal emitida por un homopolímero de 6 bases debería ser el doble que la de uno de 3 bases, sin embargo, la relación de la intensidad de esa señal no es lineal, y la precisión con la que se estima el homopolímero disminuye con la longitud de éste. Una solución a este problema es el aumento de la cobertura, aunque esto no resuelve los homopolímeros mayores de 10 nt. Al igual que ocurre en la secuenciación Sanger, la tasa de error aumenta con la posición en la secuencia. En el caso de la tecnología 454 esto es debido a la reducción de la eficiencia de la enzima o a la pérdida de ésta, lo que provoca una reducción en la intensidad de la señal (Kircher & Kelso, 2010; Metzker, 2010).

I.1.1.2 *Plataforma Solexa/Illumina*

Esta tecnología se fundamenta en una secuenciación por síntesis con dos características: nucleótidos de terminación reversibles, con cada una de las bases marcada con un fluorocromo distinto, y una DNA polimerasa capaz de incorporar ese tipo de nucleótidos. Primero se lleva a cabo la construcción de una genoteca mediante cualquier método que acabe con una mezcla de fragmentos flanqueados por adaptadores. Después se generan los subgrupos de amplicones con una estrategia diferente a la que usa la tecnología 454, mediante PCR en puente (Adessi *et al*, 2000). Los clones de cada subgrupo quedan inmobilizados en la superficie de un sustrato sólido, se linearizan mediante desnaturalización, y se secuencian a partir de un cebador universal que hibrida en uno de los adaptadores (figura I.3). La fluorescencia del nucleótido (de distinto color para cada base) incorporado en cada ciclo de secuenciación se mide con una cámara acoplada al sistema (Shendure & Ji, 2008; Ansorge, 2009; Voelkerding *et al*, 2009; Kircher & Kelso, 2010).

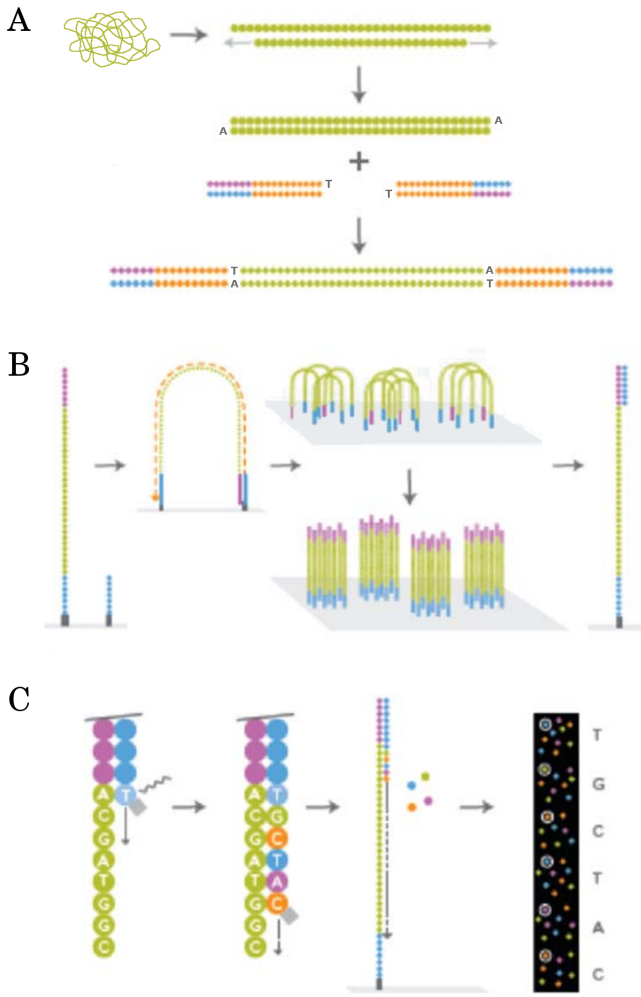


Figura I.3: **Proceso de secuenciación llevado a cabo por la plataforma Solexa de Illumina.** (A) Construcción de una genoteca mediante fragmentación del DNA y ligación de adaptadores a los fragmentos. (B) Creación de los grupos de amplicones mediante PCR en puente sobre una superficie sólida. Tras la amplificación, se desnaturaliza la muestra y se hibrida el cebador con el que tendrá lugar la secuenciación. (C) Secuenciación por síntesis en la que los nucleótidos están marcados con un fluorocromo en su extremo 3′ (cada base con un fluorocromo de diferente color) y la señal se mide con una cámara acoplada al sistema. Adaptada de Ansorge, 2009.

La longitud de las lecturas está limitada por múltiples factores que causan la pérdida de señal y el desfase entre la incorporación del nucleótido y la captación de la imagen. El principal tipo de error es la sustitución (más que las inserciones o deleciones), y los homopolímeros no son un problema tan acusado como en la tecnología 454 (Shendure & Ji, 2008). La longitud de las lecturas (entre 35-150 nt) es muy inferior a la de las lecturas obtenidas mediante la tecnología 454 (Lee & Tang, 2012), lo cual complica el ensamblaje de genomas completos (tema que trataremos más adelante, en el apartado I.1.3).

I.1.1.3 *Plataforma SOLiD/Applied Biosystems*

Ésta fue la tercera tecnología disponible en el mercado, y su principal diferencia con las anteriores plataformas es que se basa en la secuenciación mediante ligación. Tras la creación de la genoteca (fragmentos de DNA flanqueados por un adaptador), se generan los subgrupos de amplicones sobre la superficie de bolitas de 1 µm con el método de PCR de emulsión (Dressman *et al*, 2003). Despúes de la emulsión y la desnaturalización del DNA, las bolitas con los clones se depositan en un sustrato sólido de vidrio. La secuenciación comienza con la unión de un cebador al adaptador entre unas posiciones determinadas (cebador n en la figura I.4), y se realiza a través de una ligasa en vez de una polimerasa. La estrategia seguida por esta plataforma conlleva la unión de unas sondas de 8 bases (octámeros) marcadas con fluorocromos que contienen las dos primeras bases específicas y el resto de bases degeneradas. Las bases específicas de estos octámeros consisten en una de las 16 posibles combinaciones mostradas en la figura I.4A, y su identidad se relaciona con un fluorocromo distinto. En cada ciclo de secuenciación se produce la ligación de un octámero, se graba el color de la fluorescencia que emite y se corta químicamente entre las posiciones 5 y 6 (eliminando así el fluorocromo para el siguiente ciclo y generando el grupo 5'-PO_4 para la unión del siguiente octámero). Durante la extensión a partir de un cebador se realizan siete ciclos de ligación, lo que compone

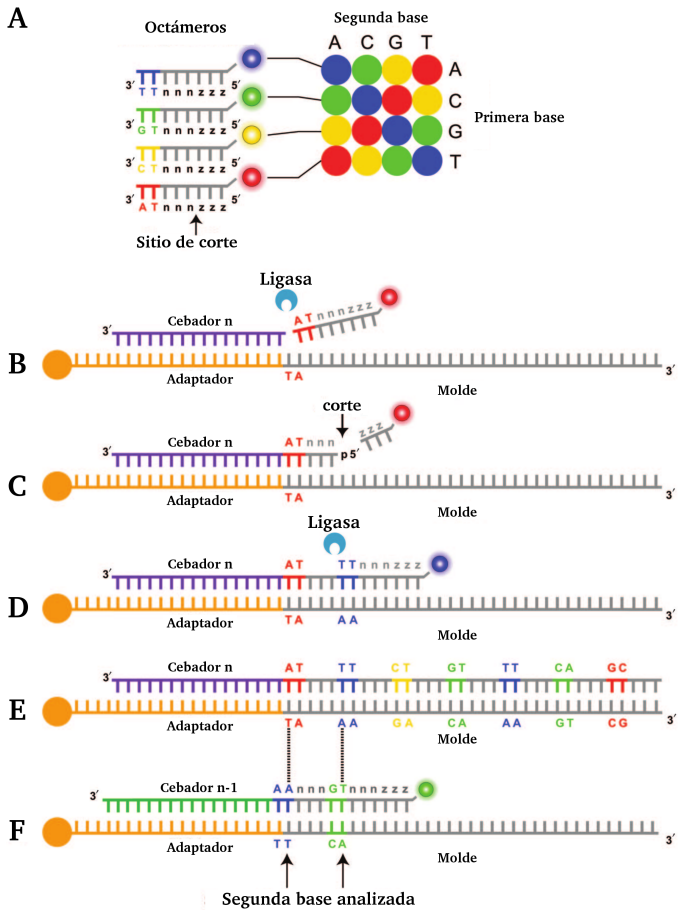


Figura I.4: **Proceso de secuenciación llevado a cabo por la plataforma SOLiD de Applied Biosystems.** (A) Código de color del fluorocromo que se encuentra unido al extremo 5′ de los octámeros usados en la secuenciación por ligación, cuya identidad viene determinada por las dos primeras bases del octámero (bases específicas). Se señala el sitio donde se produce el corte del octámero antes de pasar al siguiente ciclo de secuenciación. (B) Primer ciclo de secuenciación, en el que se incorporan todos los octámeros a la muestra y sólo hibrida el correspondiente a las dos bases siguientes al cebador usado (n). (C) Tras la ligación del octámero, se capta la señal del fluorocromo y se produce el corte entre las posiciones 5 y 6 del octámero. (D) Ligación del octámero durante el segundo ciclo de secuenciación. (E) La extensión del cebador n se completa tras siete ciclos de ligación, lo que constituye una ronda de secuenciación. Después se desnaturaliza la muestra y se liga un nuevo cebador al adaptador (en este caso el cebador n-1). (F) Segunda ronda de secuenciación, en la que se lleva a cabo la extensión del cebador n-1. En total se realizan cinco rondas de secuenciación, con distintos cebadores (n-2, n-3, n-4). Adaptada de Voelkerding *et al*, 2009.

una ronda de secuenciación. Tras ésta se produce una desnaturalización de la muestra para eliminar el cebador (n) y el producto extendido y que tenga lugar la segunda ronda utilizando otro cebador de la misma longitud (n-1). La secuenciación conlleva cinco rondas, en las que la posición donde hibridan los cebadores en el adaptador se ve incrementada un nucleótido en el extremo 3′ y disminuida en el 5′ (n, n-1, n-2, y así sucesivamente). Esto permite determinar la secuencia de DNA de la muestra mediante interpretación de los resultados de la ligación en cada caso teniendo en cuenta las 16 posibles combinaciones de las bases específicas (Shendure & Ji, 2008; Ansorge, 2009; Voelkerding *et al*, 2009; Kircher & Kelso, 2010).

Tomando como referencia la secuencia del adaptador, el sistema de codificación seguido por la plataforma SOLiD permite detectar errores de la máquina y aplicar una corrección para reducir la tasa de error media. En ausencia de una secuencia de referencia, cualquier conversión del color errónea en la lectura del fluorocromo causa que la secuencia aguas abajo del error sea incorrecta, y en este caso la tasa de error media es mayor que para la plataforma Illumina. Los errores más comunes en esta técnica son causados por la amplificación que se lleva a cabo durante el proceso, por la proximidad de las bolitas que portan los clones (que pueden crear lecturas falsas), y por problemas con los fluorocromos (como que se produzca un corte incompleto del octámero o que decaiga la intensidad de la señal). Otro problema añadido para la secuenciación de genomas completos es la longitud de las lecturas obtenidas (entre 25-50 nt), incluso menor que las obtenidas con Illumina (Kircher & Kelso, 2010; Lee & Tang, 2012).

I.1.2 Técnicas de TGS

La principal ventaja de esta nueva generación de técnicas es la eliminación de la etapa de amplificación de la muestra, la cual supone una de las causas de errores de las tecnologías de SGS. Esto se ve compensado por el sistema de detección de la fluorescencia que usan las técnicas de TGS, mu-

cho más sensible que el de las plataformas descritas en el apartado anterior (Schadt *et al*, 2010).

Entre las tecnologías consideradas de tercera generación encontramos las plataformas HeliScope, de Helicos Biosciences (Harris *et al*, 2008), y PacBio, de Pacific Biosciences (Eid *et al*, 2009), así como nuevas técnicas que se están desarrollando basadas en otras aproximaciones como la secuenciación de DNA a tiempo real usando transferencia de energía de resonancia de fluorescencia (FRET), de VisiGen Biotechnologies, el escaneo directo de DNA usando microscopía electrónica de transmisión (TEM), de Halcyon Molecular, o la secuenciación de DNA mediante el uso de nanoporos (Schadt *et al*, 2010; Yahaya, 2012; Mardis, 2013).

I.1.3 Ensamblaje de genomas completos mediante datos derivados de las técnicas de SGS

El elevado volumen de datos generado con la tecnología NGS ha presentado un desafío bioinformático para su almacenamiento, tratamiento y análisis. Los datos obtenidos mediante las imágenes durante el proceso de secuenciación tienen que ser convertidos en secuencias o lecturas a las que se les asigna un valor de calidad. Este parámetro se usa para eliminar las lecturas que presentan una calidad baja y recortar las bases que, dentro de una lectura, tengan baja calidad, para así mejorar la precisión del alineamiento (Voelkerding *et al*, 2009).

Una de las aplicaciones de la tecnología NGS es la secuenciación de genomas completos. Con el fin de reconstruir el genoma, las lecturas obtenidas tras la secuenciación deben ensamblarse, generando así secuencias continuas denominadas *contigs*. La limitación de este proceso reside en los métodos de secuenciación, los cuales producen lecturas relativamente cortas comparadas con el tamaño total del genoma. Este problema ha forzado a que la mayoría de ensambladores desarrollados utilicen varias aproximaciones heurísticas, y, dependiendo de la estrategia en la que se basen, se clasifi-

can en una de las tres categorías principales: aproximación *greedy*, aproximación OLC (*overlap-layout-consensus*) y aproximación del gráfico *de Brujin* (Miller *et al*, 2010; Lee & Tang, 2012).

El ensamblaje de los genomas se puede realizar mediante programas que usan un genoma de referencia (*mapping* o ensamblaje por referencia), como es el caso de Bowtie, SOAP o Mosaik, o mediante ensambladores que prescinden de éste (ensamblaje *de novo*), entre los que encontramos Celera, Mira, Velvet o Newbler. El ensamblaje *de novo* se ve facilitado por las lecturas tan largas obtenidas con la tecnología 454 comparadas con aquéllas obtenidas a partir de las tecnologías Illumina y SOLiD. Este motivo, junto a la baja tasa de error de identificación de bases, hacen que la tecnología 454 sea la más usada en la secuenciación de genomas completos y en la identificación de polimorfismo de nucleótidos simples (SNPs; Kircher & Kelso, 2010). No obstante, para conseguir un ensamblaje *de novo* óptimo, lo más efectivo es unir la información obtenida con tecnologías complementarias, como 454 e Illumina, con las que se produce un ensamblaje *de novo* con una calidad similar al conseguido mediante el método Sanger, y donde los errores de homopolímeros son solventados gracias a la elevada cobertura proporcionada por Illumina (Forde & O'Toole, 2013).

En este trabajo hemos llevado a cabo la secuenciación de un genoma bacteriano, con la tecnología 454, y un ensamblaje *de novo* mediante el programa Newbler, que utiliza una aproximación OLC. Por este motivo, las explicaremos en mayor detalle.

El problema que conlleva un ensamblaje *de novo* con lecturas cortas es la presencia de regiones repetidas en el genoma, que provocan la formación de numerosos *contigs* a los que se les debe asignar una posición y orientación relativa para así generar estructuras de mayor longitud, denominadas andamiajes (*scaffolds*). Este proceso se ha visto facilitado por el desarrollo de un método basado en la construcción de genotecas *paired end* (Ng *et al*, 2006; Jarvie & Harkins, 2008). La información que proporciona esta genoteca permite determinar *contigs* que deben ser contiguos pero estar separados por

una región de secuencia repetida. La construcción de la genoteca *paired end* se realiza de manera paralela a la genoteca que se usará en la pirosecuenciación directa (figura I.5). Básicamente, la técnica consiste en crear fragmentos de entre 2-4 kb a partir del DNA genómico, ligarles en los extremos una secuencia conocida llamada conector, circularizar la molécula a través del conector, y llevar a cabo una rotura aleatoria de esas moléculas. Así se generan fragmentos de entre 500-800 pb que siguen el proceso de secuenciación mediante la tecnología 454 que hemos explicado anteriormente (apartado I.1.1.1). Hay que destacar que la rotura aleatoria de las moléculas circularizadas producirá dos tipos de fragmentos: los que contengan el conector y puedan ser usados para la construcción de los andamiajes (ya que las secuencias que se encuentran aguas arriba y abajo del conector físicamente distan 2-4 kb dentro del genoma), y los que no contengan el conector o éste se encuentre en un extremo de la lectura (que se incluirán en el total de lecturas obtenidas, pero no aportarán información sobre la posición relativa de dos *contigs*). Inicialmente se desarrolló esta genoteca *paired end*, de 3 kb, aunque hoy en día también se realizan genotecas *paired end* de 8 kb y 20 kb (Lee & Tang, 2012).

El problema de los ensamblajes *de novo*, causado por las regiones repetidas en los genomas, se podría solventar con el aumento de la longitud de las lecturas obtenidas con la tecnología 454 (Jiang *et al*, 2012). Aun así, las técnicas de TGS tienen el potencial para acabar con las limitaciones que encontramos en las técnicas de SGS respecto a la calidad de los ensamblajes, puesto que tanto la precisión del método como la longitud de las lecturas son mayores, eliminando la necesidad de usar genotecas *paired end* para construir el andamiaje (Schadt *et al*, 2010).

La formación de los andamiajes conlleva la introducción de "Ns" en la secuencia, con una longitud entre los *contigs* unidos que es estimada a partir de las lecturas pareadas (obtenidas mediante la genoteca *paired end*). Por tanto, para la consecución de la secuencia completa del genoma se requiere llevar a cabo un proceso denominado *finishing*, con el que se rellenan todos los huecos entre *contigs* generados por el ensamblador. Los métodos

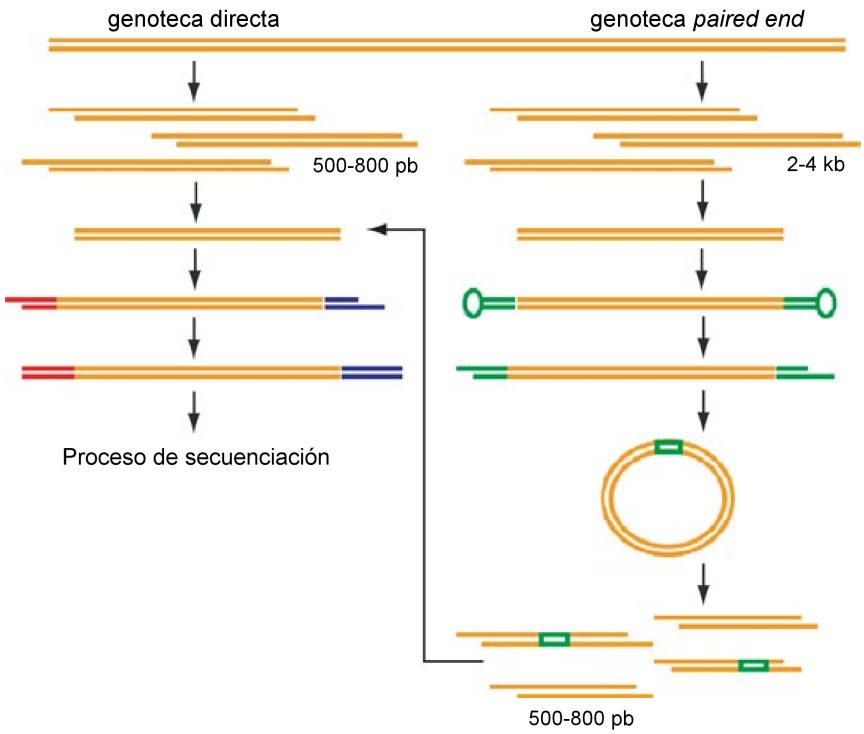


Figura I.5: **Proceso de construcción de los dos tipos de genotecas usadas en la tecnología 454 de Roche.** A la izquierda se muestran las etapas de la construcción de una genoteca directa, mientras que a la derecha se indican las de una genoteca *paired end*. En ambas se produce inicialmente una fragmentación del DNA, aunque difieren en el tamaño de los fragmentos resultantes. Para generar la genoteca *paired end* se liga un conector que permite la circularización de la molécula, tras lo que se produce una rotura del DNA aleatoria. Los fragmentos obtenidos se procesan del mismo modo que los fragmentos de la genoteca directa: en los extremos se ligan unos adaptadores con los que se lleva a cabo el proceso de secuenciación (descrito en el apartado I.1.1.1). Adaptada de Wiley *et al*, 2009.

experimentales más utilizados para realizar esta tarea se fundamentan en reacciones de PCR, mediante el diseño de cebadores que flanquean la zona de "Ns". Sin embargo, se están desarrollando aproximaciones computacionales que se basan en el uso de las lecturas pareadas para llevar a cabo un ensamblaje local en esas zonas conflictivas (Lee & Tang, 2012).

I.2 GENOMAS BACTERIANOS

El avance y abaratamiento de las nuevas tecnologías de secuenciación ha permitido incrementar el número de proyectos llevados a cabo para secuenciar genomas tanto de bacterias y arqueas como de eucariotas. En la figura I.6A se muestra la aparición de las plataformas de SGS descritas anteriormente y la progresión del número de genomas completamente secuenciados y publicados. La base de datos GOLD (*Genomes Online Database*; www.genomesonline.org), a principios de mayo de 2013, presentaba un total de 4.327 genomas (completos y borradores permanentes), de los cuales 3.957 eran bacterianos, 187 de arqueas y 183 de eucariotas (figura I.6B). En menos de un año se ha completado el proyecto de secuenciación de 1.110 nuevos genomas bacterianos (en octubre de 2012 se encontraban secuenciadas 2.847 bacterias) y los proyectos clasificados como "en progreso" han duplicado su número (8.559 en mayo de 2013 frente a los 4.226 en octubre de 2012; Forde & O'Toole, 2013).

La distribución del total de proyectos de secuenciación de genomas bacterianos presentes en la base de datos GOLD a principios de mayo de 2013, representada por filos bacterianos, pone de manifiesto que las proteobacterias son el filo con el que se están llevando a cabo más proyectos de secuenciación (figura I.6B). En concreto, las α-proteobacterias presentan cerca de 600 genomas completamente secuenciados, y es una de las clases bacterianas más estudiadas. Las especies pertenecientes a las α-proteobacterias presentan unas características genómicas muy heterogéneas y gran versatilidad para adaptarse a diferentes hábitats (figura I.7). Por este motivo, dicha clase constituye un excelente sistema modelo para estudiar la evolución de los genomas bacterianos y los rasgos genómicos implicados en la adaptación al medio ambiente (Pini *et al*, 2011).

El genoma de los organismos evoluciona por adquisición de nuevas secuencias y por reordenación de las ya existentes, generando así, además, diversidad genética. En bacterias, los elementos extracromosómicos mueven el material genético mediante transferencia horizontal, ya sea a través de

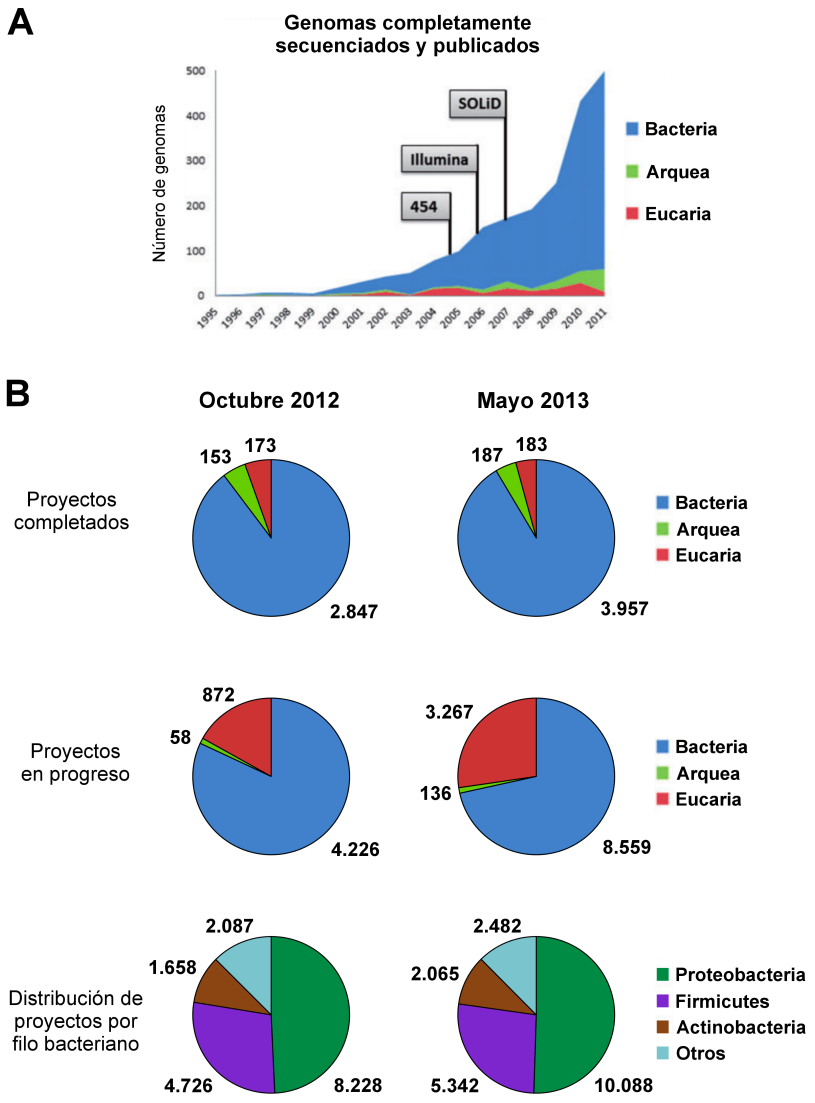


Figura I.6: **Comparación del número de genomas secuenciados en octubre de 2012 y mayo de 2013.** (A) Gráfica que muestra el número de genomas (de los tres reinos, bacteria, arquea y eucaria) completamente secuenciados cada año desde 1995. Se indica el año de aparición de las técnicas de SGS comentadas en el apartado I.1.1. (B) Número de proyectos completados y en progreso de cada reino hasta el 12 de octubre de 2012 (a la izquierda) y el 2 de mayo de 2013 (a la derecha). Debajo se muestra la distribución de todos los proyectos por filo bacteriano hasta esas mismas fechas. Adaptada de Forde & O'Toole, 2013.

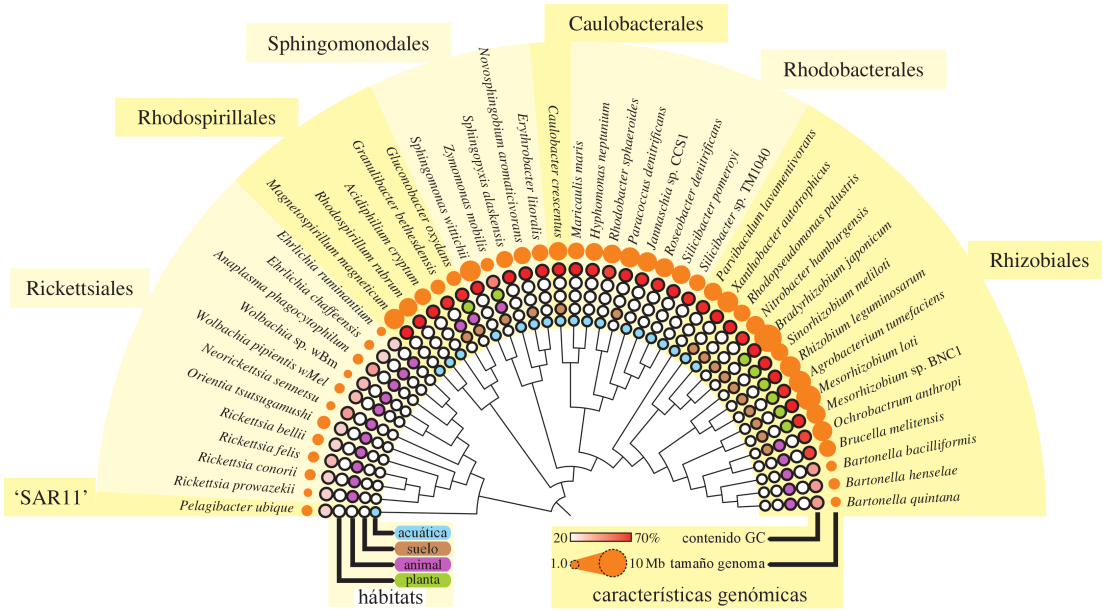


Figura I.7: **Relación filogenética de las especies pertenecientes a las α-proteobacterias.** Para cada especie se muestra el hábitat donde se encuentra y el contenido GC y el tamaño de su genoma. Adaptada de Ettema & Andersson, 2009.

plásmidos, por conjugación, o mediante fagos, por infección. Sin embargo, en eucariotas, sólo algunos virus pueden transferir información genética entre individuos durante el ciclo infectivo (Lewin, 2007).

La diferencia de tamaño en el genoma de las bacterias pertenecientes a las α-proteobacterias refleja la adaptación que han sufrido a los distintos ambientes, con la pérdida o ganancia de cientos de genes. Las bacterias adaptadas al citoplasma de células eucarióticas (con un ambiente relativamente estático) han evolucionado hacia genomas pequeños y con bajo contenido GC, mientras que las bacterias de vida libre adaptadas al suelo (con un ambiente nutricionalmente más variable) presentan genomas mucho más grandes y con un alto contenido GC (figura I.7). La reducción del tamaño de los genomas puede ocurrir por pérdida de genes innecesarios (como pseudogenes o elementos genéticos móviles), por pérdida de genes

involucrados en procesos que la bacteria ya no lleva a cabo (por ejemplo, biosíntesis de metabolitos, los cuales importa del citoplasma eucariótico) o por transferencia génica hacia el genoma del hospedador (figura I.8A). Por otro lado, el tamaño de los genomas bacterianos puede aumentar por adquisición de replicones accesorios (cuyo tamaño y contenido génico varía

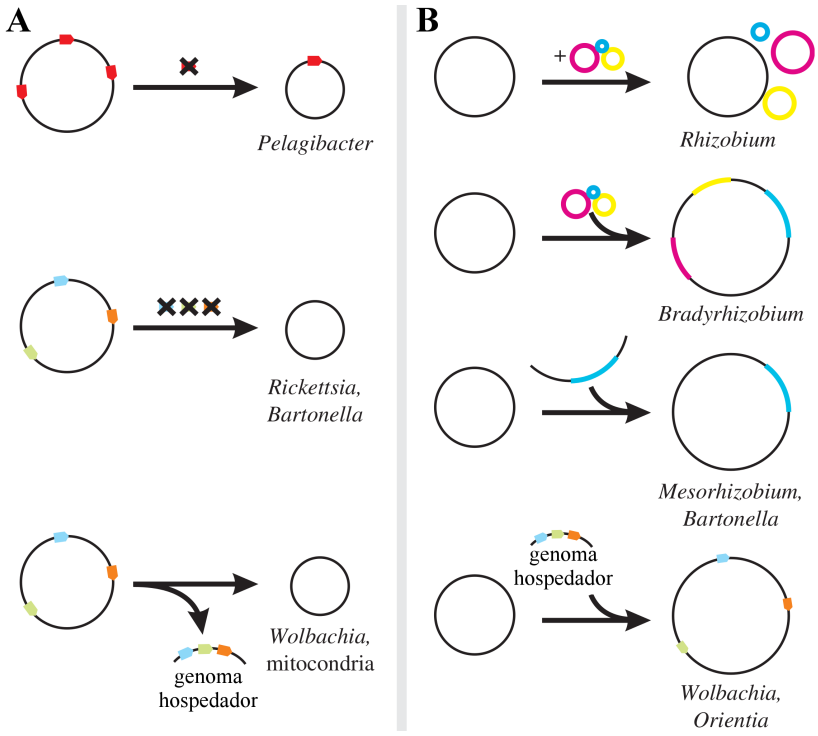


Figura I.8: **Mecanismos de reducción y expansión de los genomas de las α-proteobacterias.** A modo de ejemplo se indica la especie bacteriana en la que ha ocurrido cada mecanismo. (A) Representación de los mecanismos que llevan a la reducción de los genomas: pérdida de genes innecesarios, pérdida de genes involucrados en procesos que la bacteria ya no lleva a cabo y transferencia de genes al hospedador. (B) Representación de los mecanismos que llevan a la expansión de los genomas: adquisición de replicones accesorios, integración al cromosoma principal de replicones accesorios, transferencia de islas genómicas y adquisición de genes del hospedador. Adaptada de Ettema & Andersson, 2009.

considerablemente entre cepas, facilitando así la especificidad de nichos), por integración al cromosoma principal de replicones accesorios (evitando así la pérdida de éstos en la población), por transferencia de islas genómicas entre especies bacterianas o por adquisición de genes desde el genoma del hospedador (figura I.8B; Ettema & Andersson, 2009).

I.3 BACTERIAS DEL ORDEN RHIZOBIALES

Dentro de la clase α-proteobacterias (figura I.7) encontramos bacterias que viven en ambientes acuáticos (como *Caulobacter sp.* y varias especies del orden Rhodobacterales), bacterias que son patógenos de plantas (*Agrobacterium*), bacterias que son parásitos intracelulares de animales (como *Rickettsia* y *Brucella*), y bacterias que viven en el suelo (como muchas de las especies del orden Rhizobiales). Esta clase bacteriana tiene un interés especial debido a que se cree que son el grupo ancestral de las mitocondrias. De hecho, el orden Rickettsiales es el que se cita más a menudo como grupo bacteriano del que emergen las mitocondrias (Williams *et al*, 2007). Un tipo de bacterias muy importante a nivel agronómico es el formado por aquéllas que son capaces de establecer simbiosis mutualistas específicas con plantas leguminosas. En este grupo se enmarcan muchas de las especies pertenecientes al orden Rhizobiales, a las que colectivamente se les denomina rizobios. Entre ellas encontramos bacterias de los géneros *Mesorhizobium*, *Bradyrhizobium*, *Rhizobium*, *Sinorhizobium* (o *Ensifer*), *Azorhizobium* y *Methylobacterium* (Pini *et al*, 2011).

La importancia de los rizobios reside en su capacidad de fijar nitrógeno atmosférico, sin embargo, este proceso requiere gran cantidad de energía. Por otro lado, el nitrógeno en forma molecular N_2 presente en la atmósfera no puede ser incorporado por los vegetales, siendo uno de los factores limitantes de su crecimiento. El gran interés biológico de la relación simbiótica rizobio-leguminosa se basa en la elevada eficiencia de fijación de nitrógeno que se obtiene con esa interacción, puesto que la actividad fotosintética de la planta es una fuente de energía que permite fijar del orden de 10 a 20

veces más nitrógeno que en el caso de los rizobios fijadores de vida libre (Vance, 1998). Además, la amplia dispersión geográfica de estas plantas y su considerable empleo en alimentación, así como la protección de los cultivos y del medio ambiente al evitar el uso indiscriminado de fertilizantes industriales, hacen de la relación simbiótica un tema de interés a gran escala.

Durante la simbiosis, el rizobio induce la formación de nódulos en raíces de leguminosas, donde se establece en el interior de las células vegetales y se diferencia morfológicamente a bacteroide, la forma especializada de la bacteria que lleva a cabo la fijación del nitrógeno atmosférico y su transformación en amonio (una molécula asimilable por la planta). Este proceso tiene lugar bajo la expresión coordinada de diversos genes relacionados con la simbiosis presentes en ambos organismos. Los flavonoides (compuestos aromáticos secretados por la leguminosa) inician la producción de factores de nodulación por parte del rizobio. Estos factores de nodulación, a su vez, activan múltiples respuestas en la planta que la preparan para ser invadida por la bacteria. La formación de nódulos fijadores de nitrógeno productivos conlleva una asociación específica rizobio-leguminosa. Así, cada rizobio es capaz de establecer simbiosis con una o muy pocas especies vegetales, normalmente muy relacionadas filogenéticamente. Uno de los sistemas modelo para estudiar esta interacción es el formado por la bacteria *Sinorhizobium meliloti* y algunas leguminosas del género *Medicago*, incluyendo *M. sativa* (alfalfa; Jones *et al*, 2007).

I.3.1 *Sinorhizobium meliloti*

La primera bacteria del género *Sinorhizobium* completamente secuenciada fue *S. meliloti* 1021 (Galibert *et al*, 2001), elegida como cepa modelo. Como hemos comentado, la secuenciación de genomas bacterianos ha aumentado exponencialmente en los últimos años (figura I.7), y hasta mayo de 2013 encontramos seis genomas de cepas de *S. meliloti* totalmente secuenciados además del de 1021. Éstas son SM11 (Schneiker-Bekel *et al*, 2011),

AK83 y BL225C (Galardini *et al*, 2011b), GR4 (Martínez-Abarca *et al*, 2013), Rm41 (Weidner *et al*, 2013) y 2011 (www.ncbi.nlm.nih.gov/bioproject/PRJNA193772). El genoma de la cepa CCNWSX0020 (Li *et al*, 2012b) aparece como borrador en la base de datos del NCBI (www.ncbi.nlm.nih.gov/genome/1004), en la que también encontramos nuevos proyectos de secuenciación de varias cepas de *S. meliloti* (figura I.9).

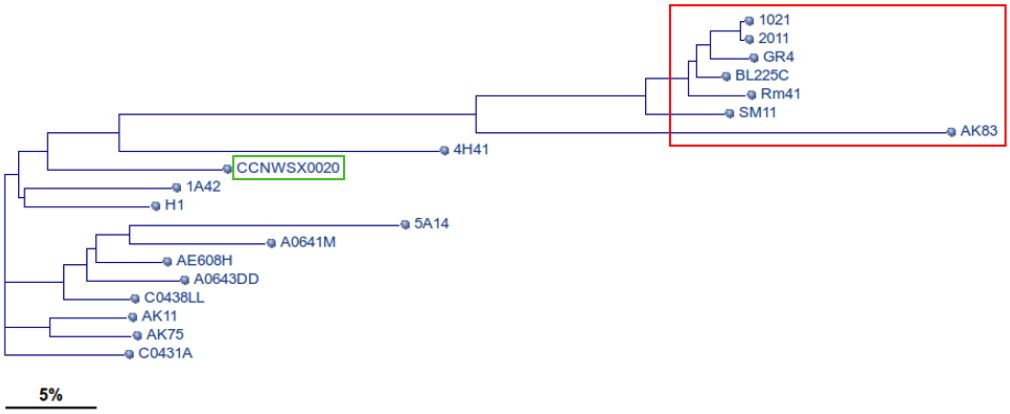


Figura I.9: **Dendograma de la especie *S. meliloti*.** Enmarcadas en rojo y en verde se muestran las cepas que han sido totalmente secuenciadas y que se encuentran como borrador de genoma respectivamente. El resto de cepas presentan un proyecto de secuenciación en progreso. Imagen obtenida el 2 de mayo de 2013 de la base de datos del NCBI (www.ncbi.nlm.nih.gov/genome/1004).

En general, el genoma de los rizobios es multipartito (figura I.8B), compuesto habitualmente por un cromosoma y varios plásmidos (hasta 11 en *R. leguminosarum*) de diferente tamaño (desde 45 kb hasta alrededor de 2'5 Mb, siendo los plásmidos de pequeño tamaño poco comunes en rizobios). Todas las cepas de *S. meliloti* secuenciadas portan al menos dos replicones además del cromosoma, denominados megaplásmidos o plásmidos simbióticos por contener genes implicados en la simbiosis. En la cepa 1021, el megaplásmido de menor tamaño se denomina pSymA, y porta los genes necesarios para la nodulación (genes *nod*) y la fijación de nitrógeno (genes *fix* y *nif*). El megaplásmido de mayor tamaño, denominado pSymB, se caracteriza por

contener grupos de genes involucrados en la síntesis de exopolisacáridos (requeridos para una nodulación efectiva), aunque también presenta genes asociados a otros procesos esenciales, como los genes *dctA*, *dctC* y *dctBD* del sistema de transporte de dicarboxilatos necesario para la fijación de nitrógeno o genes implicados en el catabolismo y transporte de azúcares. Este megaplásmido es considerado un cromosoma verdadero por contener regiones necesarias para la viabilidad de la célula, mientras que pSymA es considerado un plásmido puesto que se pueden obtener células curadas de él (Barloy-Hubler & Jebbar, 2009; López-Guerrero *et al*, 2012).

Además de estos plásmidos simbióticos, en el genoma de *S. meliloti* podemos encontrar plásmidos dispensables para la simbiosis, denominados plásmidos accesorios o crípticos, que aparecen en bajo número de copias. La inestabilidad de este tipo de plásmidos se ve contrarrestada por la presencia de genes necesarios para el crecimiento o la supervivencia, lo cual hace que se mantengan en la población. En concreto, la cepa *S. meliloti* GR4 contiene dos plásmidos crípticos: pRmeGR4a, de 115 MDa, y pRmeGR4b, de 140 MDa (Toro & Olivares, 1986). Éste último presenta una región, denominada nfe (*nodule formation eficiency*), implicada en la competitividad para la nodulación (Toro & Olivares, 1986; Sanjuan & Olivares, 1989; Soto *et al*, 1993; García-Rodríguez & Toro, 2000). Por tanto, la importancia de esta bacteria, que ha sido estudiada desde hace más de 30 años (Casadesús & Olivares, 1979), radica en el elevado grado de competitividad que presenta cuando se compara con otras cepas del mismo género.

I.4 ORGANIZACIÓN GENÓMICA

Los procariotas tienen genomas compactos, y en ellos los procesos celulares que interactúan directa o indirectamente con el DNA afectan y dan forma a la estructura genómica. El cromosoma de estos organismos es relativamente uniforme en términos de densidad génica y, en rizobios, presenta un mayor grado de conservación que los plásmidos tanto a nivel de secuencia como de sintenia (conservación del orden de los genes). A diferencia

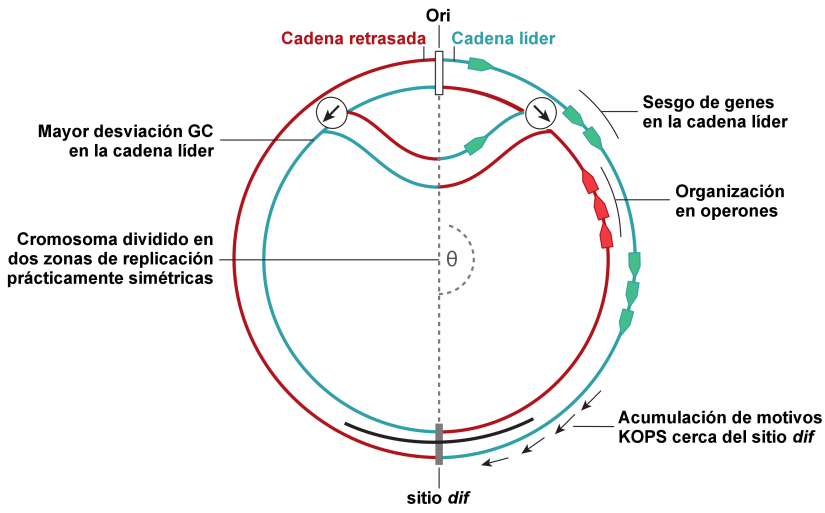


Figura I.10: **Elementos de la organización genómica en procariotas.** Se indican el origen (ori) y término (sitio *dif*) de la replicación, así como la cadena líder y retrasada. También se muestran algunas de las características de los genomas bacterianos. Adaptada de Rocha, 2008.

de los genomas eucarióticos (que tienen una típica organización de exón-intrón), en procariotas los genes generalmente se cotranscriben en unidades llamadas operones (figura I.10). Estos operones codifican genes que normalmente interactúan físicamente, y pueden agruparse en superoperones (el superoperón ribosomal es el más grande, formado por unos 50 genes). Es interesante destacar el sesgo que existe en la presencia de genes en ambas cadenas (encontrándose un mayor número de genes en la cadena líder), así como en el contenido de guaninas y citosinas (en la mayoría de genomas bacterianos la cadena líder tiende a ser rica en G y la cadena retrasada en C; Rocha, 2008; Koonin, 2009).

Durante la replicación, las translocasas (como FtsK en *E. coli* o SpoIIIE en *B. subtilis*) dirigen el DNA hacia las células hijas gracias al reconocimiento de motivos que apuntan al sitio de terminación de la replicación (sitio *dif*). En *E. coli*, esos motivos se denominan KOPS (FtsK *Orientating Polar Sequences*), y su densidad es mayor en la cadena líder y aumenta cerca del sitio *dif* (figura I.10; Bigot *et al*, 2005). La localización de estos motivos, así

como el análisis de la desviación GC (basada en el sesgo de G y C entre ambas cadenas; Lobry, 1996), permiten la identificación del origen y término de replicación en los genomas bacterianos (Sernova & Gelfand, 2008). El patrón de distribución mostrado por estos parámetros permite dilucidar el origen y término de replicación en el cromosoma mejor que en los plásmidos, cuya replicación se rige por el gen *repC* codificado por el operón *repABC* (Cevallos *et al*, 2008).

I.5 MOVILOMA

Los elementos genéticos móviles (MGEs) son uno de los factores que más influyen en la generación de diversidad genética, y, junto al DNA que movilizan, constituyen el llamado moviloma de un genoma. Desde el descubrimiento de los elementos transponibles a finales de los años 40 por Barbara McClintock, se han identificado muchos elementos móviles tanto en eucariotas como en procariotas (Hernandez-Lucas *et al*, 2006). Los MGEs se definen como una región de DNA de longitud variable que tiene la capacidad de moverse entre genomas o dentro del mismo genoma, y que portan la información necesaria para transferirse y recombinarse con el genoma hospedador. Los MGEs se pueden dividir en varias categorías en base a su mecanismo de movilidad y al tipo de DNA que los compone. Las principales categorías son: elementos transponibles (TEs), que incluye retrotransposones (de tipo LTR, no-LTR y retovirus), transposones de DNA y secuencias de inserción (ISs); plásmidos (conjugativos y no conjugativos); bacteriófagos, como profagos y fagos filamentosos; y moléculas con autoescisión, donde se incluyen los intrones del grupo I y del grupo II (Siefert, 2009; Toussaint & Chandler, 2012).

Los retrotransposones no-LTR (*non-long-terminal-repeat*) son una de las clases más abundantes de TEs. Tanto la organización como el mecanismo de transposición de estos retroelementos no-LTR están muy estrechamente relacionados con los descritos para los intrones del grupo II presentes en bacterias y orgánulos (Malik *et al*, 1999). Por esta razón, se han propuesto

a los intrones del grupo II como los antecesores evolutivos de retrotransposones no-LTR. Ambos comparten dos características fundamentales. La primera es que se replican obligatoriamente a través de intermediarios de RNA usando mecanismos dependientes de reverso transcriptasas. La segunda es que no parece que confieran una obvia ventaja selectiva a sus hospedadores, sino que son deletéreos o, en el mejor de los casos, neutrales. En un sentido evolutivo, a menudo se piensa en los retroelementos como parásitos genéticos (*selfish* DNA), y por tanto, la movilidad puede ser una consecuencia necesaria de su naturaleza parasítica (Medhekar & Miller, 2007).

Los intrones forman parte de la organización genómica de eucariotas (intrones espliceosómicos), aunque también se encuentran en procariotas (intrones del grupo I y del grupo II), e incluso se han descubierto un grupo de intrones específicos de arqueas. Se ha propuesto que los intrones del grupo II son los ancestros de los intrones espliceosómicos debido a la similitud que presentan el mecanismo de escisión y la estructura formada tras la escisión de ambos tipos de intrones (Koonin, 2006). Esto, junto a diversas aplicaciones biotecnológicas que comentaremos más adelante (apartado I.10), ha suscitado el interés por el estudio de los intrones del grupo II.

En los siguientes apartados nos centraremos en explicar dos de los elementos genéticos móviles más expandidos por los genomas bacterianos y que son objeto de estudio de esta Tesis Doctoral, las secuencias de inserción y los intrones del grupo II.

I.6 SECUENCIAS DE INSERCIÓN

Las secuencias de inserción (ISs) son los elementos transponibles autónomos más simples que se encuentran en los genomas bacterianos. Las ISs tienen un tamaño de entre 0'7 y 3'5 kb, y contienen, al menos, un ORF que codifica una transposasa (figura I.11). En sus extremos, la mayoría de ISs presentan repeticiones invertidas (IRs) terminales imperfectas de entre 10

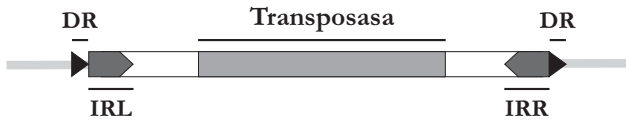


Figura I.11: **Estructura de las secuencias de inserción (ISs).** Se indican los distintos elementos de las ISs: la transposasa que codifica, las repeticiones invertidas terminales izquierda (IRL) y derecha (IRR) y las repeticiones directas (DR).

pb y 40 pb de longitud. La movilidad de las ISs ocurre por un proceso llamado transposición, en el cual se requiere la transposasa codificada por ellas además de otras proteínas del hospedador. Durante este proceso, la transposasa se une al DNA diana y a las IRs, permitiendo la inserción de la IS. Tras la transposición, las ISs generan repeticiones directas (DRs) de la secuencia diana, cuya longitud varía entre 2 pb y 14 pb y es específica para cada familia de ISs (Mahillon & Chandler, 1998; Cerveau *et al*, 2011).

Se han descrito más de 4.000 ISs, ampliamente distribuidas por los genomas bacterianos y también encontradas en arqueas (Siguier *et al*, 2012). Los plásmidos bacterianos con menos de ~20 kb generalmente no presentan ISs, y, dependiendo de su tamaño, el porcentaje de ISs puede variar desde un 5 % hasta un 40 % en el caso del megaplásmido pW100 de *Shigella flexneri*. Sin embargo, en los cromosomas bacterianos la densidad de ISs está generalmente por debajo del 3 % salvo en unos pocos casos (Siguier *et al*, 2006a). Las ISs son un factor importante en la variabilidad genética puesto que pueden promover reordenaciones genómicas y generar mutaciones por su inserción en genes o secuencias reguladoras. Ésta es una de las razones por las que se asume que la transposición normalmente se encuentra regulada en bacterias, sobre todo en las de vida libre (Gil & Latorre, 2012).

Las ISs se clasifican en más de 25 familias en base a los rasgos que comparten. Una familia se define como un grupo de ISs con transposasas relacionadas, IRs similares, conservación del sitio catalítico y conservación de la organización (Siguier *et al*, 2012). En el género *Sinorhizobium* se han descu-

bierto más de 30 ISs diferentes, pertenecientes a varias familias. Entre ellas se encuentra la familia IS*66*, cuyos miembros se limitan al filo Proteobacteria (Gourbeyre *et al*, 2010). En este filo bacteriano las ISs son particularmente abundantes en los plásmidos simbióticos y en las islas cromosómicas simbióticas (Lozano *et al*, 2010).

I.7 INTRONES DEL GRUPO II

Los intrones del grupo II son RNAs catalíticos capaces de autoescindirse de una molécula precursora. Estos elementos se identificaron inicialmente en orgánulos de plantas, hongos y otros eucariotas inferiores (Michel *et al*, 1989), y posteriormente se describieron en bacterias y algunos genomas de arqueas (Ferat & Michel, 1993; Toro, 2003; Lambowitz & Zimmerly, 2004). Los intrones del grupo II mejor estudiados son Ll.ltrB, encontrado en *Lactococcus lactis* (Mills *et al*, 1996; Shearman *et al*, 1996), y RmInt1, descubierto en *S. meliloti* (Martínez-Abarca *et al*, 1998).

Los intrones del grupo II tienen una longitud de entre 100 nt y 3 kb. Generalmente constan de dos componentes: un RNA, denominado ribozima, con una estructura secundaria dispuesta en seis dominios de diferente longitud (dominios DI-DVI), y una proteína, denominada IEP (*Intron Encoded Protein*), con actividad madurasa y reverso transcriptasa (RT); aunque también se pueden encontrar intrones del grupo II carentes de proteína (Lehmann & Schmidt, 2003). De acuerdo a diferencias en la estructura del RNA del intrón y a estudios filogenéticos de la ribozima y la proteína se han definido varias subclases de intrones del grupo II.

I.7.1 Clasificación, distribución y evolución de los intrones del grupo II

Aunque todos los intrones del grupo II comparten la estructura básica dispuesta en seis dominios, se pueden hacer divisiones en base a diferencias estructurales de la ribozima (características detalladas en el siguiente

apartado, I.7.2). Inicialmente, cuando no se conocían intrones del grupo II en bacterias y arqueas, se definieron dos subclases, intrones IIA y IIB, que, a su vez, se dividían en IIA1, IIA2, IIB1 y IIB2 (Michel *et al*, 1989). Posteriormente, y con el descubrimiento de nuevos intrones, se han clasificado en tres subclases principales, IIA, IIB y IIC (Toor *et al*, 2001), de las cuales la IIB se diferencia en cinco grupos (IIB1-IIB5; Toro, 2003).

Por otro lado, también se ha llevado a cabo una clasificación de los intrones del grupo II en base a estudios filogenéticos de la proteína que codifican. Como ocurre con la ribozima, el aumento del número de intrones descritos conlleva la aparición de nuevas clases. Primeramente se definieron seis clases de intrones: de tipo mitocondriales (ML), de tipo cloroplastidiales (CL; que se dividen en CL1 y CL2), y cuatro clases de intrones bacterianos (A-D; Toor *et al*, 2001). Más tarde se añadieron las clases bacterianas E (Toro *et al*, 2002) y F (Simon *et al*, 2008), aunque ésta última se ha puesto en duda con la reciente definición de cuatro subgrupos nuevos (G1-G4; Toro & Martínez-Abarca, 2013). Las ribozimas de los intrones del grupo II se unen específicamente a la proteína que codifican, formando un complejo ribonucleoprotéico necesario para la movilidad de estos elementos. Por este motivo, se piensa que existe una coevolución entre el RNA del intrón y la IEP (hipótesis del retroelemento ancestral), generándose así linajes filogenéticos (Toor *et al*, 2001; Toro, 2003; Toro *et al*, 2007). Cada grupo de intrones definido en base a la IEP se engloba dentro de una subclase de la ribozima (figura I.12). Así, los intrones ML se asocian a la subclase IIA, los intrones bacterianos C a la IIC, los CL a la IIB1 y IIB2, los intrones bacterianos D a la subclase IIB3, los B a la IIB4, los E a la IIB5, y el resto de intrones bacterianos (A y G1-G4) se consideran de la subclase IIB a pesar de no presentar todas las características de los miembros incluidos en ella. Hay que destacar que en los genomas bacterianos se han encontrado intrones del grupo II de todos los linajes, mientras que mitocondrias y cloroplastos sólo contienen intrones de las clases ML y CL (figura I12; Toro *et al*, 2007; Simon *et al*, 2009).

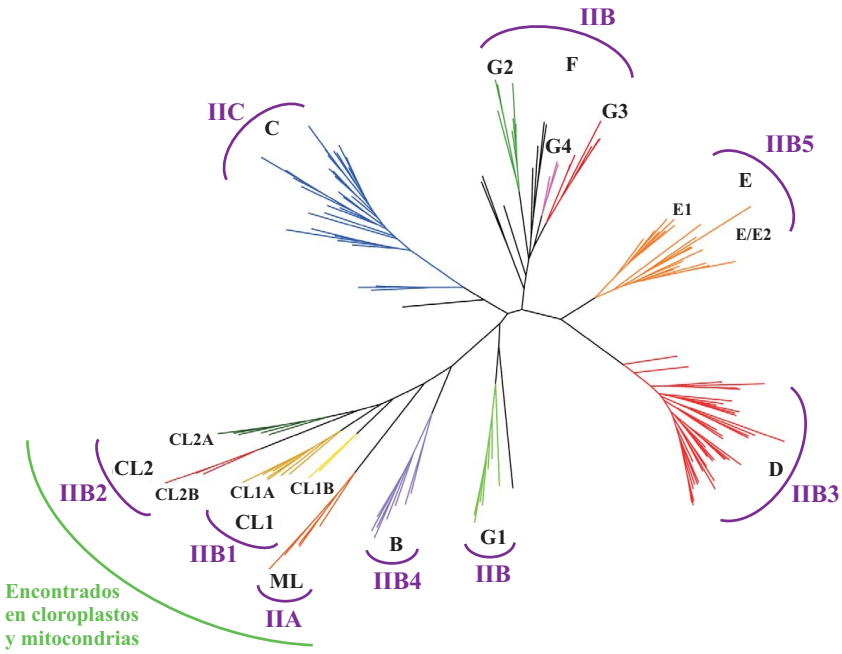


Figura I.12: **Linajes de los intrones del grupo II.** Se muestra la relación filogenética de los intrones del grupo II en base a sus IEPs (clases bacterianas, ML y CL). En morado se indica el subgrupo en base a la estructura del RNA con el que se corresponde cada clase. Las bacterias presentan intrones de todos los linajes, sin embargo, en mitocondrias y cloplastos sólo se han encontrado los linajes indicados en verde. Adaptada de Toro & Martínez-Abarca, 2013.

En mitocondrias y cloroplastos de hongos y plantas los intrones del grupo II son muy abundantes, sin embargo, su presencia en arqueas es rara, y los pocos intrones encontrados se piensa que han sido adquiridos desde bacterias por eventos de transferencia horizontal relativamente recientes. En el reino bacteriano están presentes en aproximadamente el 25 % de los genomas secuenciados, generalmente en bajo número de copias y como elementos activos con ribozima y proteína funcionales. Por el contrario, en orgánulos se encuentran frecuentemente estructuras de RNA degradadas que carecen de ORFs o bien codifican una IEP degenerada que no promueve la movilidad del intrón. Todavía no se han descrito intrones del grupo II en

genomas nucleares de eucariotas, sin embargo, los intrones espliceosómicos y retrotransposones están ampliamente distribuidos y son muy abundantes en eucariotas (Pyle, 2010; Lambowitz & Zimmerly, 2011).

I.7.2 Estructura de la ribozima

Los intrones del grupo II se caracterizan por tener una estructura secundaria muy conservada a pesar de las diferencias que puedan presentar las estructuras primarias. Esto ha permitido establecer una estructura secundaria consenso para este tipo de elementos, organizada en seis dominios alrededor de una estructura central (figura I.13). En ella se forman una serie de interacciones terciarias, en las que están involucrados motivos conservados del intrón, que dan lugar a una estructura terciaria catalíticamente activa. Algunas de esas interacciones son del tipo Watson-Crick (α-α′, β-β′, γ-γ′, δ-δ′, ϵ-ϵ′, IBS1-EBS1, IBS2,EBS2 e IBS3-EBS3), otras son interacciones tetrabucle-receptor de geometrías conocidas (ζ-ζ′, η-η′ y θ-θ′) y también se encuentran otras interacciones menos definidas no Watson-Crick (λ-λ′, κ-κ′ y μ-μ′; Pyle, 2010). Las características y funciones asociadas a cada dominio se describen a continuación.

El DI es el de mayor longitud (figura I.13) y, al igual que el DV, es imprescindible para la catálisis (Fedorova & Zingler, 2007). Se han descrito cuatro interacciones conservadas importantes para la actividad catalítica del intrón: las uniones ζ-ζ′ y κ-κ′ son críticas para la formación de la estructura terciaria, puesto que permiten el ensamblaje del DI al DV, la interacción ϵ-ϵ′ es fundamental en el posicionamiento del primer nucleótido del intrón en el sitio catalítico y la λ-λ′ juega un papel directo en la catálisis (Boudvillain & Marie Pyle, 1998; Boudvillain *et al*, 2000). Otras interacciones esenciales para el correcto plegamiento y estabilidad del intrón son la α-α′, la cual está muy conservada (Michel *et al*, 1989), la β-β′, que no está presente en todos los intrones del grupo II (Michel & Feral, 1995), la θ-θ′, que además de estabilizar la molécula acerca los DI y DIII, y la δ-δ′, descrita únicamente en intrones del grupo IIB y conocida como región de coordinación, la cual

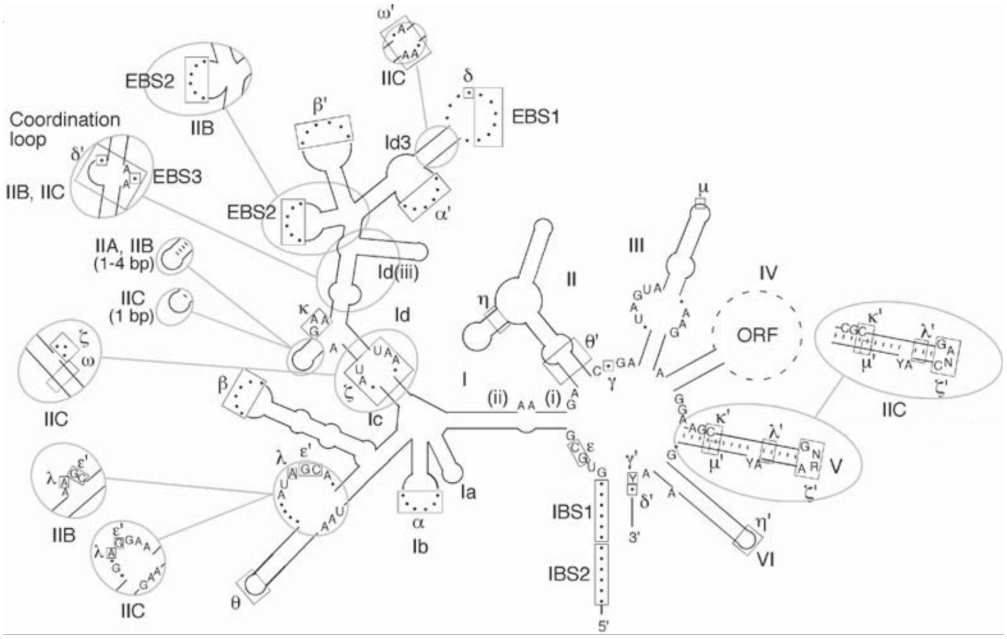


Figura I.13: **Estructura secundaria de los intrones del grupo II.** Esquema (no representado a escala) de la estructura secundaria de un intrón de la clase IIA, mostrándose en círculos las variaciones más notables en los intrones IIB y IIC respecto a esta estructura. Las cajas indican las secuencias involucradas en interacciones terciarias (letras griegas, IBSs y EBSs). Adaptada de Lambowitz & Zimmerly, 2011.

facilita que las secuencias de reconocimiento de los dos exones y la A protuberante del DVI se encuentren próximas espacialmente (Hamill & Pyle, 2006).

Además de estas interacciones terciarias intradominio e interdominio (figura I.13), en el DI se encuentran las secuencias de reconocimiento de los exones (EBS; *Exon Binding Sites*), que se unen mediante apareamiento de bases a zonas complementarias presentes en los exones (IBS, *Intron Binding Sites*). Estas uniones comprenden regiones de distinta longitud y posición dependiendo de la subclase a la que pertenezca el intrón (figura I.14). Las EBS1 y EBS2 interaccionan con las regiones IBS1 e IBS2 contenidas en la secuencia que está aguas arriba del intrón, denominado exón 1 (Jacquier

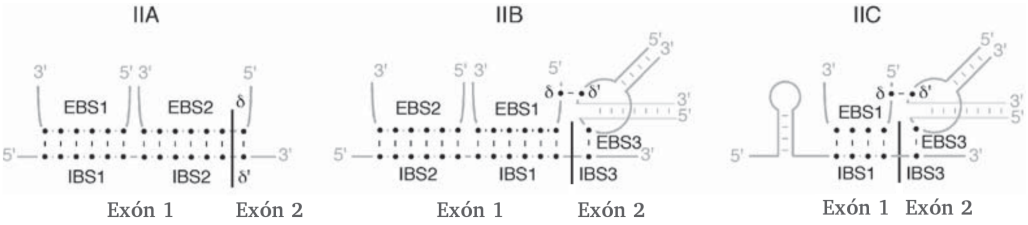


Figura I.14: **Interacciones IBS-EBS en los intrones de la clase IIA, IIB y IIC.** Se muestran los apareamientos de bases que tienen lugar en cada tipo de intrones. IBS: secuencia de unión al intrón; EBS, secuencia de unión al exón. Adaptada de Lambowitz & Zimmerly, 2011.

& Michel, 1987). Esto ocurre en los intrones de la subclase IIA y IIB, sin embargo, los intrones IIC no presentan EBS2 (Toor *et al*, 2006). Los intrones IIB y IIC también tienen en este dominio la EBS3, que aparea con la IBS3 presente en la secuencia aguas abajo del intrón, denominado exón 2. En el caso de la subclase IIA, este último apareamiento no existe, sino que es una interacción homóloga (δ-δ') la que acerca el DI al exón 2 (Costa *et al*, 2000).

El DII es la región menos conservada entre intrones del grupo II, tanto a nivel de estructura primaria como secundaria. Este dominio forma contactos terciarios esenciales con el DI (θ-θ') y el DVI (η-η'), que son importantes para un correcto plegamiento y estabilización de la molécula durante toda la reacción de escisión y que facilitan el acercamiento de los exones (Fedorova *et al*, 2003; Fedorova & Pyle, 2005).

Entre los dominios II y III se encuentra una región rica en purinas muy conservada en todos los intrones del grupo II, denominada J2/3, que forma parte del centro catalítico de la molécula (Fedorova *et al*, 2003).

El DIII está bastante conservado y, aunque no es un elemento imprescindible para la catálisis, mejora la eficiencia de escisión y de inserción del intrón. Este dominio establece una interacción terciaria con el DV (μ-μ'), que acerca el DIII al centro catalítico (Fedorova & Pyle, 2005).

El DIV es la región más variable entre intrones del grupo II. No presenta interacciones con el resto de dominios, ni está involucrado en la catálisis, pero en algunos intrones contiene un ORF que codifica una proteína conocida como IEP. Esta proteína ayuda al plegamiento de la ribozima y también se une a ella durante los procesos de escisión y de inserción del intrón (Fedorova & Zingler, 2007). Las características de la IEP se describen en el apartado I.7.3.

El DV es el más conservado filogenéticamente y, junto al DI, es absolutamente necesario para la catálisis de los intrones del grupo II. Funcionalmente, este dominio se divide en dos regiones: el sitio de unión, donde se producen las interacciones terciarias con el resto de dominios, y la región catalítica. El motivo más conservado dentro de este dominio es la llamada triada catalítica (AGC), situada en la base del DV y esencial para su funcionamiento (figura I.13; Fedorova & Zingler, 2007; Pyle, 2010).

El DVI se encuentra poco conservado entre intrones del grupo II, pero contiene una adenosina desapareada (A protuberante) muy conservada a 7-8 nucleótidos del final del intrón que está involucrada en la primera etapa de la reacción de escisión (Fedorova & Zingler, 2007).

I.7.3 Proteína codificada por el intrón (IEP)

La mayoría de intrones del grupo II bacterianos y la mitad de los presentes en mitocondrias y cloroplastos contienen un ORF en el DIV que codifica una proteína (IEP, *Intron Encoded Protein*) necesaria para la actividad de estos elementos. Las proteínas se componen de varios dominios y, dependiendo del número de dominios que presente, tienen una longitud de entre 1'2-2 kb (Lambowitz & Zimmerly, 2011). Las IEPs se traducen de forma independiente a los exones, puesto que dentro del intrón se localizan tanto el sitio de unión a los ribosomas como un codón de iniciación de la traducción; no obstante, en mitocondrias existen IEPs cuya traducción está ligada al exón que hay aguas arriba. Estas proteínas intervienen en el co-

rrecto plegamiento de la molécula de RNA del intrón, la estabilización de la ribozima y los procesos de movilidad hacia secuencias DNA diana libres de intrón (Lehmann & Schmidt, 2003).

Las IEPs de los intrones del grupo II contienen tres dominios principales conservados, el dominio RT (reverso transcriptasa), el dominio X (madurasa) y el dominio D (de unión al DNA), y en algunos de ellos también encontramos el dominio En (endonucleasa; figura I.15). El dominio RT (de entre 250-300 aa) está compuesto, a su vez, por siete bloques conservados (RT1-RT7, de color marrón en la figura I.15) y comunes a otros retroelementos (como ejemplo se muestra la RT del virus de inmunodeficiencia humana; figura I.15G). La región RT5 contiene la secuencia altamente conservada YADD, que forma parte del centro activo de la RT. La RT de los intrones del grupo II tiene mayor longitud que la de los retrovirus debido a la presencia de pequeñas regiones entre subdominios (fragmentos de color rojo en la figura I.15) y a una extensión en la zona N-terminal (RT0) que las relaciona con retrotransposones no-LTR. El dominio X (de aproximadamente 100 aa) se caracteriza por contener secuencias conservadas y tres α-hélices que son estructuralmente análogas a aquéllas encontradas en el dominio pulgar de las RTs retrovirales. Este dominio X, junto al dominio RT, promueven la formación de la estructura activa de la ribozima y posicionan la proteína para el comienzo de la transcripción. El dominio D (de en torno a 55 aa) está poco conservado, aunque presenta dos motivos importantes de unión al DNA: una zona muy rica en aminoácidos básicos y una región cuya estructura secundaria predicha es una α-hélice. El dominio En (de unos 50-60 aa) se encarga de producir el corte en el DNA diana para así generar el cebador con el que tendrá lugar la reverso transcripción del intrón. Los intrones del grupo II que no presentan este dominio (RmInt1 es el intrón con esta característica mejor estudiado; figura I.15C) utilizan un mecanismo diferente para la inserción del intrón en el DNA diana (mecanismos que veremos en apartados posteriores). El dominio En pertenece a la familia H-N-H de las endonucleasas, aunque encontramos intrones asociados a endonucleasas del tipo LAGLIDADG (típicas de los intrones del grupo I) en mitocondrias

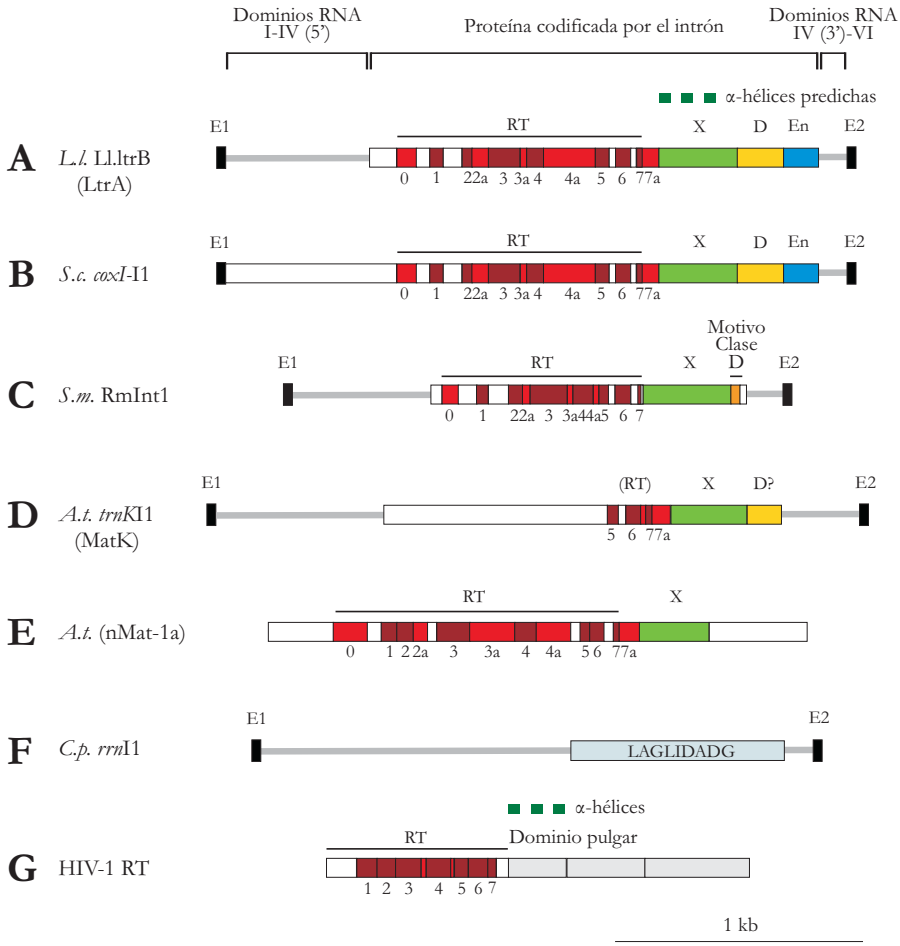


Figura I.15: **Estructura de las IEPs de los intrones del grupo II y de proteínas relacionadas**. Representación a escala de las proteínas, donde se indican los distintos dominios: en marrón el dominio RT, en verde el dominio X, en amarillo el dominio D y en azul el dominio En. Las regiones presentes entre subdominios de la RT aparecen en rojo. Encima del dominio X se señalan las α-hélices predichas, similares a las características del dominio pulgar. (A) Proteína codificada por el intrón Ll.ltrB de *L. lactis*, denominada LtrA. (B) Proteína codificada por el intrón mitocondrial *coxI*-I1 de *Saccharomices cerevisiae*. (C) Proteína codificada por el intrón RmInt1 de *S. meliloti*, la cual carece del dominio En. Se destaca el motivo de la clase D descrito para los intrones pertenecientes a este linaje. (D) Proteína codificada por el intrón *trnK*I1 de *Arabidopsis thaliana*, denominada MatK. (E) Proteína codificada por un gen nuclear en *A. thaliana*, denominada nMat-1a. (F) Proteína codificada por el intrón mitocondrial *rrn*I1 del hongo *Cryphonectria parasitica*, caracterizada por contener el motivo LAGLIDADG típico de intrones del grupo I. (G) RT del HIV-1. Adaptada de Lambowitz & Zimmerly, 2004 y 2011.

de diferentes hongos cuyas IEPs carecen de los otros dominios (figura I.15F; Lehmann & Schmidt, 2003; Lambowitz & Zimmerly, 2011). En la región C-terminal de las IEPs pertenecientes al linaje D aparece una secuencia conservada que es necesaria para la escisión y movilidad de los intrones, llamada "motivo de la clase D" (figura I.15C; Molina-Sánchez *et al*, 2010). Datos filogenéticos apuntan a que el dominio En se adquirió una sola vez por el ancestro común de las clases B, CL y ML (Toro & Martínez-Abarca, 2013).

La IEP, además de promover la escisión del intrón, interacciona con la ribozima produciendo partículas ribonucleoproteicas (RNPs) que estabilizan la molécula de intrón escindida (figura I.16). La IEP se une a un sitio de alta afinidad localizado al inicio del DIV, designado como DIVa, que contiene la secuencia Shine-Dalgarno y el codón de iniciación de la traducción, por lo que regula su propia traducción. A su vez, se producen contactos adi-

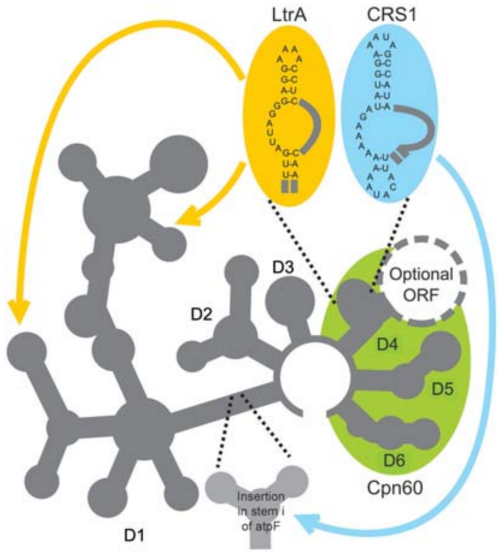


Figura I.16: **Sitio de unión a proteínas en los intrones del grupo II.** En gris se muestra un esquema de la estructura secundaria de los intrones del grupo II. Los óvalos representan alguna de las proteínas que se unen a esa estructura: LtrA en amarillo, CRS1 en azul y Cpn60 en verde. Las flechas indican interacciones con sitios de unión secundarios. Adaptada de Fedorova & Zingler, 2007.

cionales de la IEP con zonas conservadas del centro catalítico, incluyendo partes del DI, DII y DVI, que estabilizan la estructura de RNA activa (Lambowitz *et al*, 2005). La mayoría de las IEPs se unen específicamente a los RNAs que las codifican, sin embargo, algunas han evolucionado de manera que pueden llevar a cabo la escisión de múltiples intrones, generando así un "aparato común de escisión". Existen intrones relacionados dentro de un mismo genoma que se escinden con una sola IEP, aunque también se han descrito IEPs que promueven la escisión de intrones más alejados filogenéticamente. Esta función de escisión más general refleja la pérdida de especificidad con el intrón de algunas IEPs (Lambowitz & Zimmerly, 2011).

I.8 MECANISMOS DE ESCISIÓN DE LOS INTRONES DEL GRUPO II

El mecanismo de *branching* o de formación de intrón en forma de lazo es la principal ruta de escisión de los intrones del grupo II. Éste ocurre mediante dos reacciones sucesivas de transferencia de grupos fosfatos (figura I.17). En la primera reacción de transesterificación, el grupo hidroxilo (2'-OH) de la A protuberante del DVI de la ribozima es el nucleófilo que ataca

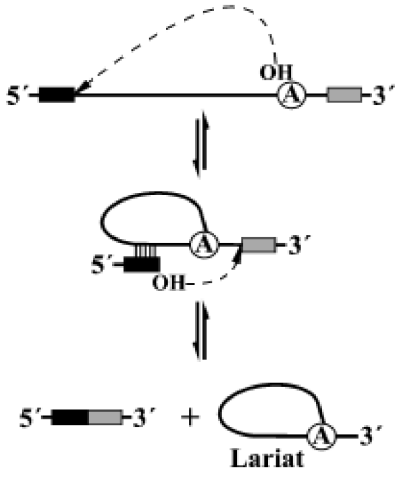


Figura I.17: **Mecanismo principal de escisión de los intrones del grupo II.** Los exones están representados por recuadros negros (exón 1) y grises (exón 2), y el RNA del intrón se muestra como una línea continua. Las flechas dobles indican la reversibilidad de las reacciones, y con flechas discontinuas se hace referencia a los ataques nucleofílicos que tienen lugar en las diferentes etapas. Adaptada de Toro *et al*, 2007.

al grupo fosfato presente en el sitio de escisión 5′. Por esta razón, la A protuberante también se conoce con el nombre de punto o sitio de ramificación o de *branch*. Tras este primer paso, la A protuberante queda unida covalentemente al primer nucleótido del intrón mediante un enlace 2′-5′ (generándose un intermediario del intrón en forma de lazo junto al exón 2), y el exón 1 permanece apareado con la ribozima (a través de las interacciones IBS-EBS). En la segunda transesterificación, es el grupo hidroxilo (3′-OH) libre del exón 1 el que actúa como nucleófilo atacando al grupo fosfato del sitio de escisión 3′. Finalmente, los dos exones quedan unidos covalentemente mediante un enlace 3′-OH/5′-P y se libera una molécula de intrón en forma de lazo (*lariat*; Lambowitz & Zimmerly, 2011). Durante este proceso, la primera reacción es el factor limitante de la escisión de los intrones, habiéndose detectado el intermediario *lariat*-exón 2 únicamente por inhibición de la segunda reacción de transesterificación (Pyle & Lambowitz, 2006).

Para este tipo de intrones se han descrito otros mecanismos de escisión minoritarios: mecanismo de hidrólisis o de formación de intrón lineal, en el que el nucleófilo involucrado en la primera reacción de transesterificación es una molécula de agua (Bonen & Vogel, 2001), y mecanismo de formación de intrón circular, el cual se basa en la unión covalente del primer y el último nucleótido del intrón mediante un enlace fosfodiéster 2′-5′ después de que el grupo hidroxilo (2′-OH) del último nucleótido del intrón ataque al sitio de escisión 5′ (Murray *et al*, 2001; Molina-Sánchez *et al*, 2006).

I.8.1 Factores del hospedador implicados en la escisión

Algunos intrones del grupo II (sobre todo los mitocondriales y cloroplastidiales) no codifican su propia IEP, por lo que su escisión depende de proteínas codificadas por otros intrones o bien de proteínas codificadas por el hospedador. Existen dos grupos básicos de proteínas que facilitan la escisión de los intrones del grupo II: las madurasas, que se unen específicamente al intrón y ensamblan la estructura activa de éste, y las chaperonas,

que ayudan al correcto plegamiento de diversas proteínas y generalmente no son específicas (Beauregard *et al*, 2008).

Un estudio detallado sobre los factores del hospedador necesarios para la escisión de los intrones del grupo II cloroplastidiales indica que éstos utilizan proteínas de unión al RNA pertenecientes a tres grandes familias específicas de plantas: CRM (proteínas para la escisión de RNA de cloroplastos y maduración de ribosomas), POROR (reconocimiento de RNA de plantas) y PPR (proteínas con repeticiones de pentatricopéptidos). Un análisis con el alga verde *Chlamydomonas reinhardti* reveló que sus dos intrones necesitan al menos 14 genes nucleares para llevar a cabo la escisión. Por tanto, los requerimientos por parte del hospedador son diferentes para cada intrón, lo cual refleja la necesidad de estudiar los distintos intrones de manera independiente (Lambowitz & Zimmerly, 2011).

Como hemos comentado anteriormente, la primera reacción de transesterificación es limitante en la escisión de los intrones del grupo II, quizás por la desestabilización del intermediario *lariat*-exón 2. Este motivo ha llevado a la idea de que estos elementos necesitan chaperonas para estabilizar las estructuras intermediarias y la ribozima activa. Se han descrito varios factores que actúan como chaperonas en la estabilización de determinados intrones. Algunos de éstos son las proteínas de la caja DEAD Mss116, Ded1 y CYT-19, y otras proteínas como StpA, Hfq y Cbp2 (Fedorova *et al*, 2010). Algo similar ocurre en los intrones espliceosómicos, donde proteínas de la caja DEAD y helicasas funcionan en múltiples reacciones para acelerar las transiciones estructurales (Pyle & Lambowitz, 2006).

I.9 MECANISMOS DE MOVILIDAD DE LOS INTRONES DEL GRUPO II

La dispersión de los intrones del grupo II, así como la de otros retroelementos, conlleva dos mecanismos asociados, uno de escisión desde la copia de la diana donde se encuentra insertado, y otro de invasión de una secuen-

cia diana libre de intrón. El proceso de movilidad se basa en la interacción que mantienen el intrón escindido y la IEP, con la que se generan RNPs que promueven la inserción del intrón (Toro *et al*, 2007). Se han descrito dos mecanismos de movilidad dependiendo de si la diana invadida es la habitual (*retrohoming*) o es un sitio ectópico (retrotransposición).

I.9.1 Proceso de *retrohoming*

A pesar de que muchos intrones son capaces de escindirse y formar RNPs, sólo se ha podido demostrar la movilidad de alguno de ellos. El mecanismo de *retrohoming* (habitualmente denominado *homing*) de los intrones del grupo II comienza con el reconocimiento del DNA diana por parte de las RNPs tras su unión de forma inespecífica al DNA. Durante este proceso participan de manera activa ambos componentes, ribozima y proteína (figura I.18A). El RNA del intrón se une a los exones mediante apareamiento de bases, generando así las interacciones IBS-EBS 1 y 2, y δ-δ′ en los intrones IIA e IBS3-EBS3 en los IIB. La principal función de la IEP parece ser el desenrollamiento de la doble hélice de DNA, y para llevarla a cabo establece contactos con las posiciones distales de ambos exones (figura I.18A; Lehmann & Schmidt, 2003).

Este proceso ha sido estudiado detalladamente en varios intrones, en los que se ha descrito la diana mínima requerida para un *homing* eficiente (figura I.18B). Los intrones mitocondriales de la levadura *Saccharomices cerevisiae coxI*-I1 (aI1) y *coxI*-I2 (aI2), y el intrón bacteriano Ll.ltrB de *L. lactis* pertenecen a la clase IIA, mientras que RmInt1, de *S. meliloti*, es un intrón IIB. Recientemente se han caracterizado dos nuevos intrones bacterianos, EcI5 de *E. coli* (Zhuang *et al*, 2009) y TeI4h de *Thermosynechococcus elongatus* (Mohr *et al*, 2010), pertenecientes a la clase IIB y con un mecanismo de movilidad similar al mostrado por Ll.ltrB. Además del correcto apareamiento de bases que tienen que presentar las interacciones de los exones con el DI de la ribozima, existen residuos en la zona distal de ambos exones importantes para el proceso de movilidad (señalados mediante rectángulos y

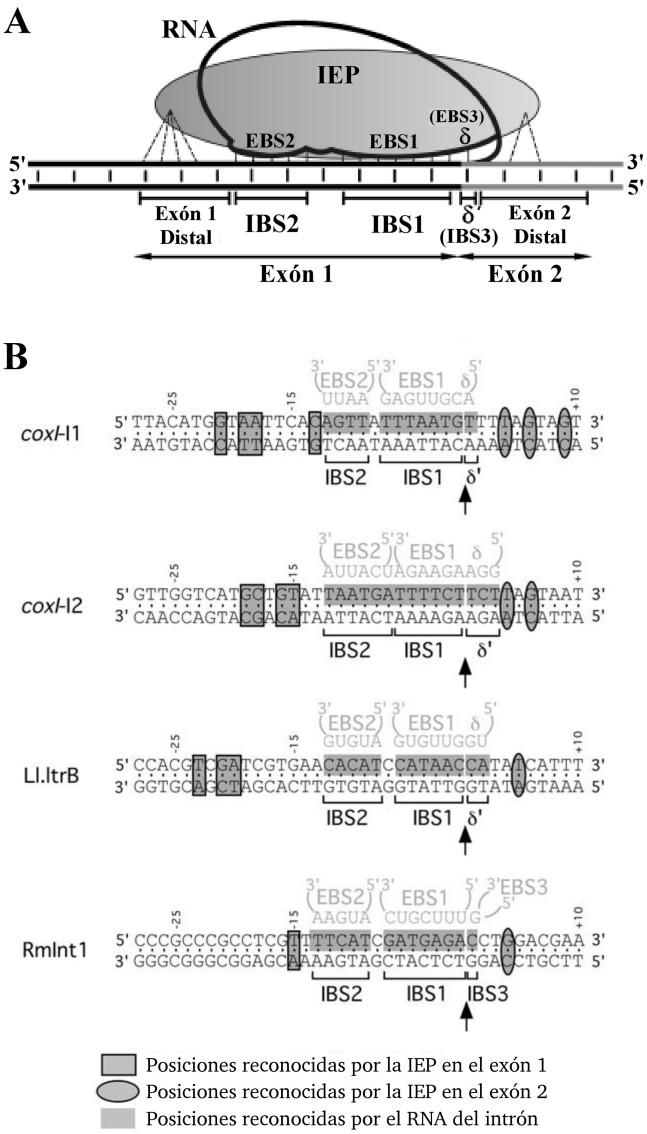


Figura I.18: **Reconocimiento del DNA diana por los intrones del grupo II.** (A) Representación de la unión de las RNPs al sitio diana. Se muestran las interacciones IBS-EBS y δ-δ′ de los exones con el RNA del intrón y la unión de la IEP a la zona distal de ambos exones (con líneas discontinuas). (B) Reconocimiento del DNA diana de algunos intrones móviles. En cada caso se destacan las posiciones reconocidas por el RNA y la IEP. La flecha indica el sitio de inserción del intrón. Los intrones *coxI*-I1, *coxI*-I2 y LI.ltrB pertenecen a la clase IIA, mientras que RmInt1 es un intrón IIB. Adaptada de Lambowitz & Zimmerly, 2004 y Toro *et al*, 2007.

óvalos en cada uno de los intrones en la figura I.18B; Jiménez-Zurdo *et al*, 2003; Lehmann & Schmidt, 2003).

Los dos intrones del grupo II bacterianos mejor caracterizados hasta el momento son el intrón de *L. lactis*, Ll.ltrB (Matsuura *et al*, 1997), y el de *S. meliloti*, RmInt1 (Martínez-Abarca *et al*, 1998). Ambos intrones presentan características en su proceso de movilidad que ha llevado a diferenciar dos tipos de mecanismos.

I.9.1.1 Homing *mediante reacción de reverso transcripción cebada por el DNA diana (TPRT)*

Este mecanismo ha sido ampliamente estudiado en el intrón Ll.ltrB, y es similar al descrito en levaduras y retrotransposones no-LTR. Las RNPs reconocen y se unen a una secuencia de DNA diana que se extiende desde la posición -25 a la +9 relativas al sitio de inserción del intrón (figura I.18B; Singh & Lambowitz, 2001). El componente ribozímico de las RNPs corta la hebra sentido en el sitio de unión de los dos exones y el intrón completo se integra mediante una reacción de escisión reversa (figura I.19). Tras la inserción, ocurren una serie de reorganizaciones conformacionales que acaban con el corte de la cadena antisentido entre las posiciones +9 y +10 (en el exón 2). El extremo 3'-OH generado tras este corte sirve de cebador para que la RT sintetice el cDNA usando como molde el RNA del intrón insertado en la hebra sentido. Tras la degradación del molde de RNA (probablemente por una RNAsa H celular), tiene lugar la síntesis del DNA complementario a la hebra antisentido (donde se encuentra el cDNA generado por la RT). Este proceso es llevado a cabo por una DNA polimerasa del hospedador, usando como cebador el extremo 3'-OH libre en el exón 1. Finalmente, la ligación de las dos hebras, y su reparación por la maquinaria celular, completa el evento de movilidad de estos intrones, el cual es independiente de recombinación homóloga. La integración del intrón mediante escisión inversa es un proceso idéntico al de escisión, en el que es necesario la formación de las

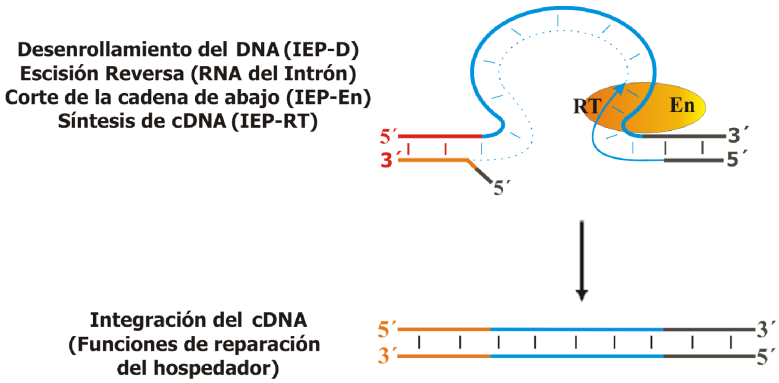


Figura I.19: **Mecanismo de movilidad mediante reacción de reverso transcripción cebada por el DNA diana (TPRT).** Tras el reconocimiento de la diana, el dominio D de la IEP (representada como un óvalo amarillo) desenrolla el DNA y se produce la escisión reversa del RNA del intrón (línea azul) en la cadena de arriba. El dominio En de la IEP corta la cadena de abajo, la cual es usada por la RT como cebador para realizar la reverso transcripción (flecha azul) del intrón integrado. Finalmente el cDNA es integrado en la nueva localización mediante las funciones reparadoras del hospedador. Adaptada de Toro *et al*, 2007.

interacciones anteriormente mencionadas (Matsuura *et al*, 1997; Cousineau *et al*, 1998; Lambowitz & Zimmerly, 2004).

I.9.1.2 Homing *independiente del dominio endonucleasa de la IEP*

El mecanismo de movilidad de tipo TPRT necesita la generación de un cebador que permita iniciar la reacción de transcripción inversa. Sin embargo, no todas las IEPs de los intrones del grupo II contienen un dominio En que produzca ese cebador mediante un corte en el DNA diana. RmInt1, el intrón de este tipo mejor caracterizado, invade su diana de manera muy eficiente (similar al intrón de *L. lactis* o a los de levadura), mostrando dos vías de movilidad (Martínez-Abarca *et al*, 2004).

El mecanismo principal está basado en una integración del RNA del intrón dentro del DNA diana de cadena sencilla durante la replicación celular

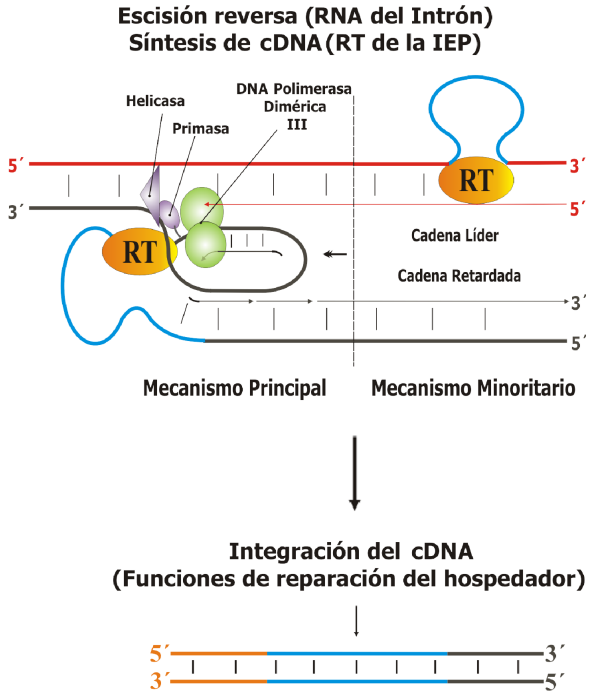


Figura I.20: **Mecanismo de movilidad (principal y minoritario) independiente de dominio endonucleasa.** En el esquema del mecanismo principal se muestran los elementos implicados en la transcripción reversa del RNA del intrón (línea azul), la cual está ligada a la replicación del DNA. En el mecanismo minoritario, el intrón sólo requiere la actividad RT de la IEP. En ambas vías, el cDNA es finalmente integrado en la nueva localización mediante las funciones reparadoras del hospedador. Adaptada de Toro *et al*, 2007.

(figura I.20). Como cebador para iniciar la transcripción reversa del intrón, la RT puede usar los pequeños fragmentos de RNA sintetizados por la primasa o los propios fragmentos de Okazaki (Martínez-Abarca *et al*, 2004). Este proceso presenta un sesgo de inserción preferencial por la cadena usada de molde para la síntesis de la hebra retardada en la horquilla de replicación. En el intrón Ll.ltrB también se ha descrito un mecanismo similar a éste (dependiente de la replicación del DNA) en mutantes con el dominio En mutado, donde el corte de la cadena antisentido se encuentra bloqueado. Aunque, a diferencia de RmInt1, las frecuencias de movilidad más altas

se obtuvieron cuando la RT utilizaba la cadena líder recién sintetizada como cebador para la reacción de transcripción inversa (Zhong & Lambowitz, 2003).

RmInt1 puede llevar a cabo un segundo mecanismo de movilidad, que es minoritario e independiente de la replicación celular (figura I.20). Este proceso podría implicar sitios diana de cadena doble, otros sitios de cadena sencilla que se encuentren así temporalmente, o cebadores alternativos de la reacción de transcripción inversa como pueden ser cortes aleatorios no específicos en el DNA, la cadena líder en la horquilla de replicación (en vez de la retrasada), o la iniciación del cebado *de novo* (Muñoz-Adelantado *et al*, 2003).

I.9.2 Proceso de retrotransposición

Los intrones del grupo II también pueden moverse a sitios no idénticos de los que proceden, proceso que ocurre a muy baja frecuencia (en RmInt1 representa un 5 % respecto al *homing* sobre su secuencia diana habitual; Martínez-Abarca & Toro, 2000). Este proceso, al igual que el *homing*, requiere la inserción del RNA del intrón mediante escisión reversa, lo que asegura la invasión de sitios a partir de los que posteriormente el intrón pueda escindirse, minimizando así el daño al hospedador (Cousineau *et al*, 2000). El proceso de retrotransposición de los intrones del grupo II bacterianos en poblaciones naturales ocurre preferentemente en dianas localizadas dentro de elementos genéticos móviles como son las ISs o los plásmidos. Esto apoya la hipótesis de que este mecanismo de movilidad es la principal estrategia de expansión de los intrones del grupo II (Muñoz *et al*, 2001).

Para que el intrón lleve a cabo el mecanismo de retrotransposición se necesita DNA de cadena sencilla, y, además, se ha observado que el proceso se produce en ausencia de actividad DNA endonucleasa (Cousineau *et al*, 2000). Estos datos sugieren que la vía de retrotransposición de los intrones del grupo II se asemeja al proceso de *homing* de los intrones que care-

cen de dominio En (apartado I.9.1.2). Aunque estudios con el intrón Ll.ltrB indican que la retrotransposición puede ocurrir por diversos mecanismos dependiendo del organismo hospedador. En *E. coli*, la retrotransposición de Ll.ltrB se produce también sobre DNA de cadena doble y es frecuente en zonas cercanas al Ori y Ter del cromosoma (posiblemente debido a la presencia de la IEP en los polos de la célula bacteriana); mientras que en *L. lactis* dicho proceso sigue la vía descrita previamente (apartado I.9.1.1; Coros *et al*, 2005).

I.9.3 Factores del hospedador implicados en la movilidad

Unido a los factores involucrados en la escisión mencionados anteriormente (apartado I.8.1), los intrones del grupo II necesitan la maquinaria celular para llevar a cabo su mecanismo de movilidad y completar la inserción del intrón en su diana. Después de que la RT realice la transcripción reversa del RNA del intrón, éste debe ser eliminado, proceso en el cual se encuentran involucradas la RNAsa H1 (*rnhA*) y la actividad endonucleasa 5'-3' de la Pol I (*polA*). Para la síntesis de la segunda cadena de cDNA se requiere la DNA polimerasa Pol III (*dnaQ*), y se piensa que el extremo del exón 1 usado como cebador en esta reacción es eliminado por la exonucleasa RecJ. La reparación de las muescas tras la síntesis del cDNA es llevada a cabo por las polimerasas reparadoras Pol II (*polB*), Pol IV (*dinB*) y Pol V (*umuDC*), y para la unión de las muescas se utiliza una DNA ligasa del hospedador. Todos estos requerimientos sugieren que el hospedador considera las últimas etapas del *homing* como una respuesta a un daño en el DNA celular (Smith *et al*, 2005; Beauregard *et al*, 2008).

Las condiciones de estrés nutricional para el hospedador (como la acumulación de ppGpp o la activación de cAMP por falta de carbono) también actúan como un estímulo que promueve el movimiento del intrón, favoreciéndose la retrotransposición frente al *homing*. Esta proliferación de los intrones permite generar diversidad y reorganización de genomas que, a

su vez, puede favorecer la respuesta del hospedador al estrés (Coros *et al*, 2009).

Además de todas esas proteínas que favorecen la movilidad de los intrones del grupo II, encontramos otras que influyen negativamente en dicho proceso. Dentro de este grupo de proteínas se incluyen las RNAsas (RNAsa I y RNAsa E), que son enzimas que degradan el RNA del intrón, y la DNasa Exo III (*xthA*), que mediante su actividad exonucleasa 3'-5' degrada la nueva hebra de cDNA generada o interfiere en su síntesis. Así, sin ningún mecanismo de silenciamiento específico, la movilidad de los intrones puede verse inhibida dependiendo del estado fisiológico de la célula (Smith *et al*, 2005; Beauregard *et al*, 2008).

I.10 APLICACIONES BIOTECNOLÓGICAS DE LOS INTRONES DEL GRUPO II

Los intrones del grupo II presentan una serie de características que los hacen adecuados para su uso como herramientas biotecnológicas en diversas áreas como medicina y agricultura, y tanto en procariotas como en eucariotas (Cui & Davis, 2007; Toro *et al*, 2007):

- Son elementos móviles capaces de insertarse específicamente en su DNA diana con una elevada eficiencia mediante un proceso independiente de recombinación homóloga.
- Reconocen el DNA diana principalmente por apareamiento de bases, por tanto, la especificidad de la diana se puede cambiar modificando los nucleótidos que forman parte de las EBSs del intrón para que sean capaces de interaccionar con las IBSs de la diana deseada.
- Pueden movilizar cualquier información genética que se integre en el intrón.

- La movilidad de estos elementos requiere la actividad de la IEP y otras proteínas del hospedador probablemente codificadas por genes conservados, lo que permite su utilización en diversos hospedadores.

Gracias a estas características, se ha conseguido reprogramar varios intrones para que se inserten en dianas de interés: Ll.ltrB de *L. lactis* (Guo *et al*, 2000), EcI5 de *E. coli* (Zhuang *et al*, 2009) y RmInt1 de *S. meliloti* (García-Rodríguez *et al*, 2011). La reprogramación de estos intrones se lleva a cabo con una forma derivada del intrón silvestre en la que la IEP se expresa aguas arriba de la parte ribozímica (intrón ΔORF), y que es una versión del intrón más eficiente (Guo *et al*, 2000; Nisa-Martínez *et al*, 2007; Zhuang *et al*, 2009). La ventaja del intrón ΔORF es que, una vez insertado, sólo puede escindirse de la secuencia que ha interrumpido mediante expresión de la IEP en *trans* (Nisa-Martínez *et al*, 2007).

Respecto a la reprogramación de los intrones, se ha desarrollado y validado un programa informático que permite encontrar los posibles sitios diana presentes en la secuencia del gen que se desea interrumpir (Perutka *et al*, 2004). De esta manera se pueden seleccionar inserciones del intrón hacia la cadena sentido o antisentido del gen. En el caso de llevar a cabo una invasión de la cadena antisentido, la inserción que se produce es definitiva o irreversible, puesto que la transcripción y posterior escisión del intrón quedan impedidas. Por el contrario, si la invasión ocurre en la cadena sentido del gen, la inserción es condicional o reversible, pudiéndose producir la transcripción del intrón y la escisión de éste si se expresa la IEP en *trans* (Nisa-Martínez *et al*, 2007).

Por otro lado, se ha diseñado un sistema de selección ligado a la movilidad del intrón para detectar la interrupción del gen de interés de una manera más eficiente. Este sistema se basa en la introducción de un gen de resistencia a trimetoprima (Tp) en el DIV del intrón ΔORF en orientación inversa a la transcripción de éste. Ese gen, a su vez, se encuentra interrumpido por un intrón del grupo I (*td*I) dispuesto en la misma orientación que la transcripción del intrón ΔORF. Así, la Tp sólo se activa tras la inserción

cromosómica del intrón ΔORF y la posterior escisión del intrón *td*I. A todo el conjunto se le ha denominado marcador de selección activado por retrotransposición (RAM; Zhong *et al*, 2003).

OBJETIVOS

Los rizobios son un grupo de α-proteobacterias de gran interés por la relación simbiótica que establecen con plantas leguminosas. Una de las especies bacterianas que viven en simbiosis mejor estudiadas es *Sinorhizobium meliloti*. Con la cepa GR4, aislada de un suelo agrícola de Granada, se lleva trabajando desde hace más de 30 años, y en ella se identificó el primer intrón del grupo II encontrado en la familia Rhizobiaceae, RmInt1. Desde su descubrimiento a finales de los años 90, RmInt1 ha sido ampliamente estudiado y caracterizado en nuestro grupo de investigación.

El objetivo de la presente Tesis Doctoral consiste en la secuenciación del genoma del endosimbionte diazotrófico *S. meliloti* GR4, así como en la caracterización de nuevos intrones del grupo II, determinando su distribución y sus posibles aplicaciones biotecnológicas. El desarrollo de este objetivo general se llevó a cabo mediante la consecución de los siguientes objetivos específicos:

1. Resolución del genoma completo de la cepa *S. meliloti* GR4. Localización de los intrones del grupo II presentes en ella.

2. Distribución de intrones del grupo II filogenéticamente relacionados con RmInt1 en Rhizobiaceae. Actividad de estos elementos móviles en diversas cepas y posible implicación de la chaperona GroEL.

3. Incremento de la colección de intrones del grupo II usados como herramientas biotecnológicas de mutagénesis. Caracterización funcional de los intrones RmInt2 y SmedInt1. Determinación del papel de la proteína codificada por el intrón (IEP) en el reconocimiento del DNA diana.

MATERIAL Y MÉTODOS

M.1 CEPAS BACTERIANAS

Las especies y cepas bacterianas utilizadas en este trabajo se encuentran descritas en la tabla M.1.

Tabla M.1: Cepas bacterianas utilizadas en este trabajo.

Cepa bacteriana	**Características**	**Referencia**
Escherichia coli DH5α	F$^-$, ø80d*lacZ*ΔM15, Δ(*lacZYA-argF*)U169, *deo*R, *recA*1, *endA*1, *hsd*R17(rK$^-$, mK$^+$), *phoA*, *supE*44, λ^-, *thi*-1, *gyrA*96, *relA*1.	Bethesda Research Lab
E. coli HB101	*supE*44, Δ(*mcrC-mrr*), *recA*13, *ara*-14, *proA*2, *lacY*1, *galK*2, *rpsL*20, *xyl*-5, *mtl*-1, *leuB*6, *thi*-1.	Promega Corporation
Sinorhizobium meliloti GR4	Aislado de suelo de Granada; *Nod*$^+$, *Fix*$^+$. Contiene 10 copias completas de RmInt1 y 7 de RmInt2.	Casadesús & Olivares (1979)
S. meliloti 1021	Mutante espontáneo a estreptomicina derivado de la cepa S. meliloti 2011; Sm^R. Posee 3 copias completas de RmInt1.	Meade *et al* (1982)
S. meliloti RMO17	Aislado de suelo de Salamanca. No contiene copias de intrones tipo RmInt1.	Villadas *et al* (1995)
Sinorhizobium medicae WSM419	Aislado de suelo de Cerdeña. Posee 4 copias completas de SmedInt1.	Reeve *et al* (2010)
S. medicae RMO15	Aislado de suelo del norte de España. No contiene copias de intrones tipo RmInt1.	Fernández-López *et al* (2005)
S. medicae RMO09	Aislado de suelo del norte de España. Posee, al menos, 3 copias completas de intrones tipo RmInt1.	Fernández-López *et al* (2005)
S. medicae RMO04	Aislado de suelo del norte de España. No contiene copias completas de intrones tipo RmInt1.	Fernández-López *et al* (2005)
S. medicae RMO02	Aislado de suelo del norte de España. No contiene copias completas de intrones tipo RmInt1.	Fernández-López *et al* (2005)
Rhizobium etli CFN42	Aislado de suelo de Méjico. No contiene copias completas de intrones tipo RmInt1.	González *et al* (2006)
R. etli CIAT652	Aislado de suelo de Costa Rica. No contiene copias completas de intrones tipo RmInt1.	González *et al* (2010)

M.2 PLÁSMIDOS Y VECTORES DE CLONACIÓN

El trabajo desarrollado en la presente Tesis Doctoral ha requerido el uso de vectores comerciales y plásmidos previamente generados en el grupo

Tabla M.2: Vectores y plásmidos utilizados en este trabajo.

Nombre	Características	Referencia
pGEM-T Easy	Vector para la clonación de productos de PCR. Contiene los promotores T7 y SP6. Ap^R.	Promega
pRK2013	Vector para la movilización en *trans* de plásmidos no autotransmisibles. Km^R.	Figurski & Helinski (1979)
pKG0	Plásmido derivado de pKG10. Construcción sin intrón y sin diana. Km^R.	Martínez-Abarca *et al* (2000)
pKGEMA4	Plásmido derivado de pKG2.5. Construcción donadora del intrón RmInt1 en su forma ΔORF. Km^R.	Nisa-Martínez *et al* (2007)
pKGEMA4DV	Plásmido derivado de pKGEMA4. Construcción donadora del intrón RmInt1 defectivo en escisión. Km^R.	Nisa-Martínez *et al* (2007)
pCm4	Plásmido derivado de pBBR1MCS. Construcción donadora del intrón RmInt1 en su forma ΔORF. Cm^R.	Nisa Martínez (2011)
pGm4	Plásmido derivado de pBBR1MCS-5. Construcción donadora del intrón RmInt1 en su forma ΔORF. Gm^R.	Nisa Martínez (2011)
pJB0.6LEAD	Plásmido derivado de pJB3Tc19. Construcción receptora del intrón RmInt1. Contiene la diana IS*Rm2011-2* en orientación inversa al sentido de la horquilla de replicación. Ap^R.	Martínez-Abarca *et al* (2000)
pJB0.6LAG	Plásmido derivado de pJB0.6LEAD. Construcción receptora del intrón RmInt1. Contiene la diana IS*Rm2011-2* en orientación directa al sentido de la horquilla de replicación. Ap^R.	Martínez-Abarca *et al* (2004)
pJBΔ129	Plásmido derivado de pJB0.6LEAD. Construcción receptora del intrón RmInt1 que carece de la región de reconocimiento para dicho intrón. Ap^R.	Martínez-Abarca *et al* (2000)

de investigación, que se detallan en la tabla M.2. Con el fin de caracterizar los nuevos intrones presentados en este trabajo, se llevó a cabo la construcción de una serie de plásmidos tanto donadores como receptores de intrón. Además, se generaron varios plásmidos para realizar ensayos de complementación de unos mutantes de 1021 para los genes *groEL* proporcionados por la Dra. Valerie Oke (miembro del departamento de Ciencias Biológicas de la Universidad de Pittsburgh, Pensilvania, USA). Los oligonucleótidos utilizados en la creación de todos estos plásmidos se muestran en la tabla M.3, y la estrategia seguida para la construcción de cada uno de ellos se describe a continuación.

Tabla M.3: Oligonucleótidos, diseñados en este trabajo, utilizados para la construcción de diferentes plásmidos.

Nombre	**Secuencia (5'-3')**
1F_RmInt2_IS17	GGGATCCCACGTGACTAGTGCCCGTCCATAG
1R_RmInt2_IS17	GGGGATCCCTAGGCGTTGGCGAGTAGC
1F_SmedInt1_IS66	GGGGATCCCACGTGTCCAAGGGTGTTCTGTGACGA
1R_SmedInt1_IS66	GGGGATCCCTAGGATCGAGCCGCGGAAG
IEP_SpeI_RmInt2	GGGGACTAGTGGAAACAGGATGACGTCGGC
IEP_SacI_RmInt2	GCCGGAGCTCTCAGGCAAACGTGCCGATC
IEP_SpeI_Smed	GGGGACTAGTGGAAACAGGATGACTTCG
IEP_SacI_Smed	GCCGGAGCTCTCAGGCAAACAGGTTCGT
dORF1_RmInt2	GCCGCTCGAGAAGCTTCGTACACCAGCCGC
dORF2_RmInt2	GCCGCTCGAGTCTCTTTGTGCACTGGCAAC
dORF1_Smed	GCCGCTCGAGAAGCTTCGTACACTTGCC
dORF2_Smed	GCCGCTCGAGCCTCTTCGTACACTGGCA
ISRm17_F*	GGCCGCCACGTGCGATCTAGACTAGTGCCCGTCCATAAACAGCAAGTCCTAGGGA
ISRm17_R*	GGCCTCCCTAGGACTTGCTGTTTATGGACGGGCACTAGTCTAGATCGCACGTGGC
Quimera-20/+5_F*	GGCCGCCACGTGCCTCGTAGTGCCCGTCCATAACTGGCCTAGGGA
Quimera-20/+5_R*	GGCCTCCCTAGGCCAGTTATGGACGGGCACTACGAGGCACGTGGC
Quimera-21/+5_F*	GGCCGCCACGTGCCTCGTTAGTGCCCGTCCATAACTGGCCTAGGGA
Quimera-21/+5_R*	GGCCTCCCTAGGCCAGTTATGGACGGGCACTAACGAGGCACGTGGC
pJB_secF	CCTCTTCGCTATTACGCCAGC
pJB_secR	CGGCTCGTATGTTGTGTGGA
GroESL3_F	CGGACTAGTTGCAGACACTGATTTCAAGCCA
GroESL3_R	CGGACTAGTTGCTGGGTAATTGACGTCTTGCTG
GroESL3_sec_F	TGCTTCGCGGCGTGGAAT
GroESL3_sec_R	AAGGGCCGATTTTGCCCGTAG

* oligonucleótidos fosforilados en su extremo 5'

M.2.1 Plásmidos donadores de intrón

Este tipo de construcciones se basan en los plásmidos desarrollados para el estudio del intrón del grupo II RmInt1 (Martínez-Abarca *et al*, 2000; Nisa-Martínez *et al*, 2007). Con el fin de generar los diferentes plásmidos donadores de los nuevos intrones RmInt2 y SmedInt1 realizamos una PCR con la DNA polimerasa *Phusion High-Fidelity* (Finnzymes; apartado M.6.1) sobre DNA genómico de GR4 y WSM419 respectivamente, usando los oligonucleótidos 1F_RmInt2_IS17 y 1R_RmInt2_IS17 para el intrón RmInt2, y 1F_SmedInt1_IS66 y 1R_SmedInt1_IS66 para SmedInt1. De esta manera, obtuvimos un fragmento con cada uno de los intrones y sus respectivas dianas, que contienen los últimos 20 nucleótidos del exón 1 y los 5 primeros nucleótidos del exón 2 (diana -20/+5 respecto al sitio de inserción del intrón). El producto de PCR generado con el intrón RmIn2 y su diana tiene un tamaño de 1.934 pb, y el del intrón SmedInt1 con su diana de 1.955 pb. Estos fragmentos, que presentan en su extremo 5′ la secuencia de reconocimiento para las enzimas de restricción *Bam*HI y *Pml*I, y en su extremo 3′ para *Bam*HI y *Bln*I, fueron introducidos en el vector pGEM-T Easy para generar las construcciones pGEMtR2 y pGEMtS1 (figura M.1). Éstos se mandaron a secuenciar con los oligonucleótidos universales T7 (5′-TAATACGACTCACTATAGGG-3′) y Sp6 (5′-ATTTAGGTGACACTATAG-3′) para comprobar que la secuencia de los intrones y sus dianas no había sufrido mutaciones. Los plásmidos donadores de intrón silvestre (pKGRmInt2 y pKGSmedInt1; figura M.1) se obtuvieron mediante digestión *Bam*HI de pGEMtR2 y pGEMtS1, clonación en el sitio *Bam*HI del plásmido pKG0 (Martínez-Abarca *et al*, 2000) y posterior selección de los plásmidos que habían introducido el intrón en sentido.

Para la creación de los plásmidos donadores de intrón en su forma derivada ΔORF tuvimos que generar otras construcciones intermedias. Llevamos a cabo dos PCRs con la DNA polimerasa *Phusion High-Fidelity* (Finnzymes; apartado M.6.1) sobre DNA de los plásmidos pGEMtR2 y pGEMtS1: una directa para amplificar la IEP con unos oligonucleótidos que producen un fragmento flanqueado por las enzimas de restricción *Spe*I en el extremo 5′ y *Sac*I en el extremo 3′ (IEP_SpeI_RmInt2 e IEP_SacI_RmInt2 para RmInt2;

IEP_SpeI_Smed e IEP_SacI_Smed para SmedInt1), y una inversa para amplificar sólo la ribozima del intrón con unos oligonucleótidos que tienen en su extremo 5′ el sitio de corte para la enzima *Xho*I (dORF1_RmInt2 y dORF2_RmInt2 para RmInt2; dORF1_Smed y dORF2_Smed para SmedInt1). El producto de PCR que porta la IEP del intrón (de 1.289 pb en ambos casos) se introdujo en el vector pGEM-T Easy, formando así los plásmidos pGEMtIEPR2 y pGEMtIEPS1 (figura M.1). El producto de la PCR inversa (de 3.821 pb en el caso de RmInt2 y de 3.840 pb en SmedInt1) se digirió con *Xho*I y se autoligó (apartado M.5.3 y M.5.5), dando lugar a los plásmidos pGEMtΔORFR2 y pGEMtΔORFS1 (figura M.1). Tanto los plásmidos que portan sólo la IEP del intrón como los que llevan sólo el ΔORF se volvieron a secuenciar con los oligonucleótidos T7 y Sp6 para continuar con una muestra que no hubiera sufrido mutaciones en su secuencia. Los plásmidos pKGEMA4R2 y pKGEMA4S1 (figura M.1) se obtuvieron mediante el intercambio de ambas partes, ribozima e IEP, con la construcción pKGEMA4 (Nisa-Martínez *et al*, 2007). Estas inserciones son dirigidas: la IEP se clonó en el sitio *Spe*I-*Sac*I de pKGEMA4 y el ΔORF en el sitio *Pml*I-*Bln*I del mismo.

A partir de los tres plásmidos donadores de intrón en su forma derivada ΔORF (pKGEMA4, pKGEMA4R2 y pKGEMA4S1) se llevó a cabo la construcción de los plásmidos donadores de intrones quiméricos (figura M.2). Cada uno de estos plásmidos, que contienen la ribozima y la proteína de intrones distintos, se generaron mediante el intercambio de una de sus partes con la ribozima o la proteína de otro de los plásmidos.

Todos los plásmidos donadores contienen el casete de resistencia a Km, y la forma de intrón que portan está bajo el promotor de Km (indicado como pKm en la figura M.1 y M.2).

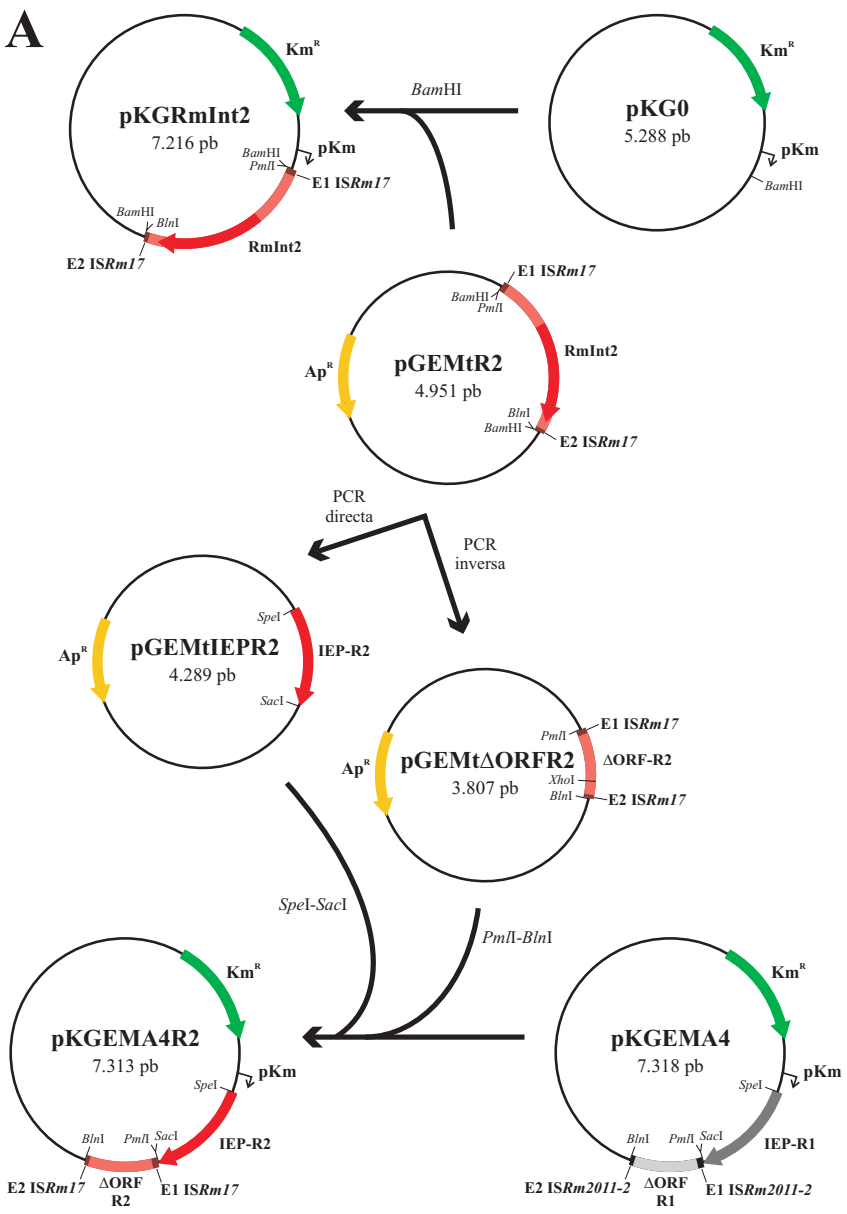
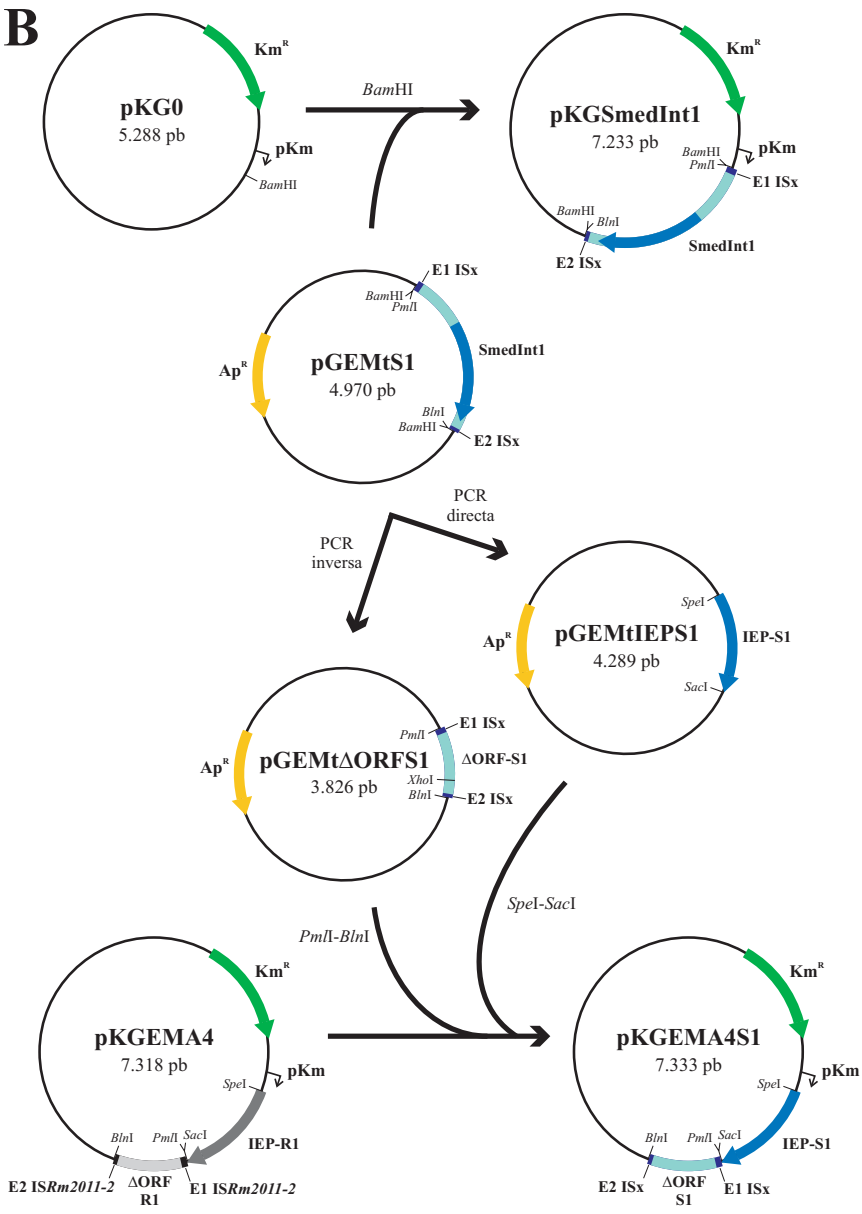


Figura M.1: **Construcción de plásmidos donadores de intrón en su forma silvestre y derivada.** El esquema muestra el proceso de creación de los plásmidos correspondientes a RmInt2 (A) y SmedInt1 (B). Se indican las enzimas de restricción implicadas en cada paso, así como el nombre y la longitud de las construcciones. Las partes ►



◄ de intrón (proteína y ribozima) e IS (exones 1 y 2) incluidas en estos plásmidos se muestran en diferentes tonos de rojo para RmInt2, de azul para SmedInt1 y de gris para RmInt1. El casete de resistencia a ampicilina aparece en amarillo, y el de kanamicina en verde. El promotor de kanamicina (pKm) se indica con una flecha.

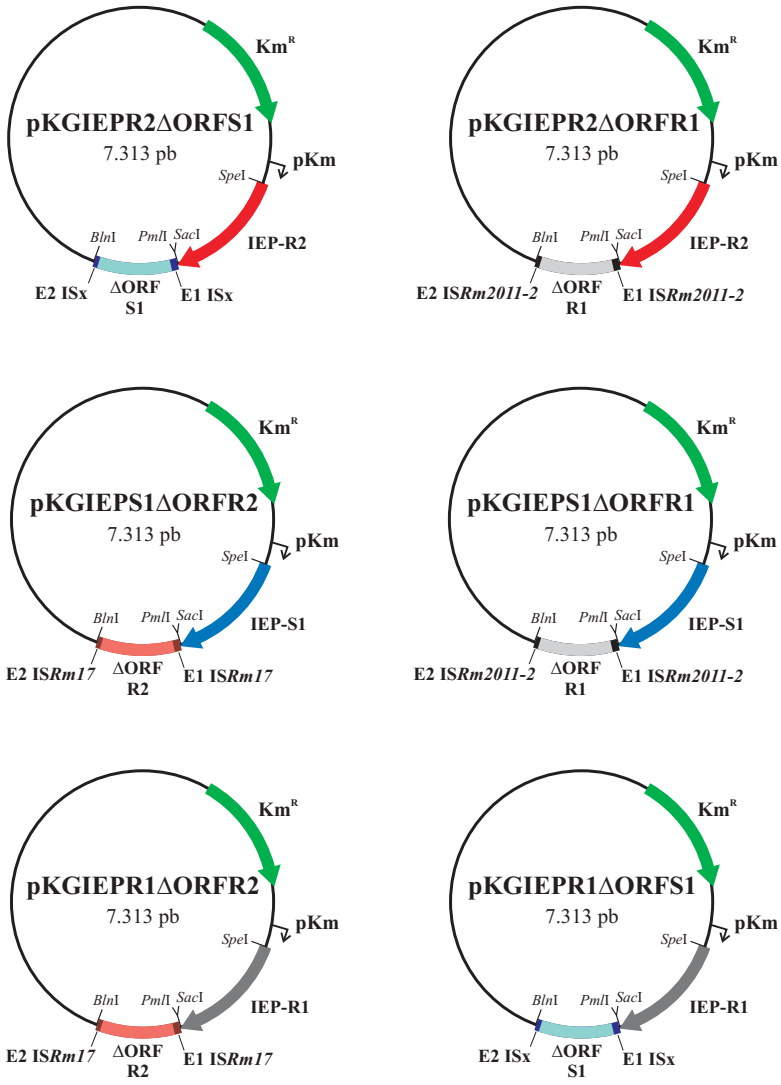


Figura M.2: **Plásmidos donadores de intrones quiméricos.** En el esquema se indican las enzimas de restricción implicadas en el intercambio de la proteína, por un lado, y la ribozima e IS (exones 1 y 2) por otro, mostradas en diferentes tonos de rojo para RmInt2, de azul para SmedInt1 y de gris para RmInt1. También se señala el nombre y la longitud de las construcciones resultantes. El casete de resistencia a kanamicina aparece en verde, y el promotor de kanamicina (pKm) se indica con una flecha.

M.2.2 Plásmidos receptores de intrón

Con el fin de evaluar la movilidad de los intrones estudiados en la presente Tesis Doctoral llevamos a cabo la construcción de una serie de plásmidos que portan la secuencia de reconocimiento de cada intrón. Para ello, partimos del plásmido que contiene la diana de RmInt1 pJB0.6LEAD (Martínez-Abarca *et al*, 2000), donde eliminamos la diana de dicho intrón mediante digestión *Spe*I-*Hind*III, enromamos los extremos del vector (apartado M.5.2) y lo autoligamos (apartado M.5.3 y M.5.5), dando lugar al plásmido pJBΔ0.6 (figura M.3A). Esta construcción es la base de todos los plásmidos receptores, que se generan por clonación de la diana de reconocimiento de cada intrón en el sitio *Not*I de ésta en ambas orientaciones, LAG y LEAD. RmInt1 presenta una tendencia de invasión a dianas localizadas en la secuencia que sirve como molde de la cadena retrasada en la horquilla de replicación (Martínez-Abarca *et al*, 2004). Por ello, es relevante diferenciar cada diana dependiendo de su posición dentro del replicón, pudiendo localizarse en la hebra usada como molde para la síntesis de la cadena retrasada o líder en la horquilla de replicación, nombradas a partir de ahora como LAG (del inglés *lagging*) o LEAD (del inglés *leading*) respectivamente.

El fragmento con la secuencia de la diana, que contiene los últimos 20 nucleótidos del exón 1 y los 5 primeros nucleótidos del exón 2 (diana -20/+5 respecto al sitio de inserción del intrón), se obtuvo de dos maneras distintas dependiendo de la construcción. En el caso de la diana IS*Sme3*, ésta se obtuvo mediante PCR con la DNA polimerasa *Phusion High-Fidelity* (Finnzymes; apartado M.6.1) sobre DNA genómico de WSM419 con los oligonucleótidos 1F_SmedInt1_IS66 y 1R_SmedInt1_IS66 (tabla M.3). El producto de PCR, de 70 pb, se introdujo en el vector pGEM-T Easy, se secuenció con el cebador universal Sp6 y se purificó tras su digestión con la enzima de restricción *Not*I (figura M.3B). El resto de dianas se generaron por anillamiento de los dos oligonucleótidos correspondientes en cada caso: ISRm17_F con ISRm17_R para crear la diana de RmInt2 (figura M.3C); Quimera-20/+5_F con Quimera-20/+5_R y Quimera-21/+5_F con Quimera-21/+5_R para construir las dianas quiméricas entre la zona distal de los exones pertenecientes

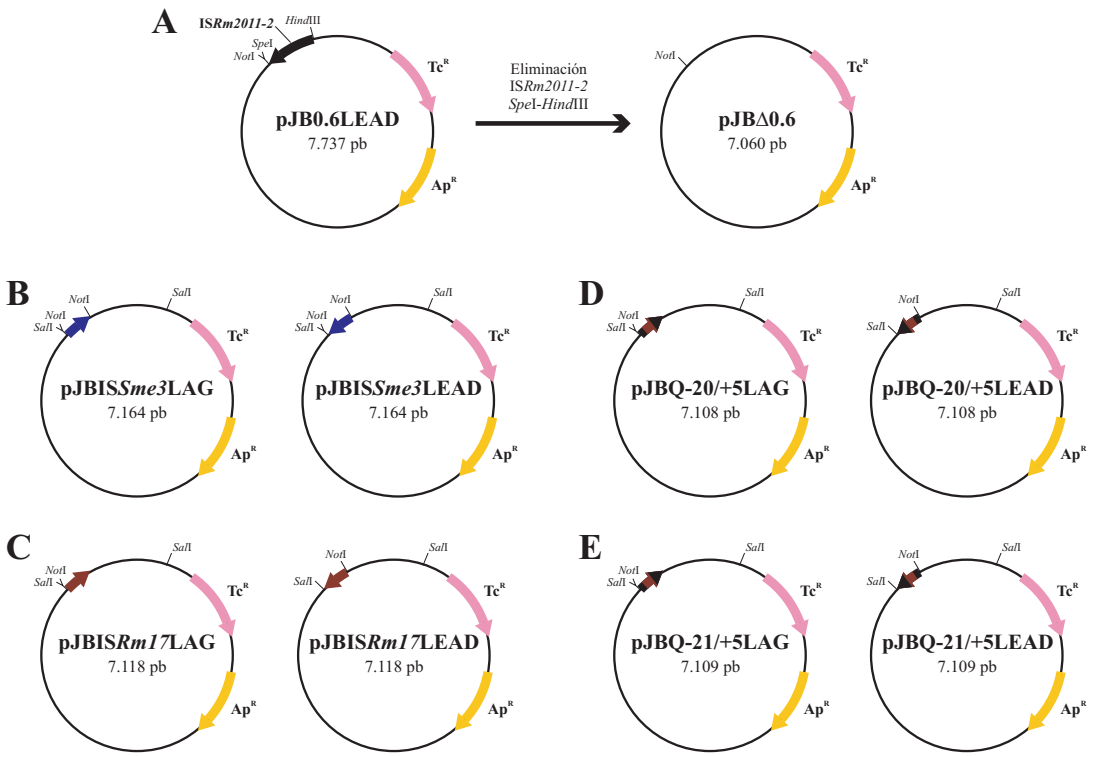


Figura M.3: **Construcción de plásmidos receptores de intrón.** (A) Construcción del plásmido base sobre el que se produce la clonación de las diferentes dianas: IS*Sme3* (B), IS*Rm17* (C) y las dianas quiméricas con IS*Rm2011-2* e IS*Rm17* (D y E). Se indican las enzimas de restricción implicadas en cada proceso, así como para evaluar los eventos de movilidad (apartado M.10). También se señala el nombre y longitud de las construcciones resultantes. El casete de resistencia a ampicilina aparece en amarillo, y el de tetraciclina en rosa.

a la diana IS*Rm2011-2* y las IBSs de la diana IS*Rm17* (figura M.3D y M.3E). Los fragmentos de DNA conteniendo estas tres últimas dianas se introdujeron directamente en el plásmido pJBΔ0.6 y se secuenciaron con unos cebadores específicos de los plásmidos pJBs que flanquean el sitio de clonación múltiple (pJB_secF y pJB_secR; tabla M.3). Los oligonucleótidos de la diana IS*Rm17* y las dianas quiméricas se diseñaron con extremos cohesivos para la diana *Not*I de tal forma que el extremo 3′ del fragmento generado tras

su anillamiento no mantuviera el sitio de corte para dicha enzima. Con esta estrategia se facilitó la identificación de la orientación de la diana una vez clonada en pJBΔ0.6. En los distintos plásmidos presentes en la figura M3 también se señala el sitio de corte de la enzima *Sal*I, usada para evaluar los eventos de movilidad de los intrones (apartado M.10).

M.2.3 Plásmidos para ensayos de complementación

La construcción de los plásmidos que contienen el operón *groESL3* en ambas orientaciones, sentido (pJBGroESL3s) y antisentido (pJBGroESL3as), se realizó a partir del plásmido pJBΔ129, que presenta los casetes de resistencia a Tc y Ap (figura M4). Llevamos a cabo una PCR con la DNA polimerasa

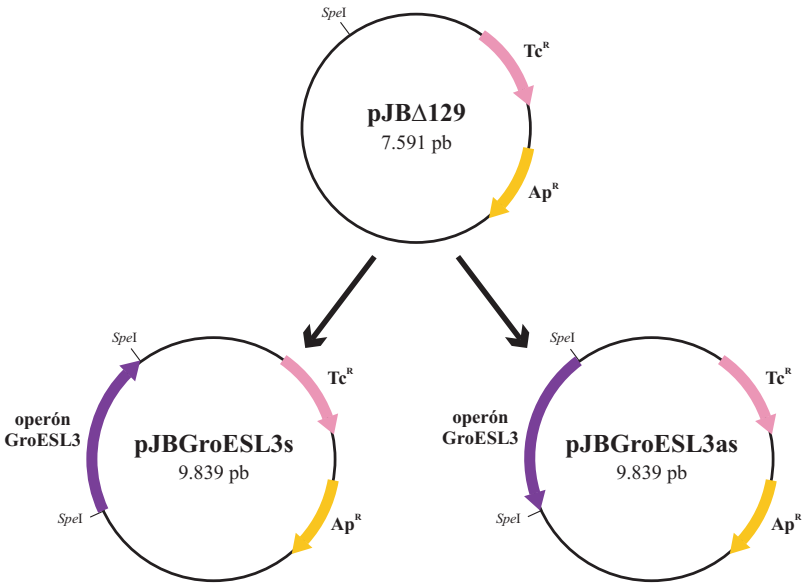


Figura M.4: **Construcción de plásmidos para la complementación de mutantes *groEL* de 1021.** En el esquema se muestran las enzimas de restricción implicadas en la creación de los plásmidos, así como el nombre y la longitud de las construcciones. El casete de resistencia a ampicilina aparece en amarillo, el de tetraciclina en rosa, y el operón *groESL3* en morado.

Phusion High-Fidelity (Finnzymes; apartado M.6.1) sobre DNA genómico de GR4 con los oligonucleótidos GroESL3_F y GroESL3_R, que contienen en su extremo 5' el sitio de corte de la enzima *Spe*I. El fragmento amplificado se introdujo en el vector pGEM-T Easy, donde se secuenció con los cebadores universales T7 y Sp6 y dos oligonucleótidos localizados dentro del operón (GroESL3_sec_F y GroESL3_sec_R; tabla M.3) para confirmar la ausencia de mutaciones, y se purificó tras su digestión con *Spe*I. Posteriormente se clonó en el sitio *Spe*I del plásmido pJBΔ129 y se seleccionaron colonias que portaban el operón *groESL3* en ambas orientaciones (figura M.4). El fragmento clonado contiene los genes *groES3* y *groEL3*, y 120 nt de secuencia aguas arriba y abajo, incluyendo así su promotor y terminador.

M.3 CULTIVOS BACTERIANOS

M.3.1 Medios de cultivo

Las cepas de *S. meliloti* se crecieron en medio TY (Beringer, 1974): 0'9 g/l de $CaCl_2 \cdot 2H_2O$, 5 g/l de triptona y 3 g/l de extracto de levadura, preparado en agua desionizada y a pH 7, y esterilizado en autoclave durante 20 minutos a 120ºC. Para obtener medio TY sólido se añadió agar (PANREAC) al 1'6 % antes de la esterilización.

Para el crecimiento de las cepas de *E. coli* se empleó el medio de cultivo descrito por Luria-Bertani (LB; Sambrook *et al*, 1989): 5 g/l de NaCl, 10 g/l de triptona y 5 g/l de extracto de levadura, preparado en agua desionizada y a pH 7, y esterilizado en autoclave durante 20 minutos a 120ºC. Para obtener medio LB sólido se añadió agar (PANREAC) al 1'6 % antes de la esterilización.

Los mutantes proporcionados por la Dra. Valerie Oke se crecieron siguiendo las instrucciones recomendadas (Bittner *et al*, 2007), en medio LB complementado con $CaCl_2$ y $MgSO_4$, ambos a una concentración final 2'5mM.

Los transconjugantes de *S. meliloti* y de los mutantes se crecieron en una modificación del medio mínimo (MM; Robertsen *et al*, 1981): 0'3 g/l de K_2HPO_4, 0'3 g/l de KH_2PO_4, 0'15 g/l de $MgSO_4 \cdot 7\ H_2O$, 0'05 g/l de $CaCl_2$, 0'006 g/l de $FeCl_3$, 0'05 g/l de NaCl, 1'1 g/l de glutamato sódico, 10 g/l de manitol, 0'0002 g/l de biotina y 0'0001 g/l de pantotenato cálcico, preparado en agua desionizada y a pH 7, y esterilizado en autoclave durante 20 minutos a 120ºC. Para obtener MM sólido se añadió agar (PANREAC) al 1'6 % antes de la esterilización.

Para comprobar que las colonias de transconjugantes seleccionadas con MM no se encontraban contaminadas con *E. coli* se utilizó un medio restrictivo para coliformes. Éste se preparó disolviendo 51 g de Endo Agar (DIFCO) por litro de agua desionizada, calentando hasta su disolución completa y esterilizando en autoclave durante 20 minutos a 120ºC.

M.3.2 Antibióticos

Los antibióticos usados en este trabajo se detallan en la tabla M.4. La adición de éstos a los medios de cultivo se hizo a partir de soluciones 100 veces concentradas preparadas en agua desionizada (todos salvo Cm) o en agua-etanol al 50 % (Cm). Las soluciones preparadas en agua se esterilizaron

Tabla M.4: Antibióticos utilizados para el crecimiento bacteriano.

Antibiótico	**Concentración final (mg/l)**		
	E. coli	***S. meliloti***	**Mutantes GroEL**
Ampicilina (Ap, Sigma)	200	200	200
Tetraciclina (Tc, Sigma)	10	10	2
Estreptomicina (Sm, Sigma)	50	250	250
Kanamicina (Km, Roche)	50	180	180
Gentamicina (Gm, Sigma)	10	50	25
Cloranfenicol (Cm, Sigma)	50	50	50

por filtración utilizando unidades Minisart®NML (Sartorius) de 0'2 μm de tamaño de poro.

M.3.3 Condiciones de cultivo y conservación de éstos

El crecimiento de las cepas de *S. meliloti* se realizó a 30ºC, mientras que las cepas de *E. coli* se incubaron a 37ºC. El tiempo de generación estimado en medio líquido a 190 rpm en agitador orbital es de aproximadamente 2'5 h y 30 min para las cepas de *S. meliloti* y *E. coli* respectivamente.

Para la conservación prolongada de los cultivos bacterianos, éstos se crecieron hasta fase logarítmica tardía, se mezclaron con glicerol 20 % (v/v) estéril en criotubos y se almacenaron a -80ºC.

M.4 AISLAMIENTO DE ÁCIDOS NUCLÉICOS

M.4.1 Extracción de DNA total

Para aislar DNA total se utilizó el kit comercial *RealPure Genomic DNA Extraction* (GE Healthcare). Se recoge 1 ml de cultivo y se centrifuga a 13.000 rpm durante 2 min. Se lava con 0'2 ml de sarcosil al 0'1 % en TE 1x y se centrifuga a 13.000 rpm durante 4 min. Para la lisis celular se añade 0'6 ml de *Lysis Solution* y se incuba 10 min a 80ºC. Se deja enfriar a temperatura ambiente y se añade 3 μl de *Solución RNAse*. Se mezcla invirtiendo 25 veces y se incuba 1 h a 37ºC. Pasado este tiempo, se adicionan 0'2 ml de *Protein Precipitation*, se agita vigorosamente durante 20 seg y se incuba en hielo 10 min. La muestra se centrifuga a 13.000 rpm durante 5 min y se recoge el sobrenadante en un tubo nuevo. Para precipitar el DNA se añade 0'6 ml de isopropanol frío (guardado a -20ºC), se mezcla totalmente invirtiendo y se mantiene durante 10 min a -20ºC. Se centrifuga 10 min a 13.000 rpm y se lava el sedimento con 0'2 ml de etanol 70 %. Finalmente, se resuspende con 50 μl de agua bidestilada y se conserva a 4ºC.

M.4.2 Extracción de DNA plasmídico

M.4.2.1 *Extracción de DNA plasmídico mediante precipitación con sales de magnesio*

Este método, usado para la obtención de DNA plasmídico de muestras de *E. coli*, se basa en el método descrito por Studier (1991). Se recogen 1'5 ml de cultivo y se centrifugan a 13.000 rpm durante 2 min. Tras descartar el sobrenadante, se añaden 100 µl de agua desionizada y se resuspende el sedimento con vórtex. Se adicionan 100 µl por tubo de una mezcla de: NaOH 0'1 M, EDTA 10 mM y SDS al 2 %. Se homogeneiza la muestra por agitación y se hierve durante 2 min. Se añaden 50 µl de MgCl2 1M y se agita vigorosamente. Se mantiene 2 min en hielo y se centrifuga 1 min a 13.000 rpm. Se añaden 50 µl de acetato potásico 5 M pH 4'8 y se mezcla en vórtex bocabajo para evitar que se levante el sedimento. Se deja 2 min en hielo y se centrifuga 5 min a 13.000 rpm. Se pasa el sobrenadante a un tubo nuevo con 600 µl de etanol 100 % frío (guardado a -20ºC), se mantiene durante 5 min a temperatura ambiente y se centrifuga 5 min a 13.000 rpm. Se tira el sobrenadante, se añaden 200 µl de etanol 70 % frío (guardado a -20ºC) y se centrigufa 5 min a 13.000 rpm. Se descarta el sobrenadante y el sedimento se seca en bomba de vacío, estufa o termobloc a 37ºC durante 10-15 min. Finalmente, se resuspende en 20-30 µl de una disolución de RNAsa 10 µg/ml con agua desionizada y se conserva a -20ºC.

M.4.2.2 *Extracción de DNA plasmídico mediante lisis alcalina*

Esta técnica, utilizada para la extracción de DNA plasmídico de muestras de *S. meliloti*, es una modificación del método descrito por Sambrook *et al* (1989). Se recogen 1'5 ml de cultivo líquido y se centrifugan a 13.000 rpm durante 2 min. El sedimento se resuspende en 200 µl de sarcosil al 0'1 % en TE 1x y se centrifuga a 13.000 rpm durante 4 min. Se descarta el sobrenadante y se resuspende el sedimento en 100 µl de lisozima fresca (guardada a 4ºC) a una concentración de 4 mg/ml más una solución

de: glucosa 50 mM, Tris-HCl 25 mM pH 8 y EDTA 10 mM. Se mantiene 5 min a temperatura ambiente y luego se añaden 200 µl de una solución de: NaOH 0'2M y SDS al 1 %. Se deja 5 min en hielo y se añaden 150 µl de acetato potásico 5 M pH 4'8. Se deja 5 min en hielo y se centrifuga a 13.000 rpm durante 5 min. Se pasa el sobrenadante (~400 µl) a un tubo nuevo y se añade un volumen de una solución de fenol-cloroformo-isoamílico 25:24:1. Se mezcla con vórtex y se separan las dos fases por centrifugación a 13.000 rpm durante 5 min. Se lleva el sobrenadante a un tubo nuevo y se añaden 300 µl de cloroformo-isoamílico 24:1. Se mezcla con vórtex y se centrifuga a 13.000 rpm durante 5 minutos. Se pasa el sobrenadante a un tubo nuevo y se añaden 2 volúmenes de etanol 100 % frío (guardado a -20ºC). Se mezcla con vórtex, se deja precipitando 5 min a temperatura ambiente y se centrifuga durante 15 min a 13.000 rpm. Se elimina el sobrenadante y se añaden 200 µl de etanol 70 % frío (guardado a -20ºC). Se centrifuga durante 5 min a 13.000 rpm y se descarta el sobrenadante. El sedimento se seca en bomba de vacío, estufa o termobloc a 37ºC durante 10-15 min. Finalmente, se resuspende en 20-30 µl de una solución de RNAsa 10 µg/ml con agua desionizada y se conserva a -20ºC.

M.4.2.3 *Extracción de DNA plasmídico mediante kit comercial*

En determinados casos, se extrajo DNA plasmídico a partir del kit comercial *Illustra plasmidPrep Mini Spin Kit* (GE Healthcare). La ventaja fundamental de usar este kit, frente a los dos métodos anteriores, es la limpieza del DNA y la ausencia de RNAsa en los pasos finales. Se recogen 1'5 ml de cultivo y se centrifugan a 13.000 rpm durante 30 seg. Para la lisis celular se resuspende con vórtex el sedimento en 175 µl de *Lysis buffer type 7*. Se añaden 175 µl de *Lysis buffer type 8* y se mezcla por inversión hasta conseguir una solución clara y viscosa. La muestra se neutraliza añadiendo 350 µl de *Lysis buffer type 9* y mezclando por inversión. Se centrifuga a 13.000 rpm durante 4 min y se pasa el sobrenadante a un tubo nuevo al que previamente se le ha colocado una columna del kit. Se centrifuga a 13.000 rpm durante 30 seg

y se tira el sobrenadante. Se lava el DNA unido a la resina de la columna con 400 µl de *Wash buffer type 1* y se centrifuga a 13.000 rpm durante 1 min. Para eluir el DNA se transfiere la columna a un tubo nuevo y se añaden entre 20-100 µl de *Elution buffer type 4* según la concentración de DNA que se requiera. Se incuba durante 30 seg a temperatura ambiente y se centrifuga a 13.000 rpm durante 30 seg. Para conservar la muestra se guarda a -20ºC.

M.4.3 Extracción de RNA total

Este método, usado para la obtención de RNA total de muestras de *S. meliloti* y *E. coli*, se basa en el método descrito por Cabanes *et al* (2000). Se lleva a cabo la inoculación de 0'5 ml del cultivo bacteriano saturado en un tubo que contiene 9'5 ml de medio con el correspondiente antibiótico y se crece hasta el final de su fase exponencial (~0,6 de A_{600}). Se recogen los 10 ml de cultivo en dos tubos de 1'5 ml mediante varias centrifugaciones de 1-2 min a 13.000 rpm. Se resuspenden las células sedimentadas con 300 µl de una solución precalentada compuesta por: SDS al 1'4 %, EDTA 4 mM y 0'4 mg/ml de proteinasa K. Se incuba durante 10 min a 65ºC, agitando de manera esporádica (cada 3 min). Pasado este tiempo, los tubos se ponen en hielo y se adicionan 150 µl de NaCl 5 M (guardado a 4ºC). Se mezcla vigorosamente, se mantiene en hielo durante 10 min y se centrifuga 15 min a 4ºC y 13.000 rpm. Se pasa el sobrenadante a un tubo nuevo y se precipita con 1 ml de etanol 100 % frío (guardado a -20ºC). El ácido nucléico precipitado se centrifuga durante 30 min a 4ºC y 13.000 rpm. El sedimento correspondiente a los 10 ml de cultivo inicial se resuspende en 85 µl de agua bidestilada y se incuba durante 1 h a 37ºC con 50 U de DNAsa I libre de RNAsas (Roche). Se añade un volumen (100 µl) de una solución de fenol-cloroformo-isoamílico 25:24:1 y se centrifuga 10 min a 4ºC y 13.000 rpm. Se pasa la fase superior a un tubo nuevo, se mezcla con un volumen (100 µl) de una solución de cloroformo-isoamílico 24:1 y se centrifuga 5 min a 4ºC y 13.000 rpm. El sobrenadante se pasa a un tubo nuevo con 600 µl de etanol 100 % frío (guardado a -20ºC) y 75 mM de acetato sódico 3 M pH 5'2 frío

(guardado a 4ºC). Se mantiene al menos 1 h a -80ºC, se centrifuga durante 30 min a 4ºC y 13.000 rpm, y se lava con 0'5 ml de etanol 70 % frío (guardado a -20ºC). El sedimento se seca a temperatura ambiente y se resuspende en 20 µl de agua bidestilada. La concentración de RNA se determinó mediante el uso de un espectrofotómetro (NanoDrop®ND-1000).

M.5 MANIPULACIÓN DE DNA

M.5.1 Digestión de DNA con enzimas de restricción

La digestión de DNA se realizó siguiendo las indicaciones de temperatura y tampón óptimos recomendados por los proveedores (New England Biolabs y Roche). Generalmente se usó 5 U de enzima por microgramo de DNA plasmídico, mientras que para la digestión de DNA total se usó 10 U de enzima por microgramo. El tiempo de incubación para una digestión completa del DNA varió desde 1 h, en el caso de DNA plasmídico, hasta 18 h, para DNA total. En las incubaciones prolongadas se añadieron 5-10 U más de enzima dependiendo de la eficiencia de ésta y la cantidad de DNA aún no digerido en la muestra. En el caso de las digestiones dobles, éstas se realizaron habitualmente de forma simultánea con las dos enzimas, eligiendo el tampón óptimo para ambas. En caso de incompatibilidad de tampones o temperatura, se realizaron digestiones sucesivas, comenzando con la enzima que requería un tampón de menor fuerza iónica o menor temperatura. Posteriormente, el DNA se resolvió en geles de agarosa (apartado M.7.1) o se purificó mediante fenolización y precipitación (apartado M.5.4).

M.5.2 Conversión de fragmentos de DNA con extremos protuberantes a extremos romos

La técnica usada para convertir un extremo 5' protuberante en un extremo romo se basa en un método descrito por Wartell & Reznikoff (1980) y modificado por Sambrook *et al* (1989). El extremo 5' protuberante se rellena

usando la actividad DNA polimerasa del fragmento de *Klenow* de la DNA polimerasa I de *E. coli*. Se incuba 0'5-1 μg del fragmento de DNA durante 15-30 min a 37ºC con una solución compuesta por: tampón 1x (formado por Tris pH 7'5 500 mM, MgCl2 100 mM, DTT 10 mM y BSA 100 μg/ml), cada dNTP 1 mM y 2 U de la enzima *Klenow* (Roche), en un volumen final de 10 μl. Para inactivar la enzima, se mantiene 10 min a 65ºC. Posteriormente, el DNA se purifica mediante fenolización y precipitación (apartado M.5.4).

Para la conversión de un extremo 3' protuberante en un extremo romo se utiliza la actividad exonucleasa 3'-5' de la DNA polimerasa del bacteriófago T4 (Sambrook *et al*, 1989). Se incuba 0'5-1 μg del fragmento de DNA durante 5 min a 37ºC con una solución compuesta por: tampón 1x (formado por Tris-HCl 10 mM, NaCl 50 mM, MgCl2 10 mM, DTT 1 mM y BSA 100 μg/ml), cada dNTP 100 μM y 1 U de T4 DNA polimerasa (New England Biolabs). La reacción se para con EDTA 10 mM, y el DNA se purifica mediante fenolización y precipitación (apartado M.5.4).

M.5.3 Defosforilación de fragmentos de DNA

Para evitar la atuoligación de vectores linearizados, se eliminaron los grupos fosfato presentes en los extremos 5' de éstos. Para ello, se incuban 5-10 μg de DNA digerido con 2 U de fosfatasa alcalina (CIP, *Calf Intestine Phosphatase*; Roche) y su tampón 1x durante 30 min a 37ºC. Se añaden otras 2 U de CIP y se mantiene 30 min a 37ºC. La reacción se para añadiendo 55 μl de una solución compuesta por tampón STE 1x (formado por Tris-HCl 1 mM pH 8, NaCl 100 mM y EDTA 1 mM) y SDS al 1 % e incubando 15 min a 68ºC. Posteriormente, el DNA se purifica mediante fenolización y precipitación (apartado M.5.4).

M.5.4 Purificación de fragmentos de DNA

La purificación de fragmentos de DNA procedentes de reacciones enzimáticas y reacciones de PCR se realizó mediante extracción con fenol-cloroformo o con el kit comercial *Illustra GFX PCR DNA and Gel Band Purification Kit* (GE Healthcare). Este mismo kit se utilizó para aislar fragmentos de DNA separados por electroforesis en geles de agarosa, siguiendo el protocolo adecuado en cada caso indicado por el fabricante.

En el proceso de fenolización, al DNA se le añade un volumen de una solución de fenol-cloroformo-isoamílico 25:24:1 y se centrifuga 5 min a 13.000 rpm. El sobrenadante se lleva a un tubo nuevo, se añade una solución de cloroformo-isoamílico 24:1 y se centrifuga 5 min a 13.000 rpm. El sobrenadante se pasa a un tubo nuevo y se añaden 3 volúmenes de etanol 100 % frío (guardado a -20ºC) y 0'1 volumen de acetato sódico pH 5'2 3 M. Se mezcla manualmente y se mantiene 5 min a temperatura ambiente. El DNA se precipita durante al menos 1 h a -80ºC (o toda la noche a -20ºC) y se centrifuga 15 min a 13.000 rpm. Se añaden 200 µl de etanol 70 % frío (guardado a -20ºC) y se centrifuga 5 min a 13.000 rpm. El sedimento se seca en bomba de vacío, estufa o termobloc a 37ºC durante 10-15 min. Por último, se resuspende en el volumen deseado de agua desionizada y se conserva a -20ºC.

M.5.5 Ligación de fragmentos de DNA

Las reacciones en las que se ligaron varios fragmentos de DNA digeridos o productos de PCR se realizaron siguiendo condiciones y relaciones molares vector:inserto variables, dependiendo del tamaño y cantidad de los fragmentos participantes. La proporción molar generalmente usada fue 1:3 (vector:inserto). Para llevar a cabo autoligaciones de un vector, el DNA se diluyó 10 veces con agua desionizada para disminuir su concentración y evitar construcciones formadas por la ligación de dos vectores. En todas estas reacciones, la ligación se realizó con la enzima T4 DNA ligasa (Roche) en un volumen final de 10-20 µl a 14-16ºC durante al menos 16 h.

M.6 AMPLIFICACIÓN DE DNA

M.6.1 Reacción en cadena de la polimerasa (PCR)

Las reacciones de PCR rutinarias, usadas para comprobar la presencia de determinados insertos, se llevaron a cabo con una *Taq* DNA polimerasa purificada en el laboratorio. Éstas se realizaron en un volumen final de 25 µl, incluyendo 10-100 ng de DNA plasmídico, 125 µM de dNTPs, 0'2 µM de cada oligonucleótido, 1x del tampón de reacción (formado por Tris-HCl 100 mM pH 8'3, KCl 500 mM y $MgCl_2$ 25 mM), 2 U de DNA polimerasa y agua desionizada hasta completar los 25 µl.

Las reacciones de PCR usadas para corroborar la unión de algunos *contigs* tras la secuenciación de la cepa *S. meliloti* GR4 se llevaron a cabo con la *Taq* DNA polimerasa purificada en el laboratorio y requirieron la adición de DMSO para evitar la formación de estructuras secundarias. Este tipo de PCR se realizó en un volumen final de 25 µl, incluyendo 10-100 ng de DNA total, 200 µM de dNTPs, 0'3 µM de cada oligonucleótido, DMSO al 6 %, 1x del tampón de reacción (formado por Tris-HCl 100 mM pH 8'3, KCl 500 mM y $MgCl_2$ 25 mM), 4 U de DNA polimerasa y agua desionizada hasta completar los 25 µl.

Las reacciones de PCR en las que era necesaria una baja tasa de errores en el producto final se llevaron a cabo con la DNA polimerasa *Phusion High-Fidelity* (Finnzymes). Dependiendo de la finalidad, éstas se realizaron en un volumen final de 25 µl (para posterior ligación del fragmento amplificado) o de 50 µl (para secuenciación directa del fragmento). En un volumen final de 25 µl se incluían 10-100 ng de DNA molde (plasmídico o total), 200 µM de dNTPs, 0'5 µM de cada oligonucleótido, DMSO al 6 %, 1x del tampón de reacción, 0'004 U de DNA polimerasa y agua desionizada hasta completar los 25 µl.

El proceso de amplificación se llevó a cabo en un termociclador *Eppendorf Mastercycler*. Las condiciones de PCR variaron dependiendo de la temperatura de anillamiento de los oligonucleótidos y de la DNA polimerasa

utilizada. En el caso de la *Taq* DNA polimerasa purificada en el laboratorio, una reacción normal consistió en un ciclo inicial de desnaturalización del DNA a 94ºC durante 2 min, seguido de 30-35 ciclos de: desnaturalización 30 s a 94ºC, anillamiento 30 s a 60ºC normalmente y extensión de la polimerasa el tiempo necesario según el tamaño del producto esperado a razón de 1 kb/30 s a 68ºC. Finalmente, un ciclo adicional de extensión durante 5 min a 68ºC.

El programa usado cuando se empleó la DNA polimerasa *Phusion High-Fidelity* (Finnzymes) consistió en un ciclo inicial de desnaturalización del DNA a 98ºC durante 1 min, seguido de 25-30 ciclos de: desnaturalización 10 s a 98ºC, anillamiento 30 s a 60ºC normalmente y extensión de la polimerasa el tiempo necesario según el tamaño del producto esperado a razón de 1 kb/15-30 s a 72ºC. Finalmente, un ciclo adicional de extensión durante 5 min a 72ºC.

M.6.2 Adenilación de amplificados de PCR

En ocasiones, el producto amplificado mediante PCR se utilizó para su clonación en el vector pGEM-T Easy, el cual presenta extremos T protuberantes. Para este proceso es necesario que el producto de PCR contenga extremos A protuberantes. Sin embargo, algunas polimerasas presentan actividad exonucleasa 3'-5' y dejan extremos romos tras la amplificación. Por este motivo, se requiere la adición de adeninas en los extremos de los fragmentos amplificados con la DNA polimerasa *Phusion High-Fidelity* (Finnzymes). Para ello, a 1-7 µl del producto de PCR previamente purificado se le añade una mezcla compuesta por: 1x del tampón de reacción, 5U de la *Taq* DNA polimerasa purificada en el laboratorio, dATP 0'2 mM y agua desionizada hasta un volumen final de 10 µl. Se incuba durante 3 min a 70ºC, y el producto se utiliza directamente en la reacción de ligación (apartado M.5.5).

M.6.3 PCR de extensión por solapamiento (OE-PCR)

Esta técnica es una variante de la PCR usada para insertar mutaciones específicas en determinados puntos de una secuencia, gracias al solapamiento de dos oligonucleótidos (los internos) en el DNA molde, al realizar dos PCRs consecutivas. En la presente Tesis Doctoral se ha utilizado la OE-PCR (*overlap extension polymerase chain reaction*) para generar un plásmido donador de intrón que contiene una proteína quimérica entre los intrones RmInt1 y RmInt2 seguida del ΔORF de RmInt2 (figura M.5). Para su construcción, primero se realizaron dos PCRs independientes con las condiciones de PCR indicadas en el apartado anterior para la DNA polimerasa *Phusion High-Fidelity* (Finnzymes): una con los cebadores EB70 (5'-ATGGTGGTCAA GCAGATGA-3') y CtailR2_R (5'-CGACCAGCTTTCGCAAGAAGAGGCTG GCGCGT-3') sobre DNA extraído del plásmido pKGEMA4, donde se amplifica la IEP de RmInt1 desde el aminoácido 94 hasta el 399 (fragmento de 918 pb); y otra con los cebadores CtailR2_F (5'-CTCTTCTTGCGAAAG CTGGTCGAGCAGCGCACCGATCTCT-3') y 105mer (5'-CCGCTCCTGAA AGCCGAT-3') sobre DNA del plásmido pKGEMA4R2, donde la amplificación comprende la región C-terminal de la proteína de RmInt2 desde el aminoácido 400, y llega hasta el nucleótido 105 de la ribozima de RmInt2 (fragmento de 223 pb; figura M.5). Estos dos fragmentos amplificados se limpiaron (apartado M.5.4) y se mezclaron en proporción 1:1. Luego se hizo una dilución 1:10 de dicha mezcla y se cogió 1 μl como DNA molde para realizar la segunda PCR con los cebadores externos, EB70 y 105mer. Así se obtuvo un producto de PCR (de 1.113 pb) que contiene la parte inicial de la proteína de RmInt1 seguida de la zona C-terminal de la IEP de RmInt2. Posteriormente, se llevó a cabo una digestión con las enzimas de restricción *Eco*RI y *Sac*I tanto de este fragmento como del vector quimérico pKGIEPR1ΔORFR2 para producir el intercambio de la IEP de RmInt1 presente en ese plásmido por la nueva proteína quimérica R1-R2 y generar así el plásmido pKGIEPQΔORFR2 (figura M.5).

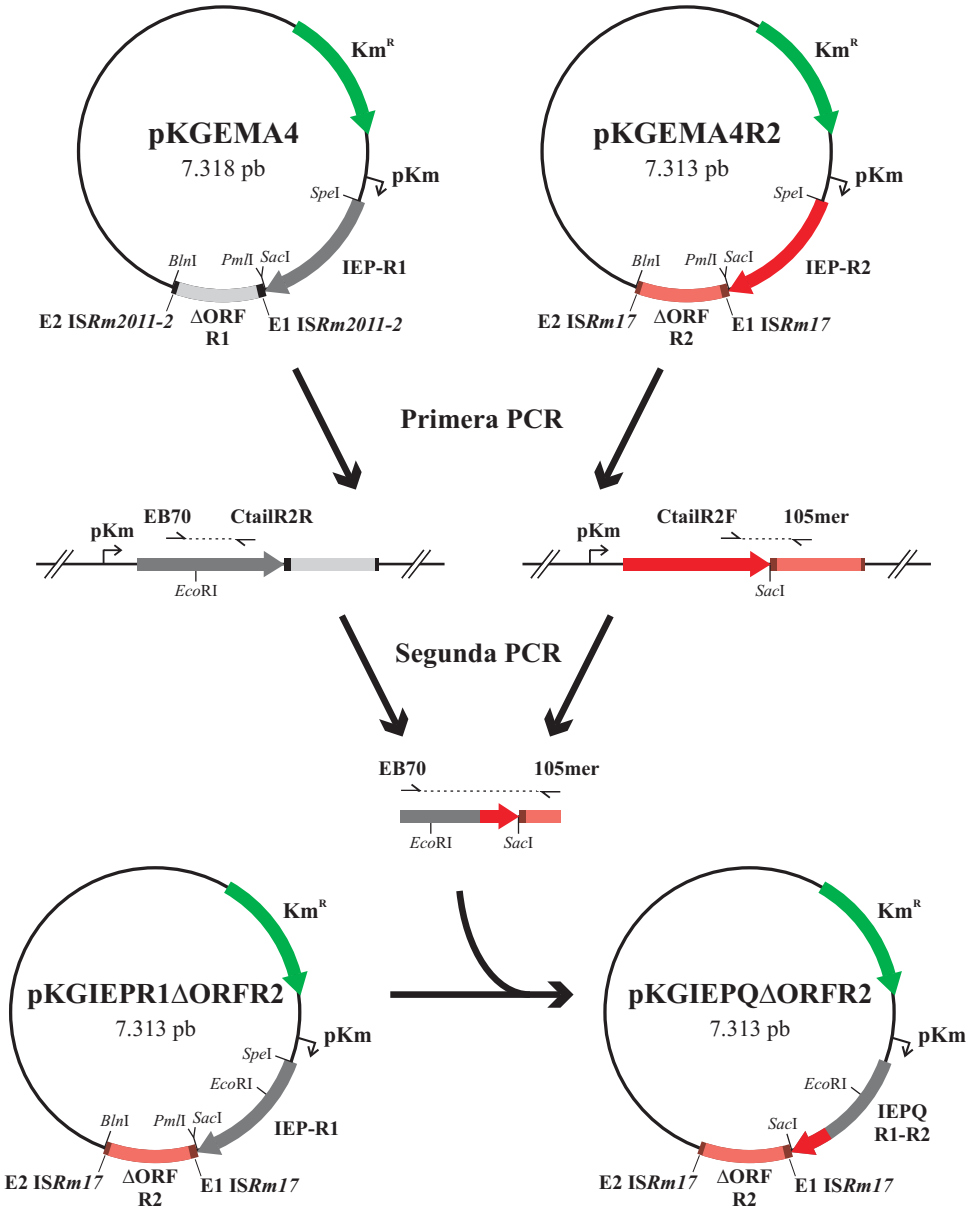


Figura M.5: **Construcción del plásmido donador de intrón con IEP quimérica.** En el esquema se muestran los plásmidos utilizados como molde para llevar a cabo la primera reacción de OE-PCR. Para la segunda reacción se utiliza una mezcla de los productos amplificados en la primera. También se indican las enzimas de restricción implicadas en la construcción del plásmido pKGIEPQΔORFR2.

M.6.4 PCR cuantitativa a partir de RNA (qRT-PCR)

La PCR cuantitativa o a tiempo real (qPCR) se trata de una variante de la PCR convencional que permite cuantificar la cantidad inicial de DNA o RNA contenida en la muestra (Higuchi *et al*, 1993). Para cuantificar muestras de RNA, la qPCR se tiene que realizar con cDNA, sintetizado a partir del RNA de las muestras estudiadas. Esta técnica se denomina PCR cuantitativa con transcripción reversa (qRT-PCR). La qRT-PCR ha sido usada en este trabajo para determinar la cantidad inicial de RNA presente en la cepa 1021 y en los mutantes de ésta para los genes *groEL* cuando se les introdujo el plásmido donador de intrón pCm4.

La síntesis de cDNA consta de una primera etapa de anillamiento del RNA con oligonucleótidos aleatorios, seguida de la formación del cDNA mediante extensión de estos oligonucleótidos con la reverso transcriptasa. En nuestro caso, partimos de 10 μg de RNA total, que se anilló con 50 ng de hexámeros de secuencia aleatoria en presencia de tampón 1x de AMV RT y DDT 1 mM. Tras calentar esta mezcla a 90ºC durante 2 min, se incubó a periodos de 1 min a temperaturas decrecientes en 5ºC hasta alcanzar los 25ºC. En este momento se añadió la mezcla de extensión, compuesta por dNTPs 0'625 mM cada uno, 15 U de inhibidor de RNAsas y 50 U de AMV RT (Roche), y se incubó 1 h a 42ºC. Después se añadieron otras 50 U de AMV RT y se incubó 2 h a 42ºC. Para la inactivación de la reverso transcriptasa se dejó a 85ºC durante 15 min.

La reacción de qRT-PCR se llevó a cabo en placas de 96 pocillos en el termociclador *iCycler iQ System* (Bio-Rad) con los oligonucleótidos E17.0 y 21.0, que amplifican la región 3' de la proteína del intrón RmInt1. A una dilución 1:10 del cDNA previamente sintetizado se le añadió la mezcla de reacción compuesta por: cada uno de los cebadores 0'2 mM, dNTPs 0'1 mM, tampón 1x de la DNA polimerasa, $MgCl_2$ 1'5 mM, una dilución 1:10.000 de Syber Green y 0'5 U de *Taq Platinum* polimerasa (Invitrogen). El programa utilizado para la amplificación consistió en un ciclo inicial de desnaturalización de 3 min a 94ºC, seguido de 40 ciclos de: desnaturalización 30 s a

94ºC, anillamiento 30 s a 68ºC y extensión 15 s a 72ºC; con un ciclo final de extensión de 3 min a 72ºC.

Para generar la curva patrón se utilizó DNA extraído mediante kit comercial (apartado M.4.2.3) del plásmido pCm4. Se pusieron controles negativos sustituyendo el cDNA por agua desionizada. De todas las muestras problema, así como del control negativo, se hicieron tres réplicas.

En este tipo de PCR se utilizó la *Taq Platinum* polimerasa (Invitrogen) para realizar la amplificación debido a que se encuentra formando un complejo con un anticuerpo que impide su funcionalidad hasta que se produce la disociación en el primer ciclo de desnaturalización. Este proceso reduce el fondo en la reacción, que es un factor determinante para la cuantificación.

M.7 ELECTROFORESIS

M.7.1 Electroforesis en geles de agarosa no desnaturalizantes

La resolución de productos de PCR, DNA plasmídico o total, y fragmentos de restricción se llevó a cabo mediante electroforesis en geles de agarosa (*SeaKem LE*, Cambrex/Iberlabo) con un porcentaje que oscila entre el 0'8-2 % (dependiendo de los tamaños a resolver) en TAE 1x (Tris-HCl 40 mM, EDTA Na_2 2 mM y ácido acético glacial 0'1142 % (v/v)). Como tampón de carga se utilizó una solución 6x compuesta por: 0'50 % de naranja G (p/v), EDTA Na_2 0'01 M y 50 % de glicerol. El revelado de este tipo de geles se realizó por inmersión de los mismos en una solución de agua con bromuro de etidio 1 mg/ml o RedGel (Biotium, Inc) al 2 % (v/v) durante 15-20 min. La visualización de los geles se llevó a cabo bajo luz UV en el transiluminador *Gel Doc 2000* (Bio-Rad), con el que se obtuvieron las imágenes para su posterior procesamiento.

M.7.2 Electroforesis en geles de agarosa desnaturalizantes

El análisis del RNA total se realizó mediante electroforesis en geles de agarosa compuestos por agarosa (Roche) al 1'4 %, MOPS 1x (MOPS 4X formado por: MOPS 80 mM, acetato sódico 20 mM y EDTA 4 mM, ajustado a pH 7 con NaOH) y formaldehído al 0'05 % (v/v). El tampón de electroforesis fue MOPS 1x. Un µl de RNA total se mezcló con 4 µl de agua desionizada y 2 µl de una solución compuesta de: 1 % de RedGel, 1'8 % de sacarosa, TBE 1x (Tris-HCl 0'089 M pH 8, ácido bórico 0'089 M y EDTA 0'002 M), 0'018 % de cianol de xileno, 36 % (v/v) de formamida y EDTA 25 mM. La visualización de los geles se llevó a cabo bajo luz UV en el transiluminador *Gel Doc 2000* (Bio-Rad), con el que se obtuvieron las imágenes para su posterior procesamiento.

M.7.3 Electroforesis en geles de poliacrilamida desnaturalizantes

El análisis de RNA y DNA marcado radiactivamente se llevó a cabo mediante geles de poliacrilamida compuestos por bisacrilamida (40 % de Acri/Bis (Amresco)) al 6 %, TBE 1x (Tris-HCl 0'089 M pH 8, ácido bórico 0'089 M y EDTA 0'002 M) y urea 8 M (Roche). Por cada 100 ml de solución de gel se añadieron 40 µl de TEMED (Sigma) y 600 µl de APS (Sigma) al 10 %. El tampón de electroforesis fue TBE 1x. Como tampón de carga se utilizó una solución 2x compuesta de: 1'8 % de sacarosa, TBE 1x (Tris-HCl 0'089 M pH 8, ácido bórico 0'089 M y EDTA 0'002 M), 0'018 % de cianol de xileno, 36 % (v/v) de formamida y EDTA 25 mM. Los geles se secaron a 80ºC y vacío en un *Gel Dryer 583* (Bio-Rad) y se expusieron en pantallas de fósforo (*Imagin Plate 2040*, Fujifilm). Tras 24-48 h, se revelaron con un sistema de escáner láser (*Personal Molecular Imager FX*, Bio-Rad) y se editaron con el programa Quantity One v4.6.2 (Bio-Rad).

M.7.4 Marcadores de peso molecular

Los marcadores de peso molecular utilizados en este trabajo han sido:

- Marcador II: DNA del fago λ digerido con la enzima *Hind*III (Roche). Está compuesto por 8 fragmentos que abarcan desde los 125 pb hasta poco más de 23 Kb. También se dispone marcado con digoxigenina, para marcar el peso molecular en las hibridaciones DNA-DNA (apartado M.11).

- Marcador III: DNA del fago λ digerido con las enzimas *Hind*III y *Eco*RI (Roche). En este caso el patrón de digestión comprende 13 fragmentos que abarcan desde los 125 pb hasta poco más de 21 Kb. También disponible marcado con digoxigenina.

- Marcador ϕ29: DNA del fago ϕ29 digerido con *Hind*III (Universidad Autónoma de Madrid). El patrón de digestión se compone de 14 fragmentos, todos menores de 5 Kb: 72, 156, 273, 453, 579, 611, 759, 1.150, 1.331, 1.933, 2.201, 2.498, 2.899 y 4.370 pb.

- Marcador pGEMT: DNA del plásmido pGEM-T digerido con las enzimas *Hinf*I y *Eco*RI (Promega). Está compuesto de 15 fragmentos que abarcan desde los pocos nucleótidos hasta las 2.500 pb.

M.8 TRANSFORMACIÓN BACTERIANA Y CONJUGACIÓN

M.8.1 Transformación por métodos químicos de células competentes

En este trabajo se han realizado transformaciones mediante choque térmico con la cepa DH5α de *E. coli*. El paso previo a la transferencia de DNA requiere prepararlas como células competentes, empleando para ello un método químico con cloruro de rubidio en condiciones de esterilidad. Las células de *E. coli* se cultivan en 100 ml de medio LB hasta una A_{600} de 0'4 (fase logarítmica). El crecimiento se detiene incubando el cultivo en hielo

durante 15 min y centrifugando 10 min a 4ºC y 6.000 rpm. El sedimento se resuspende en 32 ml de solución RF1 (formada por: RbCl 100 mM, $MnCl_2$ 50 mM, acetato potásico 30 mM, $CaCl_2$ 10 mM y 15 % de glicerol; el pH se ajusta a 5'8 con una solución de ácido acético 0'2 M; se esteriliza por filtración y se conserva a 4ºC). Esta mezcla se incuba en hielo durante 15 min y se centrifuga 10 min a 4ºC y 6.000 rpm. Se descarta el sobrenadante y las células se resuspenden en 4 ml de solución RF2 (formada por: RbCl 10 mM, MOPS 100 mM pH 6'5, $CaCl_2$ 5 mM y 15 % de glicerol; el pH se ajusta a 6'5 con una solución de KOH 1 M; se esteriliza por filtración y se conserva a 4ºC). Se reparten en alícuotas de 100 µl en tubos preenfriados en hielo, se congelan inmediatamente con nitrógeno líquido y se conservan a -80ºC. El rango de eficiencia de transformación de estas células competentes es de aproximadamente 105 células/µg DNA.

Para llevar a cabo el proceso de transformación, las células competentes se sacan del congelador de -80ºC y se ponen en hielo hasta su total descongelación (aproximadamente 20 min). En condiciones de esterilidad se añaden 50-500 ng de DNA plasmídico y se incuban en hielo con las células competentes durante 20-25 min. La muestra se somete a un choque térmico de 42ºC durante 2 min y rápidamente se pone en hielo durante 5 min. Se añaden 900 µl de LB líquido y se incuban 1 h a 37ºC en agitación a 150 rpm. Posteriormente, se siembra toda o parte de la muestra en medio LB sólido con el antibiótico adecuado y se incuba toda la noche a 37ºC.

M.8.2 Electroporación de células electrocompetentes

En este trabajo también se han realizado transformaciones mediante electroporación de células de la cepa *E. coli* DH5α. Todo el proceso de preparación de estas células como electrocompetentes se realiza en condiciones de esterilidad. Las células de *E. coli* se cultivan en 500 ml de medio LB hasta una A_{600} de 0'5. El crecimiento se detiene incubando el cultivo en hielo durante 20 min y centrifugando 15 min a 4ºC y 6.000 rpm. El sedimento se resuspende en 100 ml de glicerol al 10 % frío (guardado a 4ºC) y se centrifuga

15 min a 4ºC y 6.000 rpm. El sedimento se resuspende en 20 ml de glicerol al 10 % frío (guardado a 4ºC) y se centrifuga 15 min a 4ºC y 6.000 rpm. Por último, las células se resuspenden en 2 ml de glicerol al 10 % frío (guardado a 4ºC), se reparten en alícuotas de 50 µl en tubos preenfriados en hielo, se congelan inmediatamente con nitrógeno líquido y se conservan a -80ºC. El rango de eficiencia de transformación de estas células electrocompetentes es de aproximadamente 106 células/µg DNA.

Para llevar a cabo el proceso de electroporación, las células electrocompetentes se sacan del congelador de -80ºC y se ponen en hielo durante 15 min. En condiciones de esterilidad se añaden 50-500 ng de DNA plasmídico previamente dializado para eliminar las sales y se incuban en hielo durante 15 min. La diálisis se realiza con agua destilada utilizando filtros de nitrocelulosa VSWP de 0'025 mm (MILLIPORE). La muestra se transfiere a una cubeta de electroporación y se somete a un pulso eléctrico de 1.800 V durante 3-5 milisegundos en el electroporador *Eppendorf 2510*. Las células se pasan a un tubo nuevo con 900 µl de LB líquido y se incuban 1 h a 37ºC. Posteriormente, se siembra toda o parte de la muestra en medio LB sólido con el antibiótico adecuado y se incuba toda la noche a 37ºC.

M.8.3 Conjugación triparental entre cepas de *E. coli* y *S. meliloti*

La transferencia de DNA plasmídico desde *E. coli* DH5α hacia diferentes cepas de *S. meliloti* se llevó a cabo mediante conjugación triparental (Ditta *et al*, 1980), técnica en la que se hace uso de los genes de movilización presentes en el plásmido pRK2013 (*E. coli* HB101:pRK2013). Las conjugaciones se llevaron a cabo durante 12-16 h en medio TY sólido a 30ºC. Posteriormente, se hizo una extensión de esa mezcla en placas de MM con los antibióticos adecuados y se incubaron a 30ºC hasta la aparición de colonias (2-3 días). Se realizó una nueva selección en TY sólido con colonias aisladas haciendo una réplica de cada colonia en medio Endo-Agar sólido para comprobar la posible contaminación con coliformes y continuar con colonias de *S. meliloti* libres de *E. coli*.

M.9 EXTENSIÓN A PARTIR DE CEBADOR

Esta técnica, inicialmente descrita por Boorstein & Craig (1989) y posteriormente modificada por Marqués *et al* (1993), se ha utilizado para detectar el producto de escisión de los intrones del grupo II estudiados. Este método requiere la actividad de una reverso transcriptasa que extiende un cebador hacia el extremo 5′ del intrón para producir un DNA complementario al RNA molde de esa zona. Durante este proceso se mezcló RNA total con un exceso de un oligonucleótido complementario a una región cercana al extremo 5′ de los intrones. El oligonucleótido usado para analizar la escisión de los intrones RmInt1 y SmedInt1 fue el 97mer (5′-TGAAAGCCGATCCCG GAG-3′; Muñoz-Adelantado *et al*, 2003), que hibrida entre las posiciones 97-80 del RNA de éstos. Para evaluar la escisión del intrón RmInt2 se diseñó un nuevo cebador (debido a que el 97mer no presentaba una total complementariedad de bases): el oligonucleótido 105mer (5′-CCGCTCCTGAAAGCCG AT-3′), que hibrida entre las posiciones 105-88 del RNA de este intrón.

Para marcar radiactivamente cada oligonucleótido, se incubaron 50 pmol del cebador con 10 U de T4 PNK (New England Biolabs), tampón kinasa 1x y 10 μCi de [γ-32P]ATP (PerkinElmer) a 37ºC durante al menos 1-2 h. El oligonucleótido marcado en su extremo 5′ con 32P se limpió con una columna *Sephadex MicroSpin G25* (GE Healthcare) y se midieron las cuentas por minuto (cpm) que presentaba en el contador de centelleo (Beckman Coulter). Los marcadores de peso molecular usados en este tipo de ensayos, pGEMT y ϕ29, se marcaron de la misma manera que los oligonucleótidos.

Para la reacción de anillamiento, se incubaron 15 μg de RNA total con 0′2 pmol (300.000 cpm) del oligonucleótido marcado, tampón PIPES 10 mM pH 7′5 y NaCl 400 mM a 85ºC durante 5 min. Tras este tiempo, se enfrió rápidamente hasta los 60ºC y después se dejó enfriar lentamente hasta los 45ºC. A esta temperatura se le añadió la mezcla de extensión: tampón AMV RT 1x, dNTPs 1 mM, 60 μg/ml de actinomicina D (Sigma), 2 U del inhibidor RNAsaOUT (Invitrogen) y 7 U de AMV RT (Roche), y se incubó a 42ºC durante 1 h. La reacción de extensión se paró añadiendo a la muestra 150

µl de etanol 100 % (guardado a -20ºC) y 5 µl de acetato sódico 3 M pH 5'2. Los productos de extensión se analizaron en geles desnaturalizantes (urea 8 M) de poliacrilamida al 6 % (apartado M.7.3). El análisis y tratamiento de imágenes, así como la cuantificación de las bandas aparecidas, se realizó mediante el programa Quantity One v4.6.2. (Bio-Rad).

M.10 ENSAYOS DE MOVILIDAD DE LOS INTRONES

En la presente Tesis Doctoral se ha llevado a cabo la caracterización de dos nuevos intrones del grupo II tipo RmInt1. Entre otros estudios, se han realizado ensayos de movilidad (*homing*) de estos intrones mediante dos métodos: un sistema de doble plásmido, en el que se determina el porcentaje de invasión que producen las copias de RmInt1 expresadas desde un plásmido, y un sistema de plásmido único para evaluar la movilidad de las copias genómicas de intrón. El primer método consiste en introducir, mediante conjugación triparental (apartado M.8.3), dos plásmidos, uno donador de intrón y otro receptor de éste (con la correspondiente secuencia diana dependiendo del intrón que porta el plásmido donador), en una cepa del género *Sinorhizobium*. Generalmente, para estos ensayos se utilizó *S. meliloti* RMO17, una cepa bacteriana que no contiene copias genómicas de intrones tipo RmInt1 (figura R2.2) que puedan influir en el resultado, y en la que se ha demostrado que RmInt1 es capaz de escindirse y moverse (Martínez-Abarca *et al*, 2000; Nisa-Martínez *et al*, 2007). Aunque también se evaluó la movilidad de estos intrones en otros fondos genéticos (varias cepas de *S. meliloti*, *S. medicae* y *R. etli*). El segundo método para realizar ensayos de movilidad consiste en introducir, mediante conjugación triparental (apartado M.8.3), solamente el plásmido receptor de intrón (conteniendo la secuencia diana adecuada) en diferentes cepas bacterianas para determinar la movilidad de las copias genómicas de intrón.

Para determinar la movilidad de los intrones en la cepa *E. coli* DH5α, se llevó a cabo un ensayo siguiendo un sistema de doble plásmido. A esta cepa se le transfirió el DNA de los plásmidos donador y receptor simultánea-

Tabla M.5: Oligonucleótidos utilizados para generar las sondas de hibridación que se han usado en los ensayos de movilidad de los diferentes intrones.

Nombre	Secuencia (5'-3')	Sonda	Referencia
ε1	GGGGATCCCACGTGTCCAAGGGTGTTCTGTGACGA	RmInt1-5'	Muñoz-Adelantado, 2003
18R.0	GGGGATCCCTAGGATCGAGCCGCGGAAG		
sonda_RmInt2_F	TGGTGAGCGTTGGGTCAAAAGCCGC	RmInt2-5'	Este trabajo
sonda_RmInt2_R	CCGCTTATCGATCCTAAACGGCTTGTCT		
TetA_F	CGGCAATCATTCCGAGCATGAGTG	Tetraciclina (Tc)	Este trabajo
TetA_R	GCAGCCTTGAACGGCCTCAA		

mente mediante transformación, y, tras incubarse 1 h a 37ºC, la muestra se creció durante toda la noche en placas de LB a 30ºC (temperatura óptima para la movilidad de RmInt1; García-Rodríguez *et al*, 2011).

El DNA plasmídico, una vez extraído con el método adecuado (apartado M.4.2), se digirió con la enzima de restricción *Sal*I y los fragmentos se separaron en geles de agarosa no desnaturalizantes (apartado M.7.1) al 0'8 %. La enzima *Sal*I lineariza el plásmido donador (en el caso de que esté presente en la muestra) y divide en dos fragmentos el plásmido receptor (figura M.3), quedando en uno de ellos la secuencia diana del intrón. La diferencia de tamaños que se produce cuando el intrón invade su secuencia permite determinar los eventos de movilidad e incluso realizar la cuantificación de éstos mediante hibridación DNA-DNA con una sonda específica del plásmido receptor (sonda Tc). Los oligonucleótidos usados para generar las sondas que se han utilizado en este tipo de ensayos se detallan en la tabla M.5. Para evaluar la movilidad de los intrones, el experimento se llevó a cabo con DNA de, al menos, cuatro colonias independientes y posteriormente se calculó la media. El porcentaje de *homing* se determina aplicando la fórmula:

$$\text{Eficiencia de movilidad} = \frac{H}{H+R} \cdot 100$$

donde H es la cantidad de plásmido receptor invadido (producto de *homing*) y R es la cantidad de plásmido receptor no invadido.

M.11 HIBRIDACIÓN DNA-DNA

La técnica de hibridación DNA-DNA (*Southern Blot*) se ha usado en este trabajo para poner de manifiesto los eventos de movilidad (apartado M.10) de los intrones estudiados, así como para determinar el patrón de digestión que muestran varias cepas al digerir su DNA genómico con diferentes enzimas de restricción.

M.11.1 Transferencia alcalina por vacío

Una vez digerido el DNA plasmídico o total, éste se separa en geles de agarosa no desnaturalizantes (apartado M.7.1) al 0'8 % a 12-18 V durante 12-16 h. El gel se expone unos 30 min a luz UV para fragmentar parcialmente el DNA. Paralelamente, y para evitar el depósito de sales, la membrana de nylon a la que se fija posteriormente el DNA se humede con agua desionizada y se equilibra durante 5 min con una solución de 20xSSC (1xSSC está formado por: NaCl 150 mM, citrato sódico 15 mM, pH 7'0).

La transferencia del DNA desde el gel de agarosa a la membrana de nylon cargada positivamente (Pall Corporation) se llevó a cabo mediante un sistema de transferencia alcalina por vacío (*VacuGene* de Pharmacia). Para preparar este sistema, se coloca la membrana de nylon bajo una ventana de plástico de tamaño inferior al del gel de agarosa, y se sitúa el gel encima del conjunto con el cuidado de no formar burbujas de aire entre éste y la membrana de nylon. Se conecta la bomba de vacío a unos 55 mbar y, tras comprobar la ausencia de pérdida de vacío (que en algunos casos se consiguió sellando los bordes o grietas del gel con agarosa fundida), éste se cubre con una solución de NaOH 1 M y se deja durante 1-2 h. Posteriormente, se desmonta el sistema de transferencia, y la membrana de nylon se lava con una solución de 2xSSC durante 10 min en suave agitación y se seca a temperatura ambiente sobre papel Whatman 3MM. Finalmente, el DNA se fija a dicha membrana por calor a vacío (35 min a 120ºC y 70 cm Hg).

M.11.2 Marcaje de la sonda

Para la técnica de hibridación DNA-DNA se usó una sonda (específica para cada ensayo) marcada con digoxigenina acoplada a dUTP. Las sondas utilizadas comprenden una longitud de 200-600 pb. Éstas se sintetizan mediante PCR, utilizando como molde un DNA purificado de gel que previamente fue amplificado con la misma pareja de cebadores usados para la obtención de la sonda marcada. A la reacción de PCR, realizada con los componentes y condiciones generales (apartado M.6.1), se le añade digoxigenina-11-dUTP 0'1 mM. El fragmento amplificado se limpia con columnas del kit *MicroSpin S-300HR* (GE Healthcare) y se diluye hasta una concentración de 60 ng de DNA/ml de solución de prehibridación (5xSSC, sarcosil 0'1 %, SDS 0'02 %, formamida 50 % y solución de bloqueo (Roche) al 2 %). Suelen preparase entre 5-10 ml de sonda, un volumen suficiente para cubrir la membrana durante la hibridación. Se guardan a -20ºC, y pueden utilizarse en repetidas ocasiones (5-10 veces).

M.11.3 Hibridación, lavados y revelado

La hibridación y los lavados de la membrana de nylon se llevaron a cabo en el horno de hibridación *Hybridiser HB-1D* (Techne). En este proceso, se incuba la membrana en una solución de prehibridación (5xSSC, sarcosil 0'1 %, SDS 0'02 %, formamida 50 % y solución de bloqueo (Roche) al 2 %) a 42ºC durante al menos 2 h. Ésta se elimina y se añade la solución de hibridación, que no es más que la sonda (apartado anterior) desnaturalizada al hervirla durante 10 min y mantenerla en hielo 5-10 min. La membrana con la solución de hibridación se incuba a 42ºC durante al menos 16 h y, posteriormente, se recoge la sonda y se guarda a -20ºC para posteriores usos.

La membrana se lava dos veces con 50 ml de una solución de 2xSSC/SDS 0'1 % (p/v) durante 5 min a temperatura ambiente. Luego se lava otras dos veces con 50 ml de 0'1xSSC/SDS 0'1 % (p/v) durante 15 min a 68ºC para eliminar la sonda adherida inespecíficamente. Tras desechar esta solución,

tanto el rulo como el horno de hibridación se enfrían, para así lavar la membrana a temperatura ambiente con 50 ml de tampón de lavado [(0'3 % de Tween-20 en tampón 1 (ácido málico 0'1 M, NaCl 0'15 M, pH 7'5)] durante 5 min. Después de estos lavados, la membrana se bloquea durante 30 min a temperatura ambiente con 30 ml de tampón 2 (Blocking Reagent 1 % (p/v) en tampón 1). Se incuba 30 min con 20 ml de una dilución 1:10.000 del anticuerpo Anti-Digoxigenin-AP (Roche) en tampón 2. Para eliminar el anticuerpo no unido, se lava dos veces con 50 ml de tampón de lavado durante 15 min. La membrana se equilibra con 50 ml de tampón 3 (Tris-HCl 0'1 M, NaCl 0'1 M, MgCl2 50 mM, pH 9'5) durante 5 min, y posteriormente se incuba 5 min en 5 ml de una dilución 1:100 del sustrato quimioluminiscente CSPD (Roche) en tampón 3. La membrana se saca del rulo de hibridación y se seca completamente con papel Whatman 3MM. Se envuelve en plástico transparente y se mantiene en una estufa a 37ºC durante 15 min.

Tras todo este proceso, se hizo una exposición corta (de 4-6 h) y una larga (de ~24 h) sobre películas Kodak X-Omat que se revelaron con una solución de revelado y una solución de fijación diluidas según las indicaciones del fabricante. El análisis y tratamiento de imágenes, así como la cuantificación de las bandas aparecidas en las películas, se realizaron mediante el programa Quantity One v4.6.2 (Bio-Rad).

Una misma membrana puede utilizarse en más de una ocasión (podrían hacerse hasta cuatro hibridaciones), por lo que se realiza un proceso de *stripping* para eliminar la sonda anterior: se lava dos veces durante 15 min a 37ºC con una solución compuesta por NaOH 0'2 M y SDS 0'1 % (p/v), y después se lava 5 min con 2xSSC. En este punto, la membrana puede secarse y guardarse a temperatura ambiente, o se puede continuar con una nueva hibridación.

M.12 ANÁLISIS DE SECUENCIAS Y SECUENCIACIÓN

M.12.1 Obtención de secuencias de las bases de datos

Para conocer la distribución de los tres intrones del grupo II estudiados y de las dianas a las que se encuentran asociados realizamos un análisis mediante BLASTn con cada una de las secuencias, y trabajamos con los resultados que presentaron un porcentaje de cobertura con la secuencia analizada en cada caso mayor del 20 %. La secuencia de RmInt1, IS*Rm2011-2* e IS*Rm10-1* se consiguió en la base de datos del NCBI con los números de acceso Y11597.2, U22370 y AJ242573 respectivamente; la secuencia de IS*Rm17* se obtuvo de la base de datos ISfinder (Siguier *et al*, 2006b); y la secuencia de RmInt2, SmedInt1 e IS*Sme3* se obtuvo del genoma donde se han encontrado (GR4 y WSM419).

El número de acceso correspondiente a cada replicón de las cepas analizadas en esta Tesis Doctoral se detalla en la tabla M.6.

Tabla M.6: Número de acceso de las secuencias utilizadas en este trabajo.

Cepa bacteriana	**Replicón**	**Nº de acceso**
S. meliloti 1021	Cromosoma	AL591688.1
	pSymA	AE006469.1
	pSymB	AL591985.1
S. meliloti GR4	Cromosoma	CP003933.1
	pRmeGR4a	CP003934.1
	pRmeGR4b	CP003935.1
	pRmeGR4c	CP003936.1
	pRmeGR4d	CP003937.1
S. meliloti SM11	Cromosoma	CP001830.1
	pSmeSM11a	DQ145546.1
	pSmeSM11b	EF066650.1
	pSmeSM11c	CP001831.1
	pSmeSM11d	CP001832.1

Continúa en la página siguiente

Tabla M.6 – Continuación de la página anterior

Cepa bacteriana	**Replicón**	**Nº de acceso**
S. meliloti AK83	Cromosoma 1	CP002781.1
	Cromosoma 2	CP002782.1
	Cromosoma 3	CP002783.1
	pSINME01	CP002784.1
	pSINME02	CP002785.1
S. meliloti BL225C	Cromosoma	CP002740.1
	pSINMEB01	CP002741.1
	pSINMEB02	CP002742.1
S. meliloti Rm41	Cromosoma	HE995405.1
	pSYMA	HE995407.1
	pSYMB	HE995408.1
S. meliloti 102F51	pSymB-B152	DQ898558.1
S. meliloti C017	pHRC017	JQ665880.1
S. meliloti LPU88	ND	JQ753316.1
S. medicae WSM419	Cromosoma	CP000738.1
	pSMED01	CP000739.1
	pSMED02	CP000740.1
	pSMED03	CP000741.1
S. medicae RMO02	ND	AY608903.1
S. medicae RMO09	ND	AY608906.1
	ND	AY608907.1
Ensifer adhaerens 5D19	ND	AY248839.1
E. adhaerens R-6387	ND	AY608901.1
S. terangae ORS22	ND	AY608908.1
S. terangae ORS1009	ND	AY608905.1
S. fredii NGR234	Cromosoma	NC_012587.1
	pNGR234a	U00090.2
	pNGR234b	CP000874.1

Continúa en la página siguiente

Tabla M.6 – Continuación de la página anterior

Cepa bacteriana	Replicón	Nº de acceso
Mesorhizobium loti R7A	Cromosoma	BA000012.4
M. loti MAFF303099	ND	AL672114.1
R. leguminosarum bv. viciae 3841	Cromosoma	AM236080.1
	pRL7	AM236081.1
R. leguminosarum bv. trifolii WSM2304	Cromosoma	CP001191.1
	pRLG20	CP001192.1
R. etli 8C-3	REB01	DQ058415.1
	REB02	DQ058416.1
R. etli CE3	ND	AF176227.1
R. etli Viking1	ND	AY608902.1
R. etli CFN42	p42a	CP000134.1
	p42d	U80928.5
R. etli CIAT652	pB	CP001076.1
R. tropici CIAT899	pRtrCIAT899b	CP004017.1
Azoarcus aromaticum EbN1	Cromosoma	CR555306.1
Ochrobactrum anthropi ATCC 49188	Cromosoma 1	CP000758.1
	Cromosoma 2	CP000759.1
	pOANT03	CP000762.1

M.12.2 Secuenciación de DNA plasmídico y fragmentos de PCR

Las construcciones generadas en este trabajo se secuenciaron en el servicio de secuenciación de DNA/Genómica del Instituto de Parasitología y Biomedicina López-Neyra, y de la Estación Experimental del Zaidín. La visualización de los cromatogramas de las secuencias se realizó mediante el programa FinchTV v1.4.0. El trabajo rutinario con las secuencias (búsqueda de dianas de restricción, comparativa de secuencias, clonación *in silico*, etc.) se llevó a cabo con el programa Clone Manager Professional Suite v6.00.

M.12.3 Escalera de secuenciación mediante el método Sanger

Para corroborar el tamaño del producto de escisión del nuevo intrón RmInt2 se llevó a cabo una reacción de secuenciación con el *T7 Sequencing Kit* (USB), que se resolvió en geles desnaturalizantes de poliacrilamida (apartado M.7.3) paralelamente a las muestras obtenidas mediante extensión a partir de cebador (apartado M.9). Primeramente, se realizó la desnaturalización alcalina de 1'5-2 μg de DNA del plásmido pKGEMA4R2 y se anilló con 10 pmol del oligonucleótido 105mer (el mismo que se utilizó en la extensión a partir de cebador con el intrón RmInt2). Esta mezcla, en un volumen final de 14 μl, se incubó durante 5 min a 65ºC, 10 min a 37ºC y después 5 min a temperatura ambiente. La reacción de secuenciación se inició con la adición de la mezcla de marcaje (dATP, dGTP y dTTP 1'375 μM y NaCl 333'5 mM), 5 μCi de [α-32P]dCTP (3.000 Ci/mmol) y 2 μl de una dilución 1:4 de la T7 DNA polimerasa, y se mantuvo 5 min a temperatura ambiente. Posteriormente, pusimos 4'5 μl de esta reacción en cada uno de los cuatro tubos donde previamente pusimos los ddNTPs (840 μM de 3 dNTPs y el cuarto en una proporción 93'5 μM dNTP/14 μM ddNTP en un tampón de Tris-HCl 40 mM pH 7'5 y NaCl 50 mM) y se incubaron a 37ºC. Se añadieron 5 μl de tampón desnaturalizante 2x y se cargaron 2 μl por carril en un gel desnaturalizante de poliacrilamida.

M.12.4 Secuenciación del genoma de la cepa *S. meliloti* GR4

M.12.4.1 *Pirosecuenciación, ensamblaje y cierre del genoma de GR4*

Para la secuenciación del genoma de la bacteria *S. meliloti* GR4 se contrató el servicio de la empresa Macrogen Inc. (Korea del Sur), la cual llevó a cabo una carrera de pirosecuenciación directa (*direct shotgun*; Margulies *et al*, 2005) y una genoteca *paired end* de unas 3kb de tamaño (Ng *et al*, 2006; Jarvie & Harkins, 2008) mediante la tecnología 454 Titanium GS FLX de Roche. Se enviaron 76'5 μg de DNA total (extraído como se ha descrito en el apartado M.4.1) de un cultivo de GR4 en fase exponencial (crecido durante 2 días

a 30ºC y 190 rpm). El conector utilizado para la creación de esa genoteca fue el *linker Titanium*, de 42 nt (5'-TCGTATAACTTCGTATAATGTATGCT ATACGAAGTTATTACG-3'). Dicha empresa también realizó un ensamblaje *de novo* del genoma de GR4 mediante el programa Newbler v2.3 (Roche) usando los parámetros definidos por defecto (especificados en la figura M.6). Para la visualización del ensamblaje se utilizó el programa Tablet v1.10.03.04 (Milne *et al*, 2010).

1. Seed step
The number of bases between seed generation locations used in the exact k-mer matching part of the overlap detection.
Default value: 12

2. Seed length
The number of bases used for each seed in the exact k-mer matching part of the overlap detection (i.e. the "k" value of the k-mer matching)
Default value: 16

3. Seed count
The number of seeds required in a window before an extension is made
Default value: 1

4. Minimum overlap length
The minimum length of overlaps used by the assembler
Default value: 40

5. Minimum overlap identity
The minimum percent identity of overlaps used by the assembler
Default value: 90

6. Alignment identity score
When multiple overlaps are found, the per-overlap column identity score used to sort the overlaps for use in the progressive alignment
Default value: 2

7. Alignment difference score
When multiple overlaps are found, the per-overlapcolumn difference score used to sort the overlaps for use in the progressive multialignment
Default value: -3

Figura M.6: **Condiciones del ensamblaje llevado a cabo por el servicio de secuenciación de Macrogen.** Parámetros usados por dicha empresa para el ensamblaje *de novo* del genoma de GR4 mediante el ensamblador Newbler v2.3.

Los huecos generados entre los *contigs* y andamiajes de GR4 se completaron con *contigs* que correspondían a elementos repetidos, identificados mediante búsqueda en la base de datos ISfinder (Siguier *et al*, 2006b). Este tipo de *contigs* contienen la secuencia de más de una copia del elemento repetido al que corresponde, por lo que el programa Newbler genera una secuencia consenso de todas ellas. Para determinar la secuencia real de cada una de las copias del elemento repetido se reensamblaron las lecturas pareadas hermanas específicas de un hueco entre *contigs* adyacentes mediante el programa Geneious v4.8.5 (Drummond *et al*, 2009) con los parámetros definidos por defecto.

La obtención de la secuencia completa del genoma de *S. meliloti* GR4 requirió una serie de amplificaciones e hibridaciones DNA-DNA para confirmar la unión entre algunos *contigs* y andamiajes. Los oligonucleótidos utilizados en las reacciones de PCR, así como en la posterior secuenciación de los fragmentos amplificados correspondientes a zonas conflictivas, se describen en la tabla M.7. Por otro lado, los oligonucleótidos usados para generar las sondas que se han utilizado en las hibridaciones DNA-DNA se detallan en la tabla M.8.

Tabla M.7: Oligonucleótidos diseñados en este trabajo, utilizados para la amplificación y secuenciación de la unión de algunos *contigs* y andamiajes del genoma de GR4.

Uso	**Nombre**	**Secuencia (5'-3')**
Cierre de pRmeGR4a	pGR4a_1F	GACCACGAATTCTGGTTTCGCA
	pGR4a_1R	TGGAACACTGTTGCGAGCCTCAA
Cierre de pRmeGR4b	pGR4b_1F	ACTGTCGCGACTTCCTGCTTAGG
	pGR4b_1R	GCGCACAATCTCCCGTCCTGAA
	pGR4b_2F	GGAGTGCACAATGACTCAGGTCC
	pGR4b_2R	ACGCCAGATGCTTTCCGATGG
	pGR4b_3F	AATCTACCGCGCAAAGCTTGACAC
	pGR4b_3R	CCACCAAGTGGAAACGTCGTTCAC
	pGR4b_4F	GGTCGATCTTCTTTGTGCTGTTGG
	pGR4b_4R	TGAAACTCCTCCGGCCTACT

Continúa en la página siguiente

Tabla M.7 – Continuación de la página anterior

Uso	Nombre	Secuencia (5'-3')
Cierre de pRmeGR4b	inicioC18s	GACGCTCATGTTCGCGCTGAA
	C160s_nt200	ACCAACCTGTCGAGTATGGCATCG
Cierre de pRmeGR4c	inicioC120s_S12	CATCCACGACATCATTGCCGTCGA
	finalC120s_S12	GATGCAACGGTTTCGCACTTGAGC
	inicioC11s_S4	CTCTGGGACTGATCGCTGACGA
	finalC172s_S16	GCCTGGTACTCTGCGCCATGTA
	inicioC39s_S7	CGGAAGCACTTGAAGCGATCTAGG
	finalC46s_S7	CTATGCGACCAGAGCAACGCAC
	inicioC148s_S15	CGATCCCGTTTCACAGGCAAGC
	finalC150s_S15	CGGTCACCAATACCTCGGTGAACC
	inicioC161s	GATACGTCGCAAGGATGCCTGAG
	finalC17s	CCATCCTCACATCAAGGCGATGG
Cierre de pRmeGR4d	inicioC113s_S11	GAACCGCACTCTTTGGAGCTTTGG
	inicioC135s_S14	GGAGTTCAAGCTCTCAGGTCGCTA
	finalC142s_S14	TTTGGATCTGATGACGCCTGAGCA
	finalC117s_S11	TGCTGGATCTTTTGTGACAAGCAC
Cierre de huecos en el cromosoma	finalC83s_S10	CAGTCTGACCGTAAGCCGTCAACG
	inicioC84s_S10	CATGACTGCCCGTCAGTGCA
	finalC86s_S10	ATCGCGGTGATCGTGACCTG
	inicioC87s_S10	CATGCTCGTGGCTGAGCAGTG
	finalC88s_S10	TGATCGCGAATTCAATGGCGGCA
	inicioC89s_S10	GCACGACATGATGTTCGGCAAGC
	finalC107s_S10	CGTATTCGCAAGCGAAAGTTGC
	inicioC108s_S10	CGGAAGAGTCTTTTCAGCCCAA
	finalC184s_S17	GACGTTGAGCACGGTTTCGAC
	inicioC185s_S17	ATGGTCAGCATCTCAGTCAAGTCG
	Chr_C47_C48_F	TCCCCATCGAAATGCTACGGCA
	Chr_C47_C48_R	GGACTATGCCGATCGCAAGC
	C111_nt680_R	GGCCATCTCGACCGACAGGA
	C112_nt621_F	GCCGATATAGCCTGCCTCAA
Cierre de huecos en pRmeGR4c	finalC171s_S16	GAGGATCTCCGCTAGATCGGCTC
	inicioC172s_S16	CTTGTGCGGTGGCCGATAACG

Continúa en la página siguiente

Tabla M.7 – Continuación de la página anterior

Uso	Nombre	Secuencia (5'-3')
Secuenciación de las uniones con *contigs* 10, 37 y 38	RmInt2_nt835_F	TGGTCAAGCAGCTCATCGAACC
	RmInt2_nt1054_R	GCCATCTTTCGATGTAGAGCAACG
	RmInt2_nt1725_F	GCAAGTCGGTTCTTACAACGGCTC
	ISRm17_inicioC37	CCCTTGACGATGATGTCGGAATGC
	ISRm17_inicioC38	CGAAACGGGTTCAGCACTACC

Tabla M.8: Oligonucleótidos utilizados para generar las sondas de hibridación que se han usado para verificar la unión de algunos andamiajes de GR4.

Nombre	Secuencia (5'-3')	Sonda	Referencia
16S1	GGTGAGTGGAATTCCGAGTG	Gen ribosómico 16S	Muñoz-Adelantado, 2003
16S2	CATCTCACGACACGAGCTGA		
Rm2-1380F	GGGCAACATCCGAATGTGAC	RmInt2-3'	Este trabajo
Rm2-1620R	GTCCATAATAGCTGACCACC		
ISRm17_inicioC37	CCCTTGACGATGATGTCGGAATGC	IS*Rm17*	Este trabajo
ISRm17_inicioC38	CGAAACGGGTTCAGCACTACC		

M.12.4.2 *Anotación automática del genoma de* S. meliloti *GR4*

La anotación de los genomas secuenciados puede realizarse a través de servidores en línea o de programas que se descargan y se ejecutan en local. Médigue & Moszer (2007) resumen los programas normalmente usados para la anotación de genomas bacterianos, así como las principales bases de datos utilizadas para tal fin. Lo más conveniente para pequeños grupos de investigación que carecen de recursos computacionales es la anotación automática a través de servidores en línea (Duan *et al*, 2010; Beckloff *et al*, 2012). Basándonos en la comparativa realizada por Bakke *et al* (2009), donde evaluan tres sistemas de anotación automática de genomas, realizamos la anotación del genoma de GR4 mediante el servidor IMG ER (*Integrated Microbial Genomes Expert Review*; Markowitz *et al*, 2009) desarrollado por el instituto estadounidense JGI (*Joint Genome Institute*). Para tal fin, subimos al servidor un archivo en formato fasta de la secuencia de cada uno de los

replicones (Beckloff *et al* (2012) han descrito un protocolo para realizar el envío y la anotación de genomas en IMG ER). Tras la anotación, este servidor automático proporciona una serie de archivos que pueden ser descargados por el usuario. Éstos son:

- Un archivo en formato fasta con la secuencia de nucleótidos de todos los andamiajes
- Un archivo en formato fasta con la secuencia de aminoácidos de todas las proteínas
- Un archivo en formato fasta con la secuencia de nucleótidos de todos los genes
- Un archivo en formato fasta con las secuencias intergénicas
- Un archivo excel con la información génica

M.12.4.3 *Curación manual de la anotación automática del genoma de* S. meliloti *GR4*

Después de la anotación automática, realizamos una curación manual del genoma de GR4 siguiendo las recomendaciones indicadas por el personal del departamento de entrega de genomas de GenBank. Este proceso es necesario para eliminar los errores generados por los programas que llevan a cabo la anotación. De este modo, nos aseguramos de que los genomas presentes en las bases de datos superen los índices de calidad exigidos (Richardson & Watson, 2012). Los distintos tratamientos derivados de la curación manual del genoma de GR4 se enumeran y explican a continuación:

1. Genes de RNA.
   La anotación automática del genoma de GR4 presentaba todos los RNAs de transferencia, sin embargo, no proporcionaba información sobre el producto que generan. Por este motivo, tuvimos que buscar el producto que origina cada uno de ellos en una base de datos especí-

fica para este tipo de genes (tRNAdb; trna.bioinf.uni-leipzig.de). Mediante este proceso observamos que la anotación automática del servidor IMG ER identificó erróneamente cuatro genes como tRNAs. Estos genes, presentes uno en cada plásmido, eran de menor longitud que lo esperado para un tRNA, y finalmente se anotaron como miscRNAs con IncA como producto. La curación manual que realizamos de la anotación de los operones ribosomales fue definir la subunidad a la que corresponde cada uno de los rRNAs (23S, 16S y 5S).

2. Pseudogenes.
   El NCBI ofrece una herramienta (*Genome Submission Check Tool*) que permite examinar la anotación de un genoma para validarla antes de su envío a la base de datos de genomas microbianos (Benson *et al*, 2011). Gracias a ella, descubrimos 50 parejas de proteínas consecutivas que presentaban similitud con una misma proteína de mayor longitud tras realizar un análisis de BLASTp. Este hecho sugiere que esa pareja de proteínas puede representar un solo gen que ha sufrido un cambio en el marco de lectura o cualquier otra mutación. Por consiguiente, llevamos a cabo un análisis exhaustivo de cada pareja de proteínas, tras el que sólo 4 parejas permanecieron como dos proteínas independientes con la anotación automática original, mientras que 46 se combinaron para crear un pseudogen. Todos los pseudogenes estaban formados por dos proteínas consecutivas, sin embargo, en 3 de los casos el pseudogen se constituyó uniendo 3 proteínas consecutivas debido a que una de ellas estaba incluida en 2 parejas de proteínas consecutivas diferentes.

3. Estándares de UniProtKB para la nomenclatura de proteínas.
   La base de datos UniProtKB (*UniProt Knowledgebase*; Magrane & Consortium, 2011) recomienda seguir unas reglas de anotación específicas. Sin embargo, ninguno de los productos protéicos anotados por el servidor IMG ER estaban conectados con SwissProt (una sección de UniProtKB en la que la información de cada entrada ha sido curada manualmente). Por tanto, los términos que contradecían alguna de

esas reglas fueron eliminados.
En la anotación de un genoma no se debe hacer referencia a otros organismos para evitar ambigüedades en las bases de datos, por lo que se suprimieron de la anotación automática de GR4 los nombres de "*Bacillus*", "*Drosophila*", "*Escherichia coli*", "*Rickettsia*", "*Pseudomonas*" y "*yeast*".
Así mismo, se reemplazaron los términos "*homolog*", "*paralog*" y "*analog*", que infieren una relación evolutiva que, generalmente, no ha sido determinada para los genes anotados en un genoma, por el término "*-like protein*".
La base de datos UniProtKB recomienda que todas las proteínas que no tengan una función conocida simplemente se llamen "*hypothetical protein*". Por esta razón, se sustituyeron los nombres asignados por el servidor IMG ER de "*conserved hypothetical protein*", "*uncharacterized conserved protein*", "*uncharacterized domain*1" y "*uncharacterized protein conserved in bacteria*" por dicho término.
El servidor IMG ER atribuyó un número de comisión de enzima (*EC number*) a tres proteínas que, por otro lado, había definido como "*hypothetical protein*". Para resolver esta contradicción, realizamos un análisis de BLASTp con la secuencia de dichas proteínas y las nombramos de acuerdo a su función más probable.
Lo deseable es que todos los programas de anotación automática utilizaran los términos GO para nombrar las proteínas, ya que mejoraría la descripción de éstas y reduciría los errores sintácticos (Richardson & Watson, 2012).

4. Nucleótido de inicio del cromosoma.
   Toda proteína viene definida por un codón de inicio y un codón de parada, sin embargo, en el genoma de GR4 se encontró una región codificante que carecía de un codón de parada válido. Ésta correspondía al primer gen presente en el cromosoma, anotado como "*hypothetical protein*" en la cadena antisentido (con inicio en el nucleótido 72 y final en el nucleótido 1). Al analizar en detalle las zonas adyacentes a dicha

región, observamos que esa primera proteína y la última del cromosoma presentaban similitud con el gen *uroporphyrinogen-III decarboxylase*, pero sobre zonas distintas de éste. Debido a que la linearidad de los replicones bacterianos no es real, es recomendable que el comienzo de éstos se encuentre asociado al primer o último nucleótido de un marco abierto de lectura conocido. Por este motivo, modificamos la posición de inicio del cromosoma, que pasó a ser el nucleótido 73.

M.12.4.4 *Publicación del genoma de* S. meliloti *GR4 en la base de datos GenBank*

El NCBI dispone de una guía para el envío de genomas bacterianos (Sayers *et al*, 2011). Los pasos seguidos para la publicación del genoma de GR4 en la base de datos GenBank y obtención de un número de acceso válido fueron:

1. Registro del proyecto de genoma.
   La publicación de un genoma en la base de datos GenBank requiere inicialmente generar un proyecto (donde se introduce información general conocida sobre el microorganismo), con el que se asociará el genoma. El *project_id* asignado al genoma de GR4 es 175860 (número de acceso: PRJNA175860). En este paso también se registra el prefijo de la etiqueta que usará dicho genoma, que tiene que ser validado por la base de datos para evitar que haya dos genomas con el mismo prefijo. Éste debe tener entre 3-12 caracteres alfanuméricos y el primero no puede ser un dígito. Si en el momento de registrar el proyecto de genoma no se indica ningún prefijo, el NCBI proporciona uno por defecto. El prefijo de la etiqueta asignado al genoma de GR4 fue C770.

2. Preparación del formato de las secuencias.
   La secuencia de nucleótidos de cada replicón debe estar en formato fasta, que consiste en una sola línea de definición separada de las líneas de secuencia (una o varias) por un salto de línea. La línea de

definición empieza con un ">" seguido de, como mínimo, un identificador para la secuencia, llamado "SeqID". El identificador usado para nuestros archivos es el nombre de cada replicón (Chromosome, pRmeGR4a, pRmeGR4b, pRmeGR4c y pRmeGR4d). La línea de definición puede contener información adicional sobre la secuencia, que se introduce mediante modificadores. Además del identificador, nuestros archivos contienen los modificadores para definir el organismo y la cepa de la que proviene la secuencia. Por tanto, una de nuestras líneas de definición quedaría así:
>SeqID [organism=Sinorhizobium meliloti] [strain=GR4]

3. Preparación del formato de la anotación.
   Para enviar un genoma a la base de datos GenBank se requiere una anotación mínima, aunque se pueden incluir características adicionales. La información debe estar en una tabla de cinco columnas separadas por tabuladores en texto plano, llamada tabla de las características. Su primera línea está formada por ">Feature SeqID", donde "SeqID" es el mismo identificador usado para cada secuencia en formato fasta. Las líneas posteriores contienen los datos, distribuidos:
   - Columna 1 → localización del inicio de la característica
   - Columna 2 → localización del final de la característica
   - Columna 3 → clave de la característica
   - Columna 4 → clave del calificador
   - Columna 5 → valor del calificador.

   La tabla de las características fue generada a partir de la tabla con información génica en formato excel proporcionada por el servidor IMG ER. En esta tabla, cada característica consta de dos partes, una general donde se incluyen los datos del gen (localización, nombre y etiqueta) y otra específica donde se indican los datos de su producto (localización, nombre, función, número de comisión de enzima y *protein_id*). La información del gen debe contener como mínimo su localización y etiqueta, mientras que, dependiendo del tipo de producto, en la parte específica podrá añadirse la información total o parcialmente.

Por ejemplo, los RNAs no contienen el calificador *protein_id*, ni todas las proteínas tienen asignado un número de comisión de enzima.
A todos los genes se les debe atribuir un identificador sistemático que se indica en el valor del calificador etiqueta (*locus_tag*). Al prefijo de la etiqueta (proporcionado anteriormente) se le añade un guión bajo seguido de un número de identificación alfanumérico que es único para cada gen en todo el genoma. El identificador para los genes de cada replicón de GR4 son: GR4Chrxxxx, GR4pAxxx, GR4pBxxx, GR4pCxxxx y GR4pDxxxx, donde las "x" representan un número consecutivo desde el primer gen al último de cada replicón. Por ejemplo, en pRmeGR4d encontramos desde el gen C770_GR4pD0001 hasta el C770_GR4pD1558.
El calificador *protein_id* que se incluye en todas las proteínas tiene que ser asignado por el usuario. Debe tener el formato gnl | dbname | string, donde "dbname" es un término que hace referencia al nombre del grupo de investigación del usuario y debe ser lo más específico posible, y "string" es el valor del calificador *locus_tag* para cada proteína. La palabra usada para identificar a nuestro grupo de investigación fue NTORO. Así pues, el valor del calificador *protein_id* del primer gen encontrado en pRmeGR4d es gnl | NTORO | C770_GR4pD0001. Este tipo de identificación de los genes es importante para futuras actualizaciones del genoma y para que el NCBI pueda llevar a cabo un seguimiento interno de su base de datos.

4. Creación del archivo para el envío.
   Una vez que tenemos preparadas las secuencias de los diferentes replicones de GR4 en el formato requerido y sus respectivas tablas de las características, podemos generar el archivo definitivo para el envío del genoma a la base de datos GenBank. Para ello se utilizó un programa desarrollado por el NCBI llamado Sequin. Este programa consta de una serie de páginas a través de las cuales se incluye la información requerida, y, dependiendo de la naturaleza de los datos y de los archivos que se quieren obtener, se completan unas u otras.

Lo primero es introducir la información de los autores y del grupo de investigación que ha llevado a cabo la secuenciación del organismo, datos que se usarán como referencia en el propio acceso de la secuencia. También es necesario aportar información sobre el organismo secuenciado y el método de secuenciación utilizado. Además de esta información básica, el único archivo que cargamos en el programa para crear la entrada Sequin inicial de cada replicón es el de la secuencia en formato fasta. Una vez que visualizamos esa entrada inicial, el programa permite editarla e introducir la tabla de las características, generándose así la entrada Sequin final. Para completar el envío del genoma a la base de datos, tuvimos que mandar los archivos con la entrada Sequin final de cada replicón al departamento de entrega de genomas de GenBank. De esta manera, conseguimos los diferentes números de acceso: CP003933, CP003934, CP003935, CP003936 y CP003937 para cromosoma, pRmeGR4a, pRmeGR4b, pRmeGR4c y pRmeGR4d respectivamente.

M.13 CONSTRUCCIÓN DE ÁRBOLES FILOGENÉTICOS

En esta Tesis Doctoral hemos llevado a cabo un estudio sobre la relación evolutiva que presentan los intrones del grupo II de secuencia completa que tienen una identidad superior al 70 % con el intrón RmInt1 (número de acceso Y11597.2). La búsqueda e identificación de nuevas secuencias relacionadas con este intrón se realizó mediante un análisis de BLASTn, con la secuencia de 1.884 pb de RmInt1, en la base de datos del NCBI (www.ncbi.nlm.nih.gov/blast/Blast.cgi). Por otro lado, hemos analizado la relación que existe entre las múltiples copias de los genes *groEL* y *groES* presentes en el genoma de *S. meliloti* GR4, *S. meliloti* 1021 y *S. medicae* WSM419. La secuencia de los distintos genes *groEL* y *groES* se obtuvo del genoma de las cepas analizadas a través de la base de datos del NCBI (apartado M.12.1).

En el caso de los intrones del grupo II, se generaron árboles filogenéticos a partir de la secuencia nucleotídica de la rizobima y a partir de la secuencia de aminoácidos de la proteína que codifican. La secuencia de la ribozima (dominios I, II, III, V y VI) se obtuvo por la concatenación de dos fragmentos del intrón que comprenden los nucleótidos 1-514 y 1.802-1.884 de RmInt1, tomando los nucleótidos correspondientes a esas regiones en el resto de intrones. Para cada alineamiento se determinó el modelo teórico que mejor explica la evolución de las secuencias analizadas. En la construcción de los árboles se utilizó el modelo evolutivo GTR+G y WAG para las secuencias de nucleótidos y aminoácidos respectivamente. El estudio de los genes *groEL* y *groES* se llevó a cabo sólo con la secuencia de aminoácidos, y el modelo evolutivo utilizado para generar los árboles de ambos tipos de genes fue RtREV+G. Los modelos evolutivos se seleccionaron de acuerdo con el criterio de información de Akaike (AIC; Sugiura, 1978).

Para la construcción de los árboles filogenéticos, primero se realizó un alineamiento múltiple de las secuencias con el programa MAFFT (www.ebi.ac.uk/Tools/mafft/) y se visualizó con el programa Bioedit (Hall, 1999). La determinación del modelo teórico que mejor explica la evolución de las secuencias estudiadas se realizó mediante los programas PAUP (Swofford, 2002), ModelTest (Posada & Crandall, 1998) y MrModeltest (Nylander, 2004) para los alineamientos de nucleótidos, y con el programa ProtTest (Abascal *et al*, 2005) en el caso de los alineamientos de aminoácidos. Los árboles se generaron por el método de máxima verosimilitud (ML, *Maximum Likelihood*) utilizando el programa PhyML v2.4.4 (Guindon & Gascuel, 2003), y por inferencia bayesiana (BI, *Bayesian Inference*), usando MrBayes v3.1.2 (Ronquist & Huelsenbeck, 2003). En la construcción de los árboles se aplicó el modelo evolutivo determinado para cada alineamiento, y en los generados por el método de ML los soportes de rama se obtuvieron mediante 1.000 réplicas de bootstrap. Los árboles se visualizaron con el programa MEGA (Tamura *et al*, 2007). La conversión del formato de los distintos archivos utilizados se llevó a cabo con el programa Readseq (www.ebi.ac.uk/cgi-bin/readseq.cgi).

Para determinar si los árboles proporcionan una información consistente llevamos a cabo un test de congruencia entre matrices de distancias (CADM, Congruence Among Distance Matrices; Legendre & Lapointe, 2004) global a partir de las distancias calculadas para los árboles generados con la ribozima y la proteína (Campbell *et al*, 2011). Este test proporciona, además de un valor p (estimado usando 10.000 permutaciones), un estadístico W como estima del grado de congruencia entre las matrices analizadas, cuyo valor varía entre 0 (ausencia de congruencia) y 1 (congruencia total). Para realizar el test CADM global se utilizó el paquete ape (Paradis *et al*, 2004) del lenguaje R.

RESULTADOS Y DISCUSIÓN

CAPÍTULO 1

SECUENCIACIÓN DEL GENOMA DE LA CEPA BACTERIANA *Sinorhizobium meliloti* GR4

En los últimos años se ha secuenciado el genoma de un gran número de cepas bacterianas, el cual se verá incrementado por la mejora de las técnicas de secuenciación masiva y el abaratamiento de éstas. La diversidad de hábitats en los que se encuentran las especies pertenecientes a α-proteobacteria, entre otras características, hace que ésta sea una de las clases bacterianas más estudiadas, con aproximadamente 600 genomas completamente secuenciados (Pini *et al*, 2011).

En este capítulo se presenta la secuenciación del genoma de *S. meliloti* GR4 y las distintas aproximaciones que nos han llevado hasta la obtención de su secuencia completa.

Los datos obtenidos tras la secuenciación y el ensamblaje *de novo* del genoma de GR4 (ver Material y Métodos, apartado M.12.4.1) se detallan en la figura R1.1. Se obtuvieron un total de 1.780.998 lecturas, de las cuales 1.712.561 se ensamblaron completamente y 10.294 parcialmente, dando lugar a 174 regiones contiguas (*contigs*) de una longitud mayor a 100 pb. Gracias a las lecturas pareadas, obtenidas a partir de la genoteca *paired end*, el ensamblador Newbler fue capaz de ordenar los *contigs* y estimar la longitud de los huecos entre ellos, construyendo así 17 andamiajes (*scaffolds*) que comprenden 7.074.675 bases del genoma de GR4. La cobertura de un genoma, tras su secuenciación, representa el número de veces que un nucleótido ha sido leído, y se calcula como el número total de bases leídas entre las bases que componen el genoma (tras la formación de los andamiajes). En el caso de GR4, se obtuvo una cobertura 77x, de la cual un 17x corresponde a

1. Read

TotalNumberOfReads	TotalNumberOfBases	Assembled	Partial	Singleton	Repeat
1780998	542068492	1712561	10294	6616	50717

2. PairedRead (Paired End)

BothMapped	OneUnmapped	MultiplyMapped	BothUnmapped	DistanceAvg	DistanceDev
403771	4553	47715	46	2684.8	671.2

3. All Contig (Length >= 100bp)

Contigs	Bases
174	7003497

4. Large Contig (Length >= 500bp)

Contigs	Bases	AvgContigSize	N50ContigSize	LargestContigSize	Q40PlusBases, %Q40
155	6998274	45150	110012	529788	6997461, 99.99%

5. Scaffold (Paired End)

Scaffolds	Bases	AvgScaffoldSize	N50ScaffoldSize	LargestScaffoldSize
17	7074675	416157	1030511	2753791

Figura R1.1: **Información sobre la pirosecuenciación y el ensamblaje del genoma de GR4 realizados.** Se muestran los datos totales de la pirosecuenciación (punto 1) así como de la genoteca *paired end* (punto 2) llevadas a cabo. Los puntos 3 y 4 indican la información del ensamblaje de las lecturas obtenidas tras la pirosecuenciación. El punto 5 muestra la información de los andamiajes construidos gracias a las lecturas pareadas obtenidas con la genoteca *paired end*.

las lecturas pareadas. Los *contigs* largos ($\geqslant$ 500 pb) presentaron un tamaño medio de 45.150 pb, conteniendo el de mayor tamaño 529.788 pb. A su vez, el tamaño medio de los andamiajes fue de 416.157 pb, siendo el mayor de ellos de 2.753.791 pb. El valor N50 (cuyo significado se describe en la figura

R1.S1) fue de 110.012 pb para los *contigs*, mientras que para los andamiajes, dicho valor fue de 1.030.511 pb.

R1.1 ÍNDICE DEL GRADO DE COBERTURA

Tras un análisis detallado del ensamblaje de los *contigs*, observamos que el grado de cobertura de las lecturas solapadas en determinados *contigs* era superior a la cobertura que presentaban el resto de *contigs*. Además, pudimos apreciar este hecho dentro de un mismo *contig*, el 110 (figura R1.2). Éste tiene una longitud de 87.006 pb, y en su extremo 3' se observa un aumento de lecturas ensambladas en una zona de aproximadamente 6 kb, correspondiente al operón ribosomal. El grado de cobertura de dicha zona es tres veces superior al del resto del *contig*, coincidiendo con las tres copias de este operón presente en el genoma de otros rizobios (Galibert *et al*, 2001; Galardini *et al*, 2011b; Schneiker-Bekel *et al*, 2011).

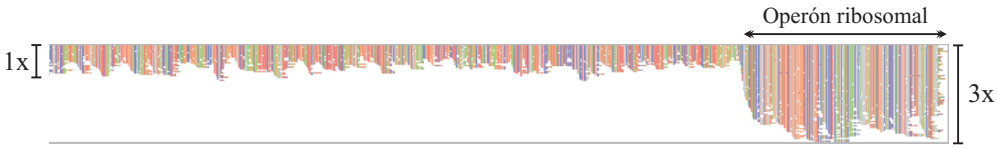


Figura R1.2: **Relación entre el grado de cobertura de una región y su número de copias dentro del genoma.** Visualización del *contig* 110 como ejemplo del grado de cobertura de una región única en el genoma (1x) y una zona repetida. La línea entre flechas indica una región de 6'3 kb, correspondiente al operón ribosomal, que tiene un grado de cobertura 3x, coincidiendo con las tres copias de este operón presentes en el genoma de GR4.

Teniendo en cuenta esta observación, la estrategia seguida para cerrar el genoma de GR4 se basó en determinar un parámetro, al que hemos denominado índice del grado de cobertura (FCI, del inglés *Fold Coverage Index*), que estima el número de copias de cualquier *contig*. El FCI, para un determinado *contig*, se calcula aplicando la fórmula:

$$FCI_i = \frac{NL_i}{LON_i} : \frac{NLT}{LONT}$$

donde NL_i y LON_i representan el número de lecturas que componen el *contig* i y su longitud, respectivamente; NLT es el número total de lecturas ensambladas en el genoma y LONT es la longitud del genoma estimada tras la creación de los andamiajes.

El valor del cociente (NL_i/LON_i) proporciona información sobre las veces que un nucleótido dentro de ese *contig* ha sido leído de media. Para determinar el FCI, este valor se pone en relación al número de veces que, en promedio, un nucleótido cualquiera del genoma ha sido leído (NLT/LONT). En nuestro caso, el número total de lecturas ensambladas (total o parcialmente; NLT) en el genoma de GR4 fue 1.722.855 (punto 1 en la figura R1.1) y la longitud estimada (LONT) de dicho genoma fue de 7.074.675 pb (punto 5 en la figura R1.1). Por tanto, la fórmula del FCI aplicada a cada *contig* i fue:

$$FCI_i = \frac{NL_i}{LON_i \cdot 0'2435}$$

El ensamblador generó los andamiajes incluyendo huecos entre *contigs* que debían ser adyacentes. Por esta razón, la longitud de los andamiajes compuestos por más de un *contig* es estimada, y el FCI de éstos se calculó, a partir de los datos referentes a longitud y número de lecturas de los *contigs* que lo componen, aplicando la siguiente fórmula:

$$FCI_j = \frac{\sum NL_k}{\sum LON_k} : \frac{NLT}{LONT}$$

donde NL_k representa el número de lecturas del *contig* k perteneciente al andamiaje j y LON_k la longitud de dicho *contig*; NLT es el número total de lecturas ensambladas en el genoma y LONT es la longitud del genoma estimada tras la creación de los andamiajes.

Tabla R1.1: Características de los andamiajes generados tras el ensamblaje del genoma de GR4.

Andamiaje	**Longitud**[a]	**Nº *contigs***[b]	**FCI**	**Nº copias**[c]	**Sintenia**
Andamiaje 9	173438	5	0'24	1	pRmeGR4a
Andamiaje 6	191956	14	0'36	1	pRmeGR4b
Andamiaje 5	2125	1	0'38	1 (1)	-
Andamiaje 11	658831	5	0'93	1	pSymB 1021
Andamiaje 14	1030511	8	0'96	1	pSymB 1021
Andamiaje 16	670337	14	0'96	1	pSymA 1021
Andamiaje 12	344444	1	0'96	1	pSymA 1021
Andamiaje 15	41156	3	0'98	1	pSymA 1021
Andamiaje 7	146125	8	1'04	1	pSymA 1021
Andamiaje 4	208439	8	1'07	1	pSymA 1021
Andamiaje 8	437883	9	1'08	1	Cromosoma 1021
Andamiaje 17	405807	5	1'10	1	Cromosoma 1021
Andamiaje 10	2753791	36	1'11	1	Cromosoma 1021
Andamiaje 1	2528	1	1'74	2	-
Andamiaje 3	2488	1	2'57	2	-
Andamiaje 2	2696	1	3'53	4 (1)	-
Andamiaje 13	2120	1	5'01	5 (1)	-

[a] en los andamiajes formados por más de un *contig* se muestra la longitud estimada por el ensamblador Newbler tras a la adición de huecos entre *contigs*
[b] número de *contigs* que componen cada andamiaje
[c] número de copias presentes en el genoma. Entre paréntesis se muestra el número de copias localizadas en los plásmidos crípticos
[d] plásmidos crípticos

De los 12 andamiajes que tienen una longitud superior a 40 kb, 10 mostraron un valor del FCI entre 0'93 y 1'11 (tabla R1.1). Un análisis de sintenia con el genoma de 1021 reveló que esos 10 andamiajes se correspondían con el cromosoma y los plásmidos simbióticos. Los andamiajes pertenecientes al cromosoma de GR4 (andamiajes 8, 10 y 17) presentaron un valor del FCI similar y, a su vez, superior al de los plásmidos simbióticos. Respecto a éstos, pRmeGR4c (equivalente a pSymA) es el que mostró valores del FCI más dispares, desde 0'96 (andamiaje 12 y 16) hasta 1'07 (andamiaje 4); el

FCI de los 2 andamiajes que componían pRmeGR4d (equivalente a pSymB) fue de 0'93 y 0'96 para el andamiaje 11 y 14 respectivamente. Llama la atención que los andamiajes 6 y 9 presentaron un FCI muy inferior al resto de andamiajes mayores de 40 kb (FCI de 0'36 y 0'24 respectivamente). Un análisis mediante BLASTn puso de manifiesto la identidad de algunas regiones de estos plásmidos con secuencias previamente publicadas presentes en los plásmidos crípticos del aislado GR4 (Mercado-Blanco & Olivares, 1994; García-Rodríguez *et al*, 2000). Así, pudimos determinar que el andamiaje 6 (de 191.956 pb) correspondía al plásmido pRmeGR4b y el andamiaje 9 (de 173.438 pb) a pRmeGR4a. Uno de los 5 andamiajes con longitud inferior a 3 kb (andamiaje 5) tenía un FCI similar al del andamiaje 6 (FCI de 0'38). Una búsqueda en las bases de datos reveló que presentaba identidad con una secuencia de inserción (IS*Rm30*). El valor del FCI del resto de andamiajes menores de 3 kb fue superior a 1'7, lo cual es indicativo de que su secuencia se encuentra más de una vez dentro del genoma de GR4.

Tras un análisis detallado del FCI de los *contigs* que componen los andamiajes (tabla R1.S1), observamos que determinados *contigs* presentaban un valor del FCI alejado del valor mostrado por el resto de *contigs* pertenecientes al mismo andamiaje. Éste fue el caso del *contig* 10 (FCI de 8'64 y 1.906 pb) dentro del andamiaje 4 (pRmeGR4c) y de los *contigs* 159 (FCI de 0'34 y 14.618 pb) y 160 (FCI de 1'57 y 1.567 pb) dentro del andamiaje 16 (pRmeGR4c). Por otro lado, encontramos *contigs* cuyo valor del FCI era ligeramente distinto del FCI del resto de *contigs* del andamiaje que componían. Éstos eran el *contig* 13 (FCI de 0'67 y 1.682 pb) dentro del andamiaje 4 (pRmeGR4c), los *contigs* 77 (FCI de 1'35 y 42.604 pb) y 110 (FCI de 1'33 y 87.006 pb) dentro del andamiaje 10 (cromosoma) y el *contig* 141 (FCI de 1'38 y 7.090 pb) dentro del andamiaje 14 (pRmeGR4d). Con los *contigs* menores de 2 kb llevamos a cabo un análisis mediante BLASTn y BLASTx. El *contig* 10 presentó identidad con una reverso transciptasa y una madurasa relacionadas con intrones del grupo II, el *contig* 13 con la proteína TraA1 y el *contig* 160 con IS*Rm23*. Estos resultados nos llevaron a considerar los *contigs* 10 y 160 como elementos repetidos dentro del genoma de GR4, y a excluir

al *contig* 159 del andamiaje 16 por su bajo valor del FCI (similar al del plásmido críptico pRmeGR4b; tabla R1.1).

El FCI ha sido calculado para todos los *contigs* mayores de 100 pb, estimando así el número de veces que tiene que ser incluido cada *contig* en el genoma de GR4 (tabla R1.S1 y R1.S2). Los *contigs* repetidos se forman por el ensamblaje de las lecturas obtenidas a partir de las múltiples copias de la región repetida a la que corresponden. Por tanto, un *contig* repetido es realmente el consenso de la secuencia de todas esas copias. En algunos casos, esto genera variaciones nucleotídicas (SNPs) puntuales además de las variaciones distribuidas aleatoriamente que aparecen a lo largo de cualquier *contig* debido a errores en la secuenciación. Existen *contigs* de elementos repetidos que presentan un elevado número de SNPs, como el *contig* 158 (34 SNPs) correspondiente a las dos copias del gen *repC* que encontramos en

Tabla R1.2: Características de algunos *contigs* correspondientes a elementos repetidos.

Contig	**Longitud**	**Nº lecturas**	**FCI**	**Nº copias**[a]	**Nº SNPs**	**Gen**[b]
Contig 112	1197	3864	13’26	11	4	IS*Rm19*
Contig 154	1882	5357	11’69	10 (1)	2	RmInt1
Contig 10	1906	4009	8’64	7	0	RmInt2[c]
Contig 156	1314	2528	7’90	6	5	IS*Rm22*
Contig 111	1341	2475	7’58	7 (1)	13	IS*Rm5*
Contig 189	1522	1004	2’71	3 (1)	9	*GroEL1/2*
Contig 125	1779	1077	2’49	2	0	Transposasa[d]
Contig 71	507	297	2’41	3 (1)	6	*NodQ*
Contig 19	951	380	1’64	2 (1)	0	IS*Rm23*
Contig 160	1567	598	1’57	2 (1)	15	IS*Rm23*
Contig 158	1126	173	0’63	2 (2)	34	*RepC*

[a] número de copias presentes en el genoma. Entre paréntesis se muestra el número de copias localizadas en los plásmidos crípticos

[b] identificación y nomenclatura de los genes según está descrito en el genoma de referencia de la cepa *S. meliloti* 1021 (http://iant.toulouse.inra.fr/bacteria/annotation/cgi/rhime.cgi)

[c] nuevo intrón del grupo II relacionado con RmInt1

[d] transposasa relacionada con la proteína identificada como AEH82072.1

los plásmidos crípticos, y otros *contigs* que, aunque se encuentren repetidos en el genoma, no contienen SNPs en su secuencia (tabla R1.2).

R1.2 CIERRE DE LOS HUECOS GENERADOS EN EL ENSAMBLAJE DEL GENOMA DE GR4 Y RESOLUCIÓN DE SNPs

Como hemos mencionado anteriormente, el ensamblaje *de novo* del genoma de GR4 dio como resultado la formación de 174 *contigs* de una longitud superior a 100 pb. Con la ayuda de las lecturas pareadas se obtuvieron

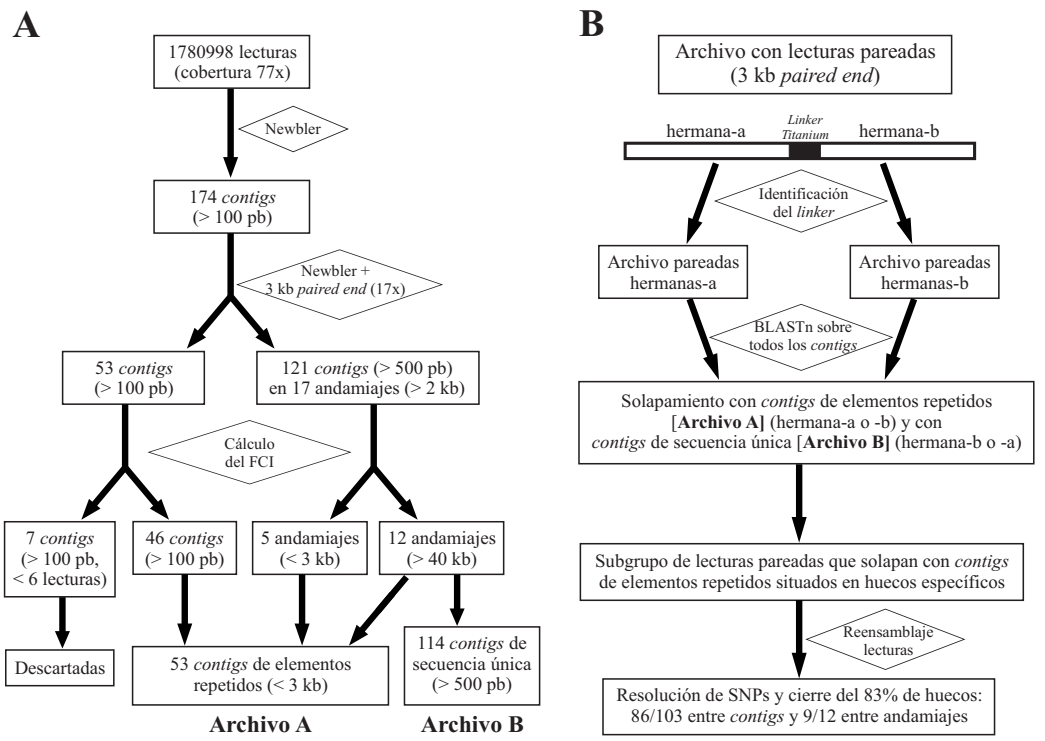


Figura R1.3: **Estrategia del cierre del genoma de GR4.** (A) Esquema de los datos iniciales y los datos obtenidos tras calcular el FCI. (B) Diagrama de flujo que muestra los pasos seguidos para cerrar el 83 % de los huecos presentes en el genoma de GR4 después del ensamblaje y resolver los SNPs gracias a las lecturas pareadas obtenidas con la genoteca *paired end*.

17 andamiajes que contenían 121 *contigs* de más de 500 pb (figura R1.3A). El cálculo del FCI permitió diferenciar dos tipos de *contigs* (los que correspondían a elementos repetidos dentro del genoma y los que presentaban una secuencia única), que se agruparon en dos archivos distintos. En un primer archivo (de elementos repetidos; archivo A en la figura R1.3A) se incluyeron 46 de los 53 *contigs* que no formaban parte de los andamiajes, los 5 andamiajes menores de 3 kb y los *contigs* de elementos repetidos identificados dentro de los andamiajes (*contig* 10 y 160; tabla R1.S1 y R1.S2). Siete *contigs* de los que no formaban parte de los andamiajes fueron descartados por presentar un valor del FCI menor de 0'1 (tabla R1.S2). El segundo archivo (archivo B en la figura R1.3A) estaba constituido por 114 *contigs* de secuencia única (todos los que componían inicialmente los andamiajes mayores de 40 kb a excepción de los *contigs* 10, 159 y 160). Esta diferenciación de *contigs* generó finalmente un borrador del genoma de GR4 con 115 huecos, de los cuales 103 se encontraban entre *contigs* de un mismo andamiaje y 12 completaban el cierre entre andamiajes (figura R1.4).

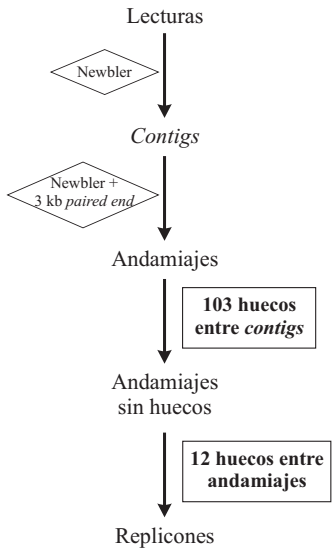


Figura R1.4: **Número de huecos resueltos para la formación de los cinco replicones que componen el genoma de GR4.**

Para el cierre de los huecos del genoma de GR4 diseñamos una estrategia basada en las lecturas pareadas obtenidas con la genoteca *paired end* de 3 kb (figura R1.3B). Primero llevamos a cabo la identificación del conector usado en la creación de la genoteca *paired end* (*linker Titanium*), y después creamos dos archivos para cada lectura diferenciando la secuencia que se encuentra aguas arriba (pareada hermana-a) y abajo (pareada hermana-b) del conector. La distancia media que separa las dos secuencias que flanquean al conector en el genoma de GR4 se estimó en 2.685 pb (punto 2 de la figura R1.1). Con objeto de identificar el *contig* donde solapaban específicamente cada una de las lecturas pareadas hermanas, se llevó a cabo un análisis mediante BLASTn de los archivos que contenían este tipo de lecturas frente a los archivos A y B previamente generados (que incluían los *contigs* de elementos repetidos y de secuencia única respectivamente). Tras esta comparación, determinamos los *contigs* de elementos repetidos que debían incluirse en huecos específicos entre dos *contigs* de secuencia única, y creamos, a su vez, subgrupos de lecturas pareadas asociadas a los *contigs* de elementos repetidos presentes en cada uno de estos huecos específicos. El reensamblaje de las lecturas contenidas en esos subgrupos permitió definir la secuencia real de los *contigs* de elementos repetidos en cada hueco, resolviendo así los SNPs presentes en alguno de éstos (tabla R1.2). En este reensamblaje posterior de las lecturas pareadas se consiguió una cobertura de, al menos, cuatro lecturas por nucleótido. Gracias a la estrategia diseñada (figura R1.3B), rellenamos el 83 % de los huecos presentes en el borrador del genoma de GR4 (figura R1.4): 86 de los 103 huecos entre *contigs* dentro de andamiajes (84 %) y 9 de los 12 huecos que completaban el cierre entre andamiajes (75 %).

La figura R1.5 muestra, a modo de ejemplo, la identificación y resolución de SNPs llevada a cabo en el *contig* 154, que se corresponde con el intrón del grupo II RmInt1 (un elemento repetido con el que se ha trabajado desde finales de los años 90 y que es objeto de estudio del segundo y tercer capítulo de esta Tesis Doctoral; Martínez-Abarca *et al*, 1998). Un análisis detallado de dicho *contig* reveló la existencia de dos SNPs localizados en las posi-

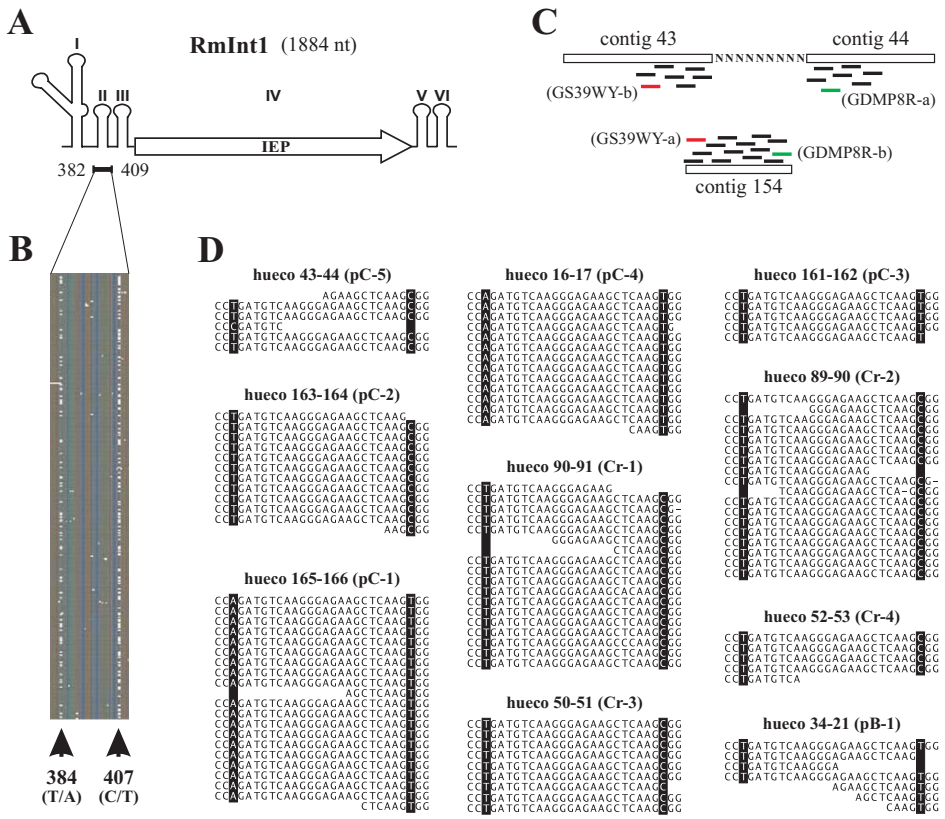


Figura R1.5: **Identificación y resolución de SNPs dentro del *contig* 154 (correspondiente al elemento repetido RmInt1).** (A) Esquema de la estructura del intrón del grupo II RmInt1 dispuesta en seis dominios (DI-DVI) y de la proteína codificada por él (IEP) presente en el DIV. (B) Visualización parcial de las lecturas asociadas al *contig* 154. Se muestran las lecturas ensambladas entre las posiciones 382 y 409 de RmInt1. Los puntos blancos indican variaciones nucleotídicas en las lecturas respecto a la secuencia consenso. Cuando estas variaciones se acumulan en una posición concreta (señaladas con una flecha) son indicativas de la presencia de SNPs. (C) Esquema de uno de los huecos resueltos donde debe incluirse el *contig* 154 según la estrategia descrita en la figura R1.3B. Las lecturas pareadas hermanas que permiten esa identificación se representan como líneas negras, mostrándose en rojo y en verde dos de estas parejas. (D) Secuencia real de las distintas copias de RmInt1 (posiciones 382-409), obtenidas gracias al reensamblaje de las lecturas pareadas provenientes de huecos específicos entre *contigs*. Cada línea en el ensamblaje corresponde a una lectura pareada. Las columnas negras señalan las posiciones que presentan variaciones nucleotídicas acumuladas. Los ensamblajes están nombrados de acuerdo a los *contigs* de secuencia única adyacentes a cada hueco, indicándose entre paréntesis el nombre de la copia de intrón según se define en las figuras R1.12 y R3.3A.

ciones 384 y 407 de RmInt1 (figura R1.5B), siendo el nucleótido consenso T y C respectivamente. El reensamblaje de las lecturas pareadas asociadas diferencialmente a las diez copias de este elemento repetido (figura R1.5D) permitió determinar la secuencia real de cada una de las copias.

R1.2.1 Cierre de los huecos entre *contigs* dentro de un mismo andamiaje

La longitud de los huecos entre *contigs* estimada por el programa Newbler para generar los andamiajes varió entre 246 pb y 2.128 pb. La estrategia de las lecturas pareadas permitió resolver 86 huecos dentro de andamiajes, incluyendo la secuencia real (sin SNPs) del *contig* correspondiente (figura R1.6). En todos los huecos incluimos un solo *contig* de elemento repetido salvo en uno, donde el subgrupo de lecturas pareadas pertenecía a dos *contigs* de elementos repetidos distintos (*contig* 111 y 112) y, gracias al reensam-

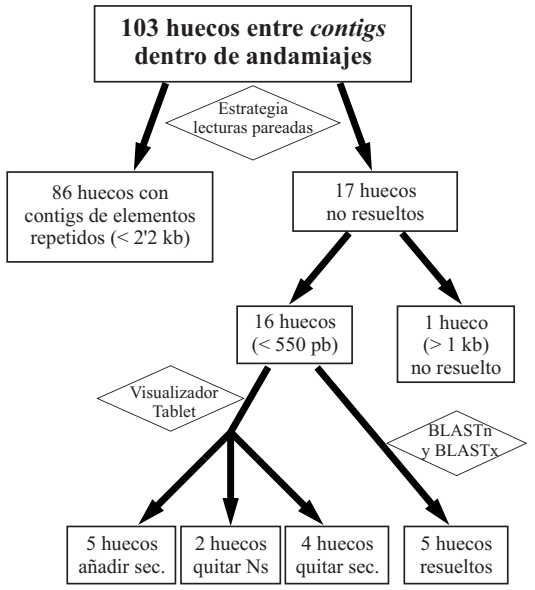


Figura R1.6: **Resolución de los huecos entre *contigs* presentes en el genoma de GR4 tras su ensamblaje.** Se muestra el número, la longitud estimada y la estrategia seguida para rellenar cada tipo de hueco.

blaje de estas lecturas, rellenamos el hueco con la secuencia exacta (para su confirmación se llevó a cabo una secuenciación Sanger de esa zona). El cierre de los 17 huecos no resueltos mediante la estrategia de las lecturas pareadas se abordó de diferente manera dependiendo de su complejidad. El hueco situado entre los *contigs* 107 y 108 dentro del andamiaje 10, con una longitud estimada de 1.185 pb, no pudo ser finalmente resuelto. Los huecos restantes presentaban una longitud estimada inferior a 550 pb. Once de ellos se resolvieron observando (con el programa Tablet) las lecturas parcialmente ensambladas en el extremo de los *contigs* de secuencia única que los flanqueaban: en 5 casos añadimos un fragmento de una longitud de entre 95 nt y 302 nt, en 2 unimos directamente los *contigs* (suprimiendo las "Ns" que el ensamblador había introducido entre ellos) y en 4 eliminamos parte de la secuencia de uno de los *contigs* por solapamiento con el *contig* adyacente. Los otros 5 huecos con menos de 550 pb estimadas necesitaron un estudio adicional por tratarse de zonas con repeticiones de pequeños motivos. Tras realizar un análisis mediante BLASTn y BLASTx frente a las bases de datos con los *contigs* que flanqueaban esos huecos, pudimos inferir la secuencia definitiva para cada hueco.

R1.2.2 Cierre de los huecos entre andamiajes

El ensamblaje del genoma de GR4 generó 17 andamiajes (tabla R1.1). Tras el cálculo del FCI, los 12 andamiajes que presentaban una longitud superior a las 40 kb fueron considerados andamiajes de secuencia única. Los plásmidos crípticos pRmeGR4a y pRmeGR4b estaban formados por un único andamiaje, el cual debía ser cerrado para crear una molécula circular. El resto de replicones estaban compuestos por varios andamiajes, por lo que el cierre de éstos requería una ordenación previa de los andamiajes dentro de cada replicón. Esta tarea se llevó a cabo con distintas aproximaciones dependiendo del tipo de hueco (figura R1.7).

La unión de los extremos del andamiaje 9 para construir el plásmido pRmeGR4a se resolvió con la estrategia de las lecturas pareadas. Para com-

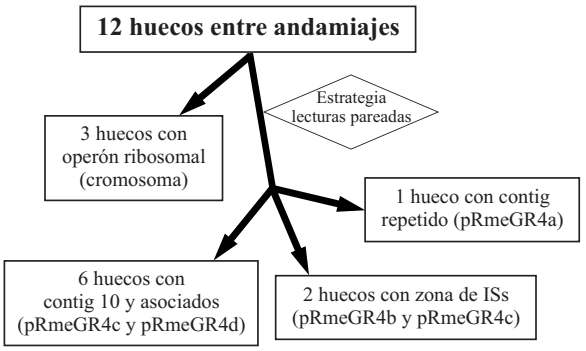
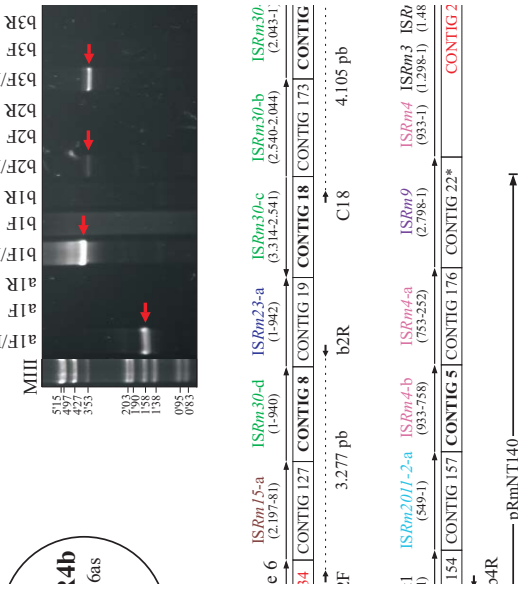


Figura R1.7: **Resolución de los huecos entre andamiajes para el cierre de los replicones.** Se indican el tipo de secuencia introducida en cada hueco y la localización genómica de los huecos.

pletar el plásmido se incluyó una copia de IS*Rm5* (de 1.340 pb), cuya secuencia correspondía al hueco específico entre los *contigs* 60 y 64 (tabla R1.S1). Dicha unión se confirmó mediante PCR, con los oligonucleótidos a1F y a1R (figura R1.8A), y posterior secuenciación Sanger del producto amplificado.

Figura R1.8: **Cierre de los plásmidos crípticos pRmeGR4a y pRmeGR4b.** (A) Esquema de la estructura de estos plásmidos, donde se indica la zona de unión del único andamiaje que los componen. (B) Detalle de los *contigs* presentes en la zona de ISs de pRmeGR4b. Los *contigs* de secuencia única que el ensamblador incluyó en este plásmido se destacan en rojo, y los que no incluyó, en negrita. El asterisco significa que el *contig* 22 en esa región se encuentra de manera parcial (sólo la IS*Rm9*). Las flechas encima de los *contigs* muestran su orientación. Encima de las flechas se indica el elemento repetido al que tiene similitud cada *contig*, mostrándose, entre paréntesis, los nucleótidos y la orientación de dicho elemento en el *contig*. Cada elemento repetido con similitud a más de un *contig* aparece en un color diferente, y los fragmentos del elemento se representan con diferentes letras. Debajo de los *contigs* se muestra la localización de los cebadores usados para confirmar la estructura de la zona de ISs propuesta para pRmeGR4b, indicándose el tamaño de los productos esperados. Más abajo se muestra la zona de solapamiento del plásmido de secuencia conocida pRmNT140. (C) Geles de agarosa al 0'8 % donde se observan los fragmentos de PCR amplificados con cada pareja de cebadores indicada en (B), así como con cada cebador de manera independiente. Las flechas señalan la banda del producto esperado en cada caso.

(Página siguiente) ►


C
a1
b1
b2
b3
a1F/R
a1F
a1R
b1F/R
b1F
b1R
b2F/R
b2F
b2R
b3F/R
b3F
b3R
MIII
515
457
427
353
203
190
158
138
095
083
24b
6as
e 6
ISRm15-a (2.197-81)
ISRm30-d (1-940)
ISRm23-a (1-942)
ISRm30-c (3.314-2.541)
ISRm30-b (2.540-2.044)
ISRm30- (2.043-1
34
CONTIG 127
CONTIG 8
CONTIG 19
CONTIG 18
CONTIG 173
CONTIG
3.277 pb
b2R
C18
4.105 pb
1
ISRm2011-2-a (549-1)
ISRm4-b (933-758)
ISRm4-a (753-252)
ISRm9 (2.798-1)
ISRm4 (933-1)
ISRm3 (1.298-1)
ISRm (1.48
154
CONTIG 157
CONTIG 5
CONTIG 176
CONTIG 22*
CONTIG 2
94R
nRmNT140

La unión de los extremos del andamiaje 6 para construir el plásmido pRmeGR4b fue más compleja (figura R1.8). El característico valor del FCI de este plásmido críptico (0'36; tabla R1.1) se utilizó para diferenciar *contigs* de copia única dentro del genoma de GR4 que debían incluirse en dicho plásmido por su similar valor del FCI (ya fueran *contigs* de elementos repetidos correspondientes a ISs con una sola copia en el genoma o *contigs* de secuencia única que el ensamblador había unido a otros andamiajes; tabla R1.S1 y R1.S2). Esto, junto con la información obtenida a través de la estrategia de las lecturas pareadas (con la que conocemos los *contigs* que flanquean un *contig* determinado), nos permitió ir cerrando esa unión de 32'6 kb formada por un puzzle de ISs y el *contig* de secuencia única 159 (de 14'6 kb, inicialmente ensamblado en el andamiaje 16). La estructura propuesta para esta zona se confirmó mediante una serie de amplificaciones llevadas a cabo con oligonucleótidos específicos de los *contigs* incluidos una sola vez en el plásmido (figura R1.8B y R1.8C). Posteriormente, tanto una secuenciación Sanger de los productos de PCR b1F/R y b4F/R como un alineamiento con la secuencia del plásmido pRmNT140 (proveniente de una librería de pRmeGR4b; Toro, 1985) corroboraron la secuencia propuesta.

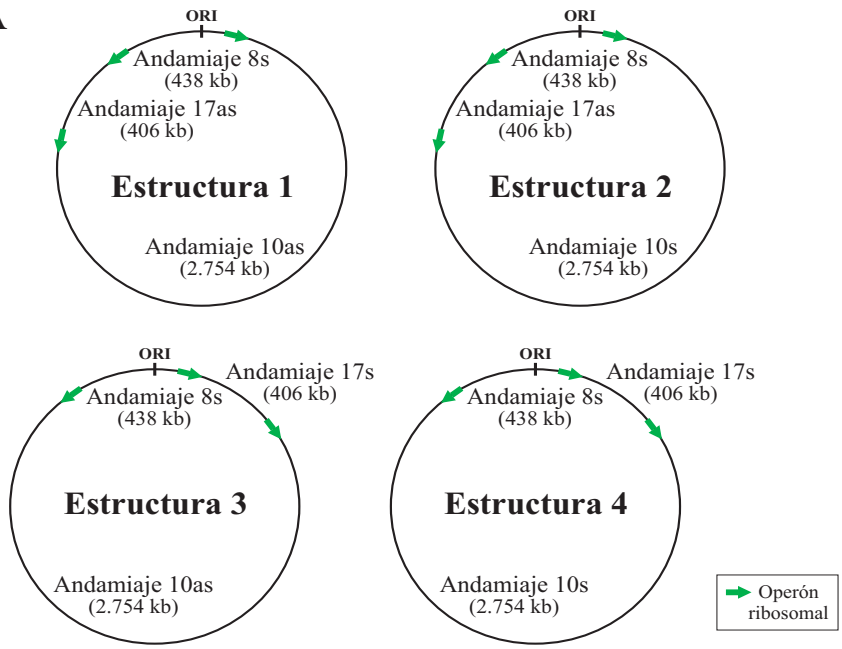
Los huecos presentes entre los tres andamiajes que componían el cromosoma de GR4 (andamiajes 8, 10 y 17) se correspondían con copias del operón ribosomal (6'3 kb; figura R1.7). Esta premisa sugería que, en principio, cualquier orden de los andamiajes era posible. Sin embargo, la orientación de este operón en los extremos de los tres andamiajes y la localización del

---

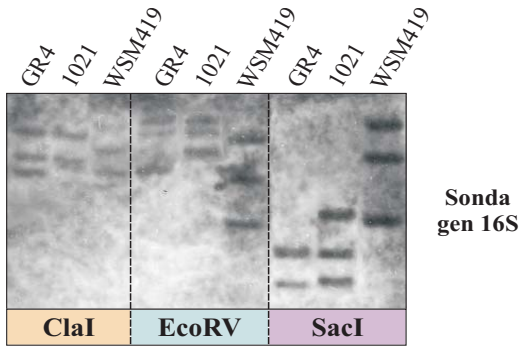
Figura R1.9: **Cierre del cromosoma de GR4.** (A) Esquema de las diferentes estructuras posibles para el cromosoma dependiendo del orden de los tres andamiajes que lo componen. (B) Hibridación con una sonda específica para el gen ribosómico 16S sobre distintas digestiones del DNA total de GR4, 1021 y WSM419: *Cla*I en color salmón, *Eco*RV en celeste, y *Sac*I en violeta. Debajo se muestran los tamaños de hibridación esperados para cada estructura cromosómica representada en (A), así como para las cepas secuenciadas 1021 y WSM419.

(Página siguiente) ►

A



B



Estructura 1	Estructura 2	Estructura 3	Estructura 4	1021	WSM419
15.785 12.596 9.550	13.152 12.596 12.181	15.920 11.933 8.852	12.596 12.176 11.933	15.836 12.597 11.273	18.930 14.682 12.075
20.471 18.085 13.059	20.471 17.695 13.447	24.698 11.972 6.938	20.471 17.690 11.972	20.472 18.136 14.782	16.981 12.165 9.203
8.196 8.027 6.919	8.196 8.027 6.919	10.003 8.025 6.918	10.003 8.027 6.916	9.919 8.028 6.970	21.281 14.990 9.685

ORI (en el andamiaje 8) permitieron reducir las estructuras posibles a cuatro (figura R1.9).

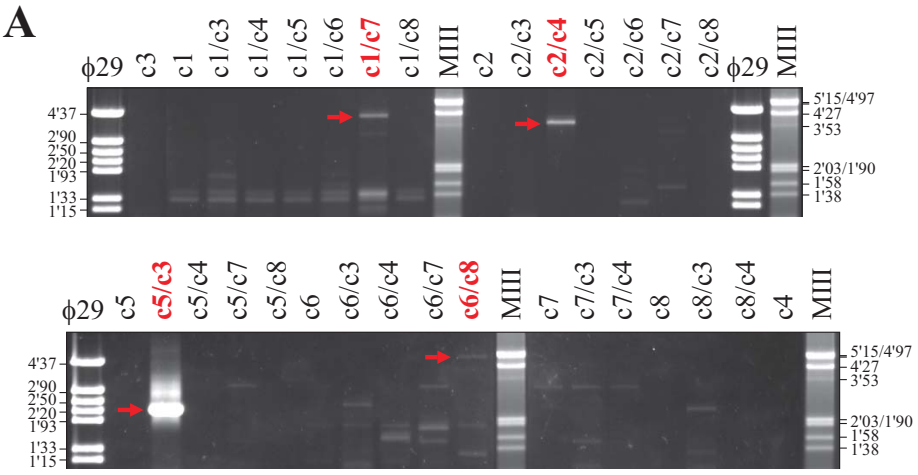
La estrategia seguida para determinar el orden y la orientación de los andamiajes en el cromosoma se basó en el perfil de hibridación generado por estas cuatro estructuras al digerir el genoma de GR4 de manera independiente con varias enzimas de restricción y, posteriormente, hibridar con una sonda específica para el gen ribosómico 16S. En la figura R1.9 aparece el perfil de GR4, 1021 y WSM419 cuando se digieren con las enzimas *Cla*I, *Eco*RV y *Sac*I, y se hibridan con la sonda del gen 16S. La tabla bajo la imagen muestra los tamaños de hibridación esperados para las distintas estructuras cromosómicas propuestas para GR4, así como para las cepas 1021 y WSM419. El perfil de estas dos cepas secuenciadas sirvió de referencia para determinar, de manera relativa, el tamaño de las bandas del perfil de hibridación de GR4. En el caso de la digestión con *Cla*I, las dos bandas superiores aparecidas en GR4 tenían un tamaño similar a las dos bandas más grandes de 1021 (de 15'8 kb y 12'6 kb respectivamente), siendo la tercera de menor tamaño en GR4 que en 1021. Este perfil descartaba las estructuras 2 y 4, no pudiendo diferenciar con exactitud las estructuras 1 y 3 al tratarse de tamaños de banda tan grandes. En la digestión con *Eco*RV, de nuevo, las dos bandas superiores de GR4 compartían tamaño con las bandas más altas de 1021 (de 20'5 kb y 18'1 kb respectivamente), lo cual descartaba la estructura cromosómica 3. El tamaño de las tres bandas encontradas en el perfil de GR4 generado con *Sac*I se parecía a las dos bandas inferiores de 1021 (de 8'0 kb y 7'0 kb respectivamente). Esto excluía las estructuras 3 y 4. Así, la única opción en la que concordaban todos los tamaños de banda era la estructura cromosómica 1. Por tanto, el orden de los andamiajes, y su orientación, dentro del cromosoma fue: andamiaje 8 en sentido (el cual contiene el ORI), andamiaje 10 en antisentido y andamiaje 17 en antisentido.

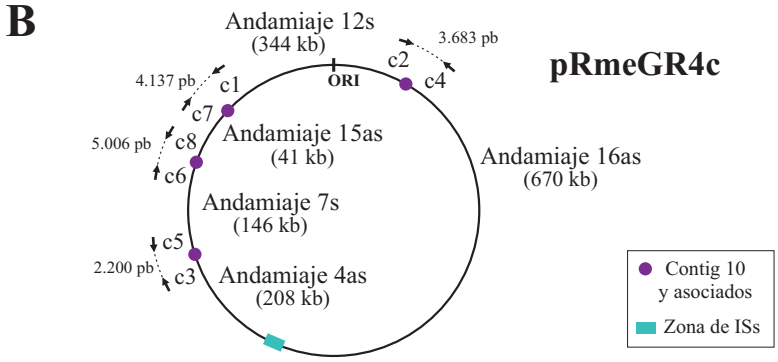
El plásmido simbiótico pRmeGR4c (correspondiente a pSymA en *S. meliloti* 1021) estaba compuesto por cinco andamiajes (4, 7, 12, 15 y 16). Con la estrategia de las lecturas pareadas descubrimos que en cuatro de los cinco huecos entre andamiajes se encontraba el *contig* 10, y que, en tres de ellos, esta-

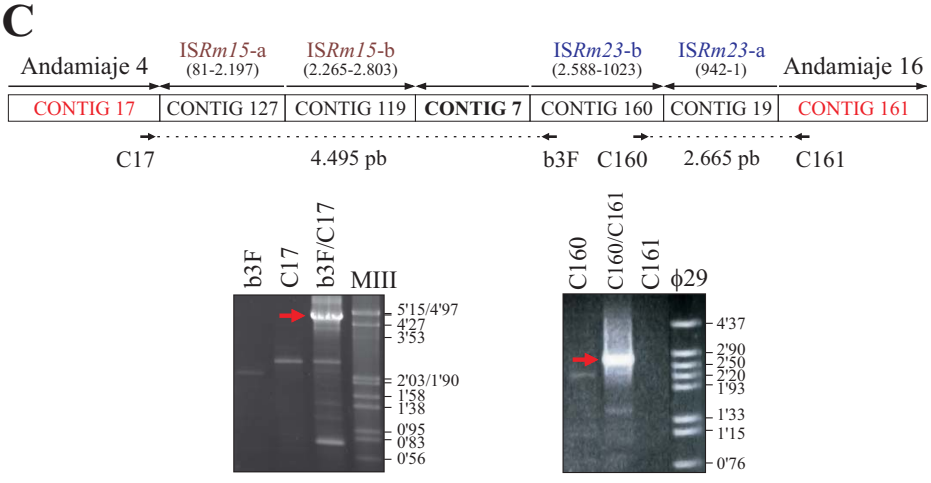
ba acompañado de los *contigs* 37 y 38 (figura R1.7). Un análisis mediante BLASTx y BLASTn de éstos respectivamente puso de manifiesto que el *contig* 10 presentaba un 74 % de identidad con el intrón del grupo II RmInt1, y los *contigs* 37 y 38 un 100 % de identidad con IS*Rm17*. A este nuevo intrón lo hemos denominado RmInt2, y su caracterización y comparación con el ya descrito RmInt1 (Martínez-Abarca *et al*, 1998) se presenta en el capítulo 3 de esta Tesis Doctoral. El hueco restante entre andamiajes se corresponde con una zona de ISs.

La estrategia seguida para el cierre de pRmeGR4c se basó en el diseño de cebadores específicos en los extremos de cada andamiaje. Un análisis de PCR con todas las combinaciones posibles de cebadores, salvo los que anillaban en los andamiajes que flanqueaban la zona de ISs, reveló el orden de los andamiajes dentro de este plásmido (figura R1.10A). Las reacciones de PCR llevadas a cabo con un solo cebador permitieron definir las bandas únicas generadas con cada pareja de cebadores (indicadas con una flecha en la figura R1.10A). Mediante secuenciación Sanger de las bandas correspondientes a cada unión, se confirmó la secuencia incluida en todos estos huecos entre andamiajes. La figura R1.10B muestra la estructura correcta del plásmido pRmeGR4c, que fue: andamiaje 12 en sentido (el cual contiene el ORI), andamiaje 16 en antisentido, andamiaje 4 en antisentido, andamiaje 7 en sentido y andamiaje 15 en antisentido. En ella también se indica la localización de los cebadores usados en las reacciones de PCR y el tamaño esperado para cada fragmento amplificado. El hueco entre los andamiajes 4 y 16 se abordó del mismo modo que la zona de ISs encontrada en pRmeGR4b, aunque en éste se incluyó un fragmento de menor tamaño (de 6'8 kb). Tras proponer un orden de los *contigs* en esa región, se estimó el tamaño de los productos de PCR generados a partir de la combinación de cebadores diseñados previamente (figura R1.10C). Los geles que aparecen en la figura R1.10C muestran la banda esperada para cada pareja de cebadores (indicada con una flecha).

El plásmido simbiótico pRmeGR4d (correspondiente a pSymB en *S. meliloti* 1021) estaba compuesto sólo por dos andamiajes (11 y 14), lo que reducía


A
ϕ29 c3 c1 c1/c3 c1/c4 c1/c5 c1/c6 c1/c7 c1/c8 MIII c2 c2/c3 c2/c4 c2/c5 c2/c6 c2/c7 c2/c8 ϕ29 MIII
4'37
2'90
2'50
2'20
1'93
1'33
1'15
5'15/4'97
4'27
3'53
2'03/1'90
1'58
1'38
ϕ29 c5 c5/c3 c5/c4 c5/c7 c5/c8 c6 c6/c3 c6/c4 c6/c7 c6/c8 MIII c7 c7/c3 c7/c4 c8 c8/c3 c8/c4 c4 MIII


B
Andamiaje 12s
(344 kb)
3.683 pb
c2
c4
4.137 pb
c1
ORI
c7
pRmeGR4c
5.006 pb
c8
Andamiaje 15as
(41 kb)
Andamiaje 16as
(670 kb)
c6
Andamiaje 7s
(146 kb)
c5
2.200 pb
c3
Andamiaje 4as
(208 kb)
Contig 10
y asociados
Zona de ISs


C
Andamiaje 4
ISRm15-a
(81-2.197)
ISRm15-b
(2.265-2.803)
ISRm23-b
(2.588-1023)
ISRm23-a
(942-1)
Andamiaje 16
CONTIG 17
CONTIG 127
CONTIG 119
CONTIG 7
CONTIG 160
CONTIG 19
CONTIG 161
C17
4.495 pb
b3F
C160
2.665 pb
C161
b3F
C17
b3F/C17
MIII
5'15/4'97
4'27
3'53
2'03/1'90
1'58
1'38
0'95
0'83
0'56
C160
C160/C161
C161
ϕ29
4'37
2'90
2'50
2'20
1'93
1'33
1'15
0'76

el número de estructuras posibles. Basándonos en la sintenia que presentaban estos andamiajes frente al plásmido pSymB de 1021, y en que ambos huecos debían contener los *contigs* 10, 37 y 38, propusimos la estructura representada en la figura R1.11A. Para confirmarla, y al igual que en la estrategia seguida para cerrar pRmeGR4c, diseñamos cebadores específicos en los extremos de los andamiajes (figura R1.11A). La reacción de PCR realizada con la pareja de cebadores d1/d3 dio como resultado un fragmento del tamaño esperado (indicado con una flecha en la figura R1.11B). Sin embargo, con la pareja d2/d4 no logramos conseguir el producto de PCR esperado (de unas 5'8 kb). En la figura R1.11C se muestra la disposición que debían tener los *contigs* 10, 37 y 38 en el hueco entre los *contigs* 142 (andamiaje 14) y 117 (andamiaje 11), estimado siguiendo la estrategia de las lecturas pareadas. La presencia de dos *contigs* 10 convergentes dentro de la región de amplificación d2/d4 puede ser el motivo por el cual no se gene-

---

Figura R1.10: **Cierre del plásmido simbiótico pRmeGR4c.** (A) Geles de agarosa al 0'8 % donde se muestran los productos de PCR de las distintas combinaciones de cebadores usados para determinar el orden de los andamiajes dentro de pRmeGR4c. En rojo se indica la pareja de cebadores que genera el tamaño de banda esperado para cada unión, cuyo producto de PCR se señala con una flecha. (B) Esquema del orden de los andamiajes que componen pRmeGR4c de acuerdo con las parejas de cebadores que han generado un amplicón del tamaño esperado. Se muestra la longitud del producto de PCR indicado con flechas en (A). (C) Detalle de los *contigs* presentes en la zona de ISs de pRmeGR4c. Los *contigs* de secuencia única que el programa Newbler incluyó en este plásmido se destacan en rojo. Las flechas encima de los *contigs* muestran su orientación. Encima de las flechas se indica el elemento repetido al que tiene similitud cada *contig*, mostrándose, entre paréntesis, los nucleótidos y la orientación de dicho elemento en el *contig*. Cada elemento repetido con similitud a más de un *contig* aparece en un color diferente, y los fragmentos del elemento se representan con diferentes letras. Debajo de los *contigs* se muestra la localización de los cebadores usados para confirmar la estructura de la zona de ISs propuesta para pRmeGR4c, indicándose, además, el tamaño de los productos esperados. Más abajo aparecen los geles de agarosa al 0'8 % donde se observan los productos de PCR esperados (señalados con flecha).

◀ (Página anterior)

ró el producto de PCR esperado. Durante el periodo de anillamiento en la reacción de PCR (tras la desnaturalización del DNA), la secuencia de ambos *contigs* 10 pueden unirse y generar una estructura secundaria que impida la amplificación. Por tanto, y a pesar de la falta de amplificación con la pareja de cebadores d2/d4, el orden de los andamiajes dentro de pRmeGR4d

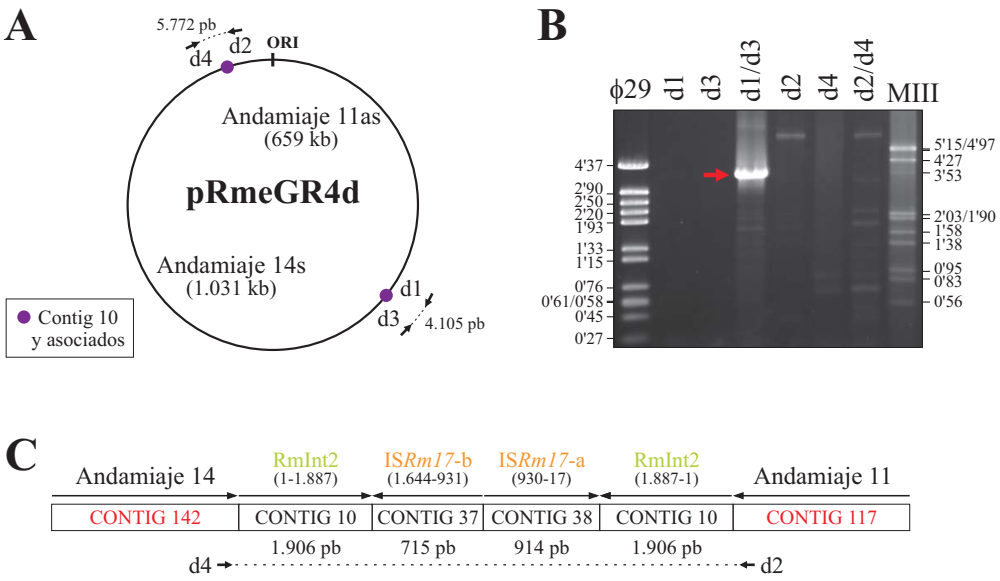
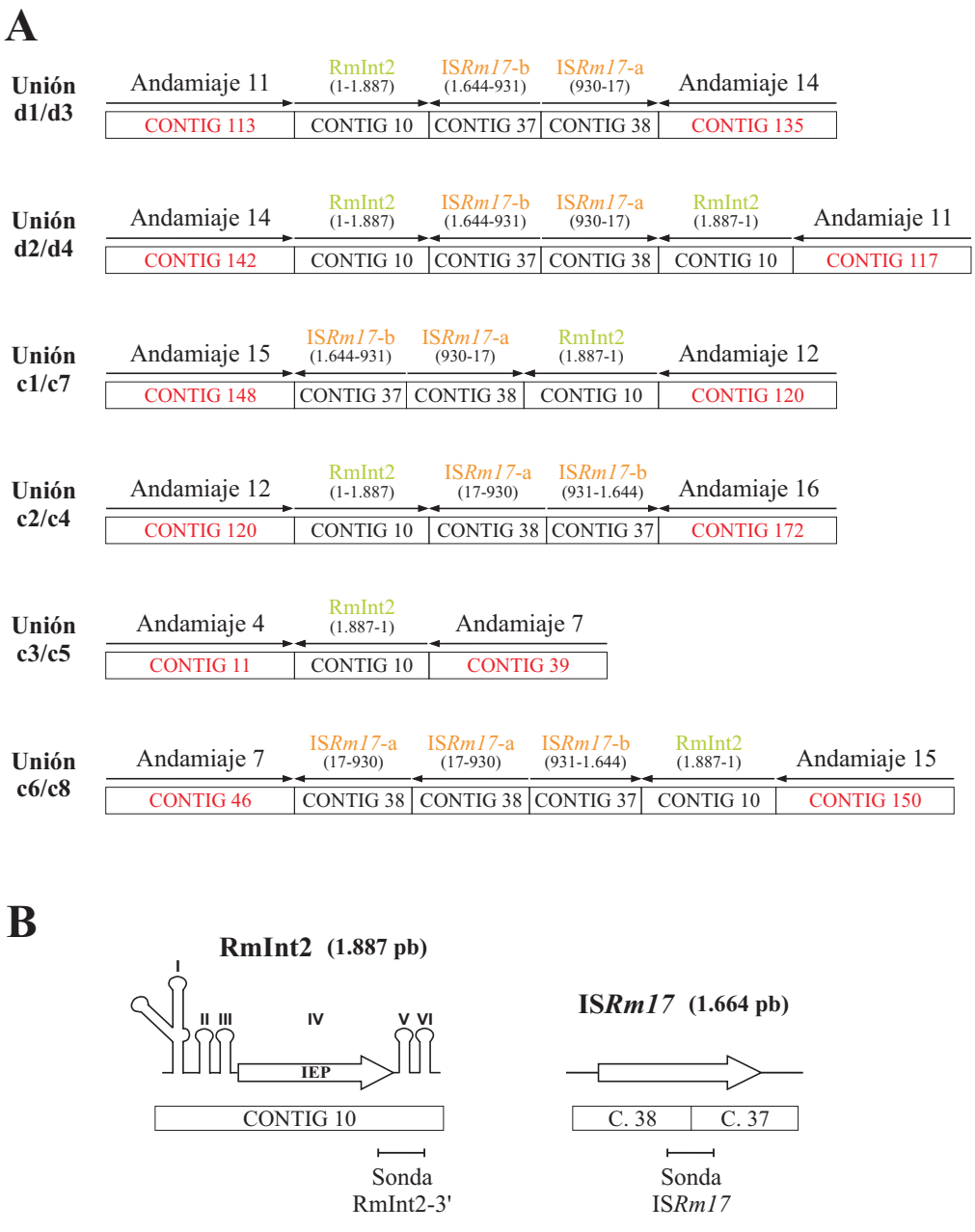


Figura R1.11: **Cierre del plásmido simbiótico pRmeGR4d.** (A) Esquema del orden de los andamiajes propuesto para este plásmido. Se muestra la localización y la longitud del producto de PCR esperado con cada pareja de cebadores. (B) Gel de agarosa al 0'8 % donde, señalado con flecha, aparece el producto de PCR generado por la pareja de cebadores d1/d3. (C) Detalle de los *contigs* propuestos para cerrar la unión entre los cebadores d4 y d2. Los *contigs* de secuencia única que el programa Newbler incluyó en este plásmido se destacan en rojo. Las flechas encima de los *contigs* muestran su orientación. Encima de las flechas se indica el elemento repetido al que tiene similitud cada *contig*, mostrándose, entre paréntesis, los nucleótidos y la orientación de dicho elemento en el *contig*. Cada elemento repetido con similitud a más de un *contig* aparece en un color diferente, y los fragmentos del elemento se representan con diferentes letras. Debajo de los *contigs* se indica el tamaño de cada uno, y se muestra la localización de los cebadores usados para confirmar la unión de los andamiajes 14 y 11.

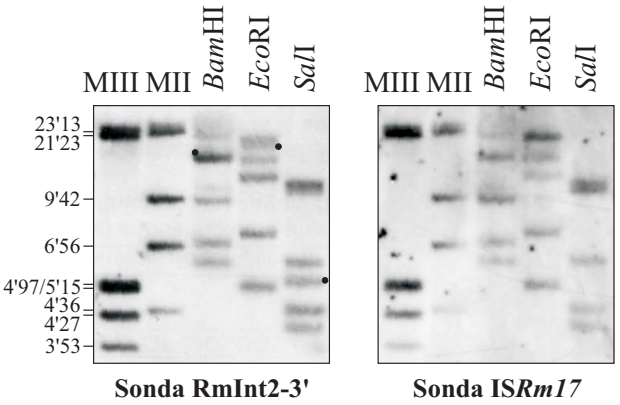
fue: andamiaje 11 en antisentido (el cual contiene el ORI) y andamiaje 14 en sentido.

Los 2 huecos entre andamiajes que cerraban pRmeGR4d, y 4 de los 5 huecos que completaban pRmeGR4c correspondían al intrón del grupo II RmInt2 (*contig* 10) y, en algunos casos, también a la IS a la que se encuentra asociado (IS*Rm17*, formada por los *contigs* 37 y 38). Sin embargo, en todos ellos, la disposición de los *contigs* fue distinta dependiendo de la orientación del elemento repetido y el número de veces que aparecía repetido dentro de cada hueco (figura R1.12A). En la unión de uno de los huecos entre andamiajes de pRmeGR4d (unión d1/d3) y de 2 huecos en pRmeGR4c (uniones c1/c7 y c2/c4) se introdujo una secuencia de 3'6 kb, correspondiente a los *contigs* 10, 37 y 38. La unión de andamiajes representada por la pareja de cebadores c6/c8 en pRmeGR4c contenía una copia adicional de IS*Rm17* parcial (*contig* 38), por lo que se introdujo una secuencia de 4'5 kb, mientras que en la unión d2/d4 en pRmeGR4d se encontraron 2 copias de RmInt2 (*contig* 10) y el hueco se rellenó con un fragmento de 5'5 kb. El hueco entre los andamiajes 4 y 7 de pRmeGR4c (unión c3/c5) se completó con la secuencia del *contig* 10, previamente excluido del andamiaje 4 por ser considerado un *contig* de elemento repetido (figura R1.12A).

El resultado de la hibridación de DNA genómico de GR4 con sondas específicas para RmInt2 e IS*Rm17*, previa digestión con las enzimas de restricción *Bam*HI, *Eco*RI y *Sal*I, corroboró la disposición de los *contigs* 10, 37 y 38 en los huecos entre andamiajes propuesta para pRmeGR4c y pRmeGR4d (figura R1.12C). Los tamaños de banda esperados para cada una de las digestiones e hibridaciones aparecen en la figura R1.12D. La unión constituida por la pareja de cebadores c3/c5, en pRmeGR4c, presentó una banda correspondiente al *contig* 10 (sonda RmInt2-3') que no apareció tras la hibridación con la sonda IS*Rm17*. La banda correspondiente al *contig* 10 en la unión d2/d4, en pRmeGR4d, presentó una intensidad doble respecto a las bandas de tamaño similar, que no se observó al hibridar con la sonda IS*Rm17*. Lo mismo ocurrió con la banda correspondiente a la unión c6/c8,


A
Unión d1/d3
Andamiaje 11
RmInt2 (1-1.887)
ISRm17-b (1.644-931)
ISRm17-a (930-17)
Andamiaje 14
CONTIG 113
CONTIG 10
CONTIG 37
CONTIG 38
CONTIG 135
Unión d2/d4
Andamiaje 14
RmInt2 (1-1.887)
ISRm17-b (1.644-931)
ISRm17-a (930-17)
RmInt2 (1.887-1)
Andamiaje 11
CONTIG 142
CONTIG 10
CONTIG 37
CONTIG 38
CONTIG 10
CONTIG 117
Unión c1/c7
Andamiaje 15
ISRm17-b (1.644-931)
ISRm17-a (930-17)
RmInt2 (1.887-1)
Andamiaje 12
CONTIG 148
CONTIG 37
CONTIG 38
CONTIG 10
CONTIG 120
Unión c2/c4
Andamiaje 12
RmInt2 (1-1.887)
ISRm17-a (17-930)
ISRm17-b (931-1.644)
Andamiaje 16
CONTIG 120
CONTIG 10
CONTIG 38
CONTIG 37
CONTIG 172
Unión c3/c5
Andamiaje 4
RmInt2 (1.887-1)
Andamiaje 7
CONTIG 11
CONTIG 10
CONTIG 39
Unión c6/c8
Andamiaje 7
ISRm17-a (17-930)
ISRm17-a (17-930)
ISRm17-b (931-1.644)
RmInt2 (1.887-1)
Andamiaje 15
CONTIG 46
CONTIG 38
CONTIG 38
CONTIG 37
CONTIG 10
CONTIG 150
B
RmInt2 (1.887 pb)
I
II III
IV
V VI
IEP
CONTIG 10
Sonda RmInt2-3'
ISRm17 (1.664 pb)
C. 38
C. 37
Sonda ISRm17

C



D

	Sonda RmInt2			Sonda IS*Rm17*		
Unión	***Bam*HI**	***Eco*RI**	***Sal*I**	***Bam*HI**	***Eco*RI**	***Sal*I**
c2/c4	6.469	4.876	3.862	6.469	4.876	3.862
c3/c5	14.480	16.709	5.083	-	-	-
c6/c8	9.064	18.786	9.766	9.064	18.786	9.766
c1/c7	5.688	13.724	5.754	5.688	13.724	5.754
d1/d3	21.257	6.929	4.284	21.257	6.929	4.284
d2/d4	14.401	11.283	10.382	14.401	11.283	10.382

Figura R1.12: **Localización del intrón del grupo II RmInt2 y de IS*Rm17* en el genoma de GR4.** (A) Esquema con el orden y orientación de los *contigs* 10, 37 y 38 en los huecos entre andamiajes correspondientes a la unión de la pareja de cebadores indicada en cada caso, representadas en las figuras R1.10 (pRmeGR4c) y R1.11 (pRmeGR4d). Los *contigs* de secuencia única se destacan en rojo. Las flechas encima de los *contigs* muestran su orientación. Encima de las flechas se indica el elemento repetido al que tiene similitud cada *contig* (diferenciados por colores), mostrándose, entre paréntesis, los nucleótidos y la orientación de dicho elemento en el *contig*. (B) Localización de la zona donde hibrida cada una de las sondas en estos elementos repetidos. (C) Hibridación del genoma de GR4 con sondas específicas para RmInt2 (*contig* 10) e IS*Rm17* (*contigs* 37 y 38). Con un punto se señala la unión c3/c5 (carente de IS*Rm17*) en la hibridación con la sonda RmInt2-3′. (D) Tamaños de banda esperados según la localización de estos elementos repetidos en los plásmidos pRmeGR4c y pRmeGR4d.

en pRmeGR4c, cuya intensidad tras la hibridación con la sonda IS*Rm17* fue el doble que al hibridar con la sonda RmInt2-3'.

R1.3 CARACTERÍSTICAS DEL GENOMA DE *S. meliloti* GR4

La información extraída del genoma de GR4 tras su anotación automática (ver Material y Métodos, apartado M.12.4.2) y curación manual (ver Material y Métodos, apartado M.12.4.3) se detalla en la tabla R1.3. El tamaño y porcentaje G+C del cromosoma y los plásmidos pRmeGR4a, pRmeGR4b, pRmeGR4c y pRmeGR4d son 3.618.794 pb (62'8 %), 175.986 pb (60'0 %), 225.725 pb (58'6 %), 1.417.856 pb (60'4 %) y 1.701.197 pb (62'4 %) respectivamente. De esas 7.139.558 pb, 6.157.924 pb componen la región codificante del genoma de GR4 (un 86'3 %), la cual está constituida por 6.707 genes de los 6.837 genes totales que presenta dicha cepa. La longitud media de los genes codificantes dentro del genoma de GR4 es 896 pb. Esta longitud varía entre replicones, siendo pRmeGR4b y pRmeGR4d los que, como media, contienen genes de menor y mayor tamaño respectivamente (655 pb y 963 pb). Un 70'3 % del total de genes se encuentran asociados a una o varias de las categorías funcionales establecidas por los conjuntos de grupos ortólogos (COG, *Clusters of Orthologous Groups*; Tatusov *et al*, 2000) de proteínas, y un 74'7 % están relacionados con familias de proteínas presentes en la base de datos Pfam (Finn *et al*, 2006).

El servidor IMG ER dispone de una herramienta para generar un mapa circular del replicón anotado en cada caso, representando así gráficamente parte de la información detallada en la tabla R1.3. En la figura R1.13 aparece el mapa generado para los cinco replicones que constituyen el genoma de GR4. En ella se muestran los genes de RNA encontrados, diferenciando entre tRNAs, rRNAs y otros RNAs. Al igual que para las otras cepas ya secuenciadas, los tres operones ribosomales que contiene GR4 se localizan en el cromosoma, donde también encontramos la mayoría de los tRNAs (53 genes) que presenta esta cepa. Los otros dos tRNAs están situados uno en el plásmido pRmeGR4c y otro en pRmeGR4d. La distribución de los

genes miscRNA encontrados en este rizobio son: 23 en el cromosoma, 1 en pRmeGR4a, 8 en pRmeGR4b, 21 en pRmeGR4c y 13 en pRmeGR4d. En la figura R1.13 se observan, además, los genes presentes en ambas cadenas, sentido y antisentido, según el código de color usado para diferenciar las categorías funcionales de los COGs, cuya distribución se recoge en la tabla R1.4. El 30 % del total de genes no tiene una función COG asignada, y casi un 20 % de los genes asociados a COGs se relaciona con funciones generales (10'21 %) o desconocidas (9'11 %). Del resto de categorías funcionales, la más representada es replicación, recombinación y reparación (11'60 %),

Tabla R1.3: Información sobre el genoma de *S. meliloti* GR4

Característica	**Cromosoma**	**pRmeGR4a**	**pRmeGR4b**	**pRmeGR4c**	**pRmeGR4d**	**Genoma**
Longitud (pb)	3.618.794	175.986	225.725	1.417.856	1.701.197	7.139.558
Contenido G+C (%)	62'8	60'0	58'6	60'4	62'4	62'0
Región codificante (%)	86'8	85'6	78'9	84'0	87'9	86'3
Número total de genes	3.413	185	264	1.427	1.548	6.837
Genes de RNA	85	1	8	22	14	130
Genes codificantes	3.328	184	256	1.405	1.534	6.707
Longitud media de genes codificantes (pb)	919	817	655	825	963	896
Pseudogenes	9	0	9	20	6	44
Genes con función predicha (%)	84'50	59'46	59'85	72'33	81'99	71'63
Genes relacionados con enzimas (%)	26'24	9'19	3'65	13'87	18'71	14'3
Genes relacionados con transporte (%)	15'29	11'89	7'3	16'01	22'57	14'6
Genes asignados a rutas KEGG (%)	30'9	13'5	8'8	17'3	24'4	19'0
Genes asignados a KO (%)	54'4	26'5	20'4	36'4	46'3	36'8
Genes asignados a rutas MetaCyc (%)	25'8	8'7	3'7	13'7	18'6	14'1
Genes asignados a COGs (%)	84'3	56'8	56'9	72'0	81'7	70'3
Genes asignados a KOGs (%)	36'3	16'8	12'0	25'3	31'3	24'3
Genes asignados a Pfam (%)	86'3	63'8	64'2	74'5	84'9	74'7
Genes asignados a TIGRfam (%)	30'5	13'0	8'0	15'0	19'7	17'2
Genes en grupos parálogos (%)	31'6	13'0	27'7	39'6	43'3	31'1
Genes codificantes de péptidos señal (%)	20'4	30'8	10'2	21'1	22'1	20'9
Genes codificantes de proteínas transmembrana (%)	2'4	29'7	9'5	23'0	23'3	17'6

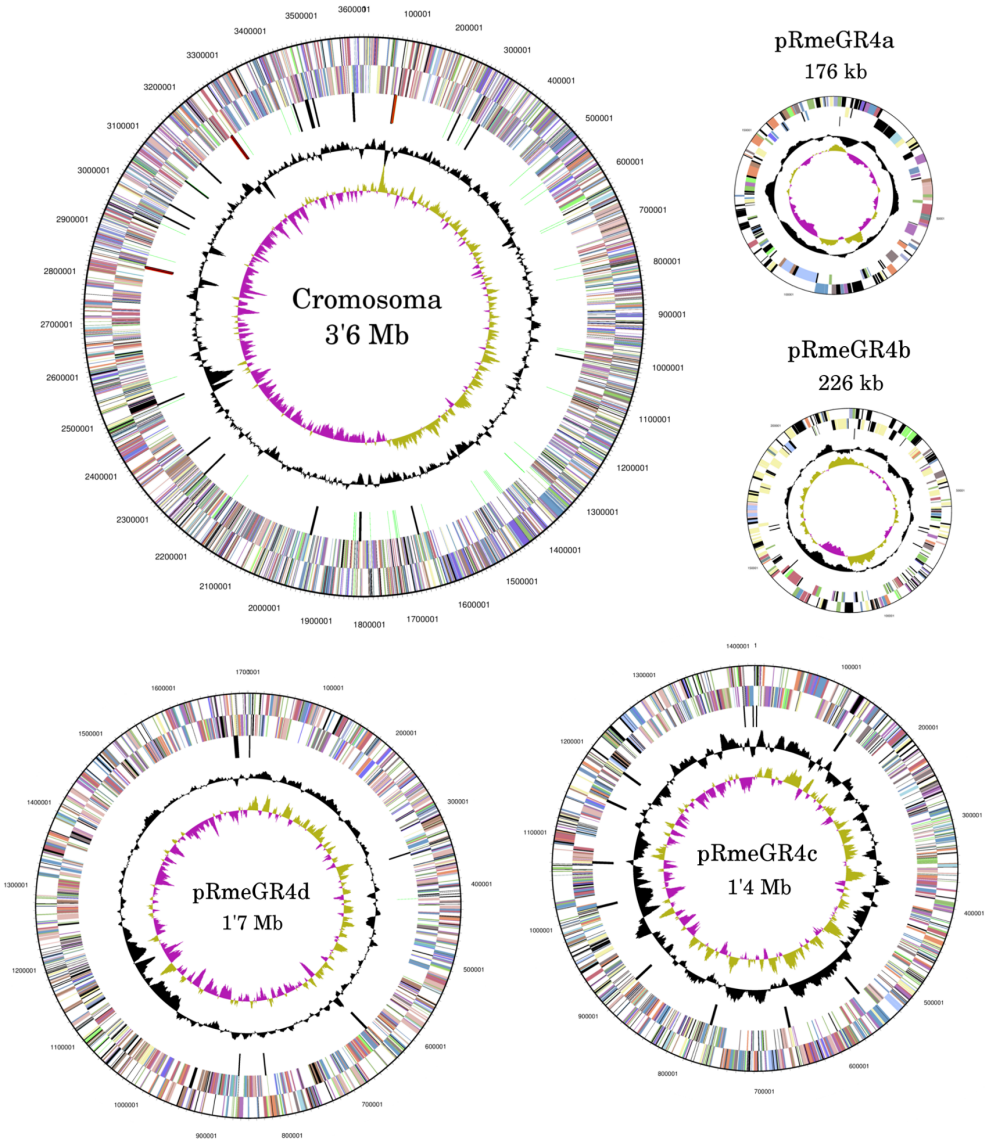


Figura R1.13: **Mapa génico de los cinco replicones del genoma de GR4 generados con el servidor IMG ER.** Diagrama del cromosoma y los plásmidos de *S. meliloti* GR4 no representados a escala. Los dos círculos más externos muestran los genes presentes en la cadena sentido (el externo) y antisentido (el interno), representados según el color del código COG mostrado en la figura R1.14B. El anillo central muestra los genes de RNA: en verde los tRNAs, en rojo los rRNAs y en negro otros RNAs. Los círculos interiores indican el contenido G+C relativo (en negro) y la desviación GC (verde/morado).

seguida de transcripción (9'71 %), metabolismo y transporte de aminoácidos (9'15 %) y metabolismo y transporte de carbohidratos (7'96 %).

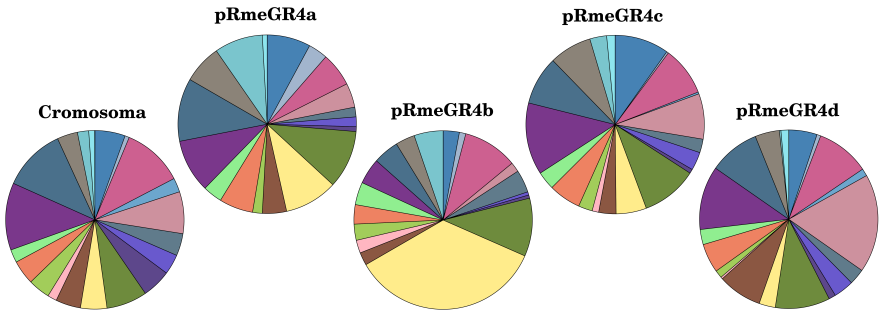
Tabla R1.4: Porcentaje de genes de la cepa *S. meliloti* GR4 que se encuentra asociado a cada una de las categorías funcionales establecidas por los COGs.

Código	**Categorías funcionales de los COGs**	**% genes**[a]
A	RNA processing and modification	0'00
B	Chromatin structure and dynamics	0'01
C	Energy production and conversion	6'25
D	Cell cycle control, cell division, chromosome partitioning	1'30
E	Amino acid transport and metabolism	9'15
F	Nucleotide transport and metabolism	0'85
G	Carbohydrate transport and metabolism	7'96
H	Coenzyme transport and metabolism	3'04
I	Lipid transport and metabolism	2'59
J	Translation, ribosomal structure and biogenesis	1'75
K	Transcription	9'71
L	Replication, recombination and repair	11'60
M	Cell wall/membrane/envelope biogenesis	4'48
N	Cell motility	1'11
O	Posttranslational modification, protein turnover, chaperones	2'52
P	Inorganic ion transport and metabolism	5'03
Q	Secondary metabolites biosynthesis, transport and catabolism	3'20
R	General function prediction only	10'21
S	Function unknown	9'11
T	Signal transduction mechanisms	5'28
U	Intracellular trafficking, secretion, and vesicular transport	3'88
V	Defense mechanisms	0'98
W	Extracellular structures	0'01
Y	Nuclear structure	0'00
Z	Cytoskeleton	0'00
NA	Not in COGs	29'66[b]

[a] porcentaje de genes asignados a un COG respecto al total de genes asociados a COGs
[b] porcentaje de genes no asociados a COGs respecto al total de genes presentes en esta bacteria

Sobre los genes que se engloban en una o varias de las categorías funcionales de los COGs, llevamos a cabo un análisis más exhaustivo. La figura R1.14 muestra la proporción de genes que se corresponde con cada una de las categorías funcionales establecidas por los COGs de proteínas en los cinco replicones de GR4, cuyo porcentaje se recoge en la tabla R1.S3. Casi un cuarto de los genes del cromosoma no tienen función conocida (11'45 %) o se incluyen en funciones generales (12'17 %); los genes más abundantes después de éstos son los relacionados con metabolismo y transporte de aminoácidos (11'07 %). La categoría funcional que engloba la traducción, estructura ribosomal y biogénesis comprende el 5'25 % de los genes del cromosoma, no superando el 1'3 % en ninguno de los otros replicones, confirmando de esta forma las características cromosómicas del replicón de mayor tamaño. En relación a los plásmidos crípticos, pRmeGR4a contiene un gran porcentaje de genes con función desconocida (11'40 %), aunque se encuentra una proporción similar de genes asociados con transcripción (10'53 %), replicación, recombinación y reparación (9'65 %) y funciones generales (9'65 %); pRmeGR4b contiene el porcentaje más elevado de genes involucrados en replicación, recombinación y reparación (35'09 %), y también presenta un alto número de genes asignados a transcripción (10'53 %) y metabolismo y transporte de aminoácidos (9'94 %). El megaplásmido simbiótico pRmeGR4c presenta gran cantidad de genes relacionados con funciones generales (13'05 %), transcripción (10'17 %) y conversión y producción de energía (9'83 %). El otro megaplásmido simbiótico, pRmeGR4d, contiene el mayor número de genes involucrados en metabolismo y transporte de carbohidratos (17'85 %). Los genes con funciones generales también se encuentran en elevada proporción (11'52 %), al igual que los genes asignados a transcripción (9'91 %) y sin función conocida (9'21 %). La categoría funcional que engloba la biogénesis de la membrana y pared celular presenta mayor porcentaje de genes en este megaplásmido (8'01 %) que en los otros replicones, donde varía entre el 2'34 % y el 4'6 %.

Sobre el genoma de GR4 también llevamos a cabo un análisis de la distribución de la secuencia octomérica GGGCAGGG (un motivo denominado



COG Code	COG Function Definition
[A]	RNA processing and modification
[B]	Chromatin structure and dynamics
[C]	Energy production and conversion
[D]	Cell cycle control, cell division, chromosome partitioning
[E]	Amino acid transport and metabolism
[F]	Nucleotide transport and metabolism
[G]	Carbohydrate transport and metabolism
[H]	Coenzyme transport and metabolism
[I]	Lipid transport and metabolism
[J]	Translation, ribosomal structure and biogenesis
[K]	Transcription
[L]	Replication, recombination and repair
[M]	Cell wall/membrane/envelope biogenesis
[N]	Cell motility
[O]	Posttranslational modification, protein turnover, chaperones
[P]	Inorganic ion transport and metabolism
[Q]	Secondary metabolites biosynthesis, transport and catabolism
[R]	General function prediction only
[S]	Function unknown
[T]	Signal transduction mechanisms
[U]	Intracellular trafficking, secretion, and vesicular transport
[V]	Defense mechanisms
[W]	Extracellular structures
[Y]	Nuclear structure
[Z]	Cytoskeleton
[NA]	Not Assigned

Figura R1.14: **Distribución de los genes codificantes que se incluyen en las distintas categorías funcionales de los COGs en los cinco replicones de *S. meliloti* GR4.** Se muestran los diagramas con la proporción de genes que codifican proteínas con COGs presentes en cada replicón de GR4 y el código de color utilizado para representar las categorías funcionales de los COGs.

KOPS, del inglés FtsK *Orienting Polar Sequences*). Su orientación presenta un sesgo hacia la cadena de replicación líder, y su frecuencia aumenta cerca del sitio de terminación de la replicación (Bigot *et al*, 2007). Este patrón de distribución se aprecia más claramente en el cromosoma de GR4 que en los plásmidos (figura R1.15). La región estimada para el origen y término del cromosoma se encuentra hacia la posición 3.600.000 y 1.700.000 respectivamente. La acumulación de copias del motivo KOPS es mayor en el término del replicón que en su inicio, apareciendo (como promedio de las 50 copias más cercanas a cada región) una copia cada 7 kb y 22 kb respectivamente. El cromosoma presenta 276 copias de este motivo, de las que 253 se encuentran orientadas desde el origen hacia el término de la replicación. Los plásmidos crípticos son los que presentan menor número de copias del motivo

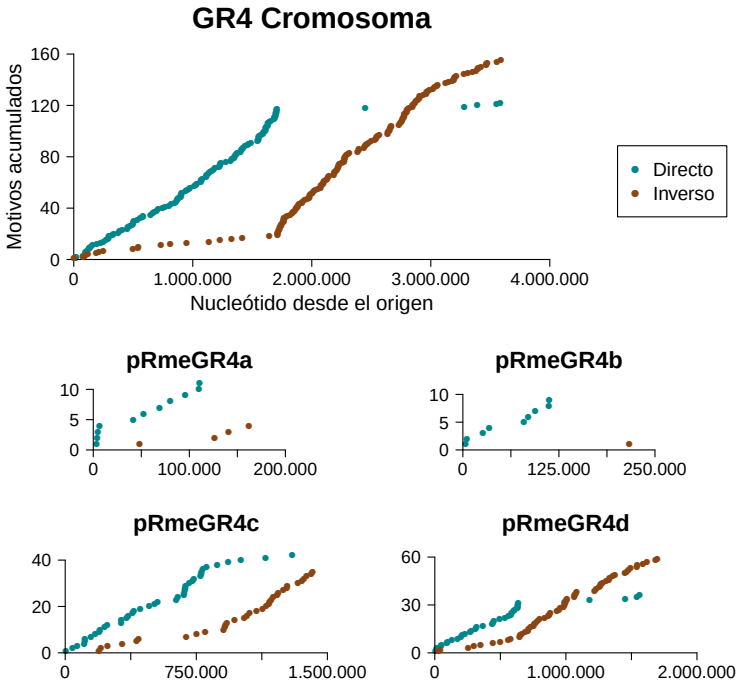


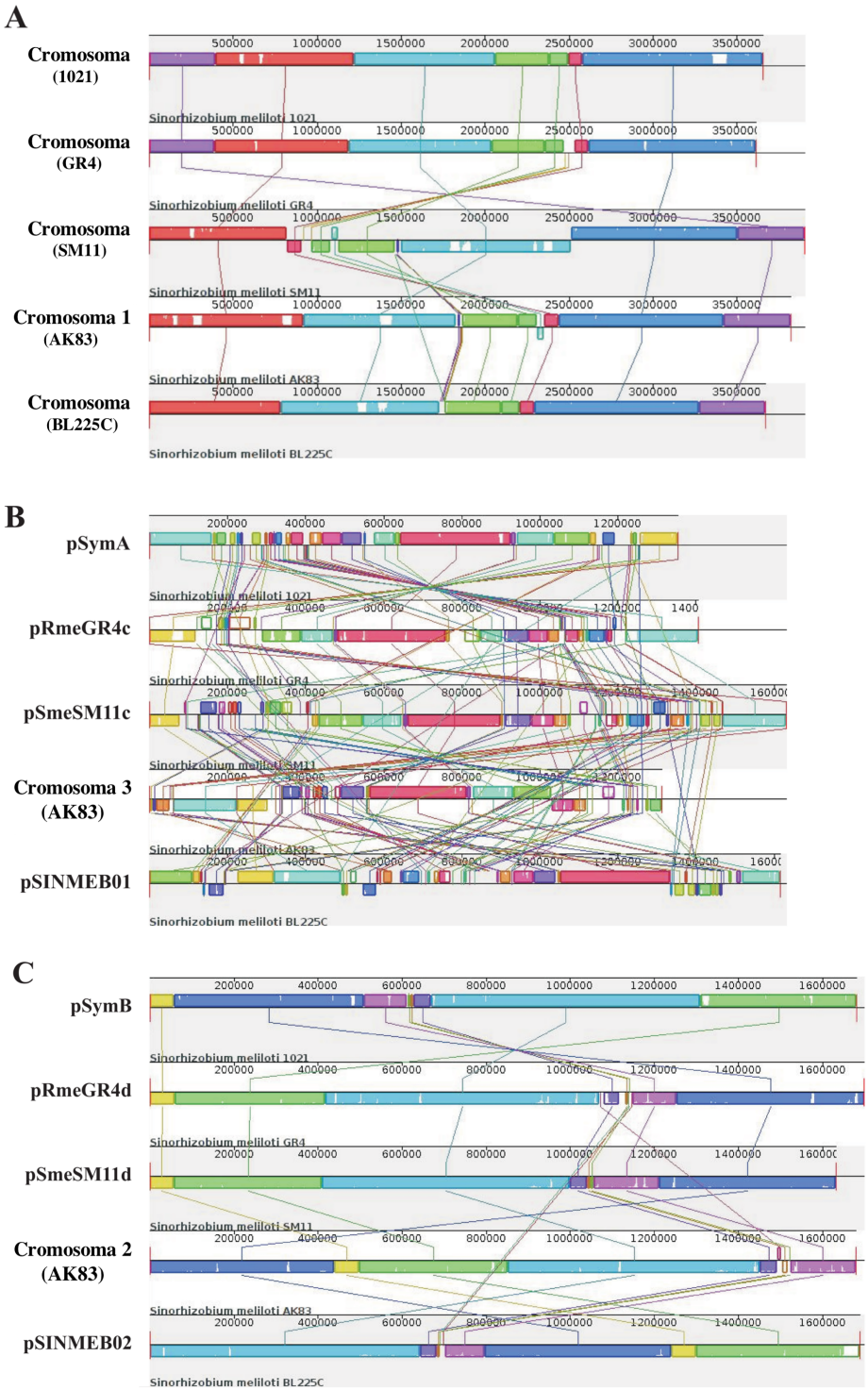
Figura R1.15: **Distribución del motivo octomérico GGGCAGGG en el genoma de *S. meliloti* GR4.** Se muestra la acumulación del número de motivos KOPS en orientación directa e inversa en todos los replicones de esta cepa.

KOPS: pRmeGR4a contiene 15 copias y pRmeGR4b 10 copias. Los plásmidos simbióticos muestran un patrón de distribución de este motivo similar al del cromosoma: en pRmeGR4c aparecen 76 copias y en pRmeGR4d 95 copias, encontrándose 62 y 81 de ellas, respectivamente, orientadas desde el origen hacia el término de la replicación.

R1.4 GENÓMICA ESTRUCTURAL DE LA ESPECIE *S. meliloti*

A principios del año 2012 se encontraban disponibles en las bases de datos el genoma completo de cuatro cepas bacterianas de la especie *S. meliloti* (1021, Galibert *et al*, 2001; SM11, Schneiker-Bekel *et al*, 2011; y AK83 y BL225C, Galardini *et al*, 2011b) y el borrador del genoma de la cepa *S. meliloti* CCNWSX0020 (Li *et al*, 2012b). Posteriormente, se ha publicado en las bases de datos el genoma de *S. meliloti* Rm41 (Weidner *et al*, 2013) y *S. meliloti* 2011. Los estudios de genómica comparada presentados en este trabajo se han llevado a cabo con las cuatro cepas completamente secuenciadas hasta 2012 y la cepa GR4.

La figura R1.16 muestra el alineamiento del genoma de las cinco cepas analizadas. En ella se observa que el cromosoma de todas las cepas presentan la misma disposición de bloques homólogos, salvo el de SM11, donde se aprecia una inversión (figura R1.16A). Los plásmidos tipo pSymA son los que presentan mayor diferencia en la distribución de los bloques homólogos entre cepas (figura R1.16B). Entre estos plásmidos también se observan diferencias en la orientación de la secuencia, hecho que se refleja en la disposición de los bloques por encima o por debajo de la línea que representa la secuencia del replicón, si está en sentido o antisentido respectivamente. Los plásmidos tipo pSymB, al igual que el cromosoma, conservan la distribución de bloques homólogos entre cepas (figura R1.16C). Sin embargo, los plásmidos crípticos comparten pocas regiones homólogas entre cepas (figura R1.16D y R1.16E), y además, esos bloques presentan zonas de bajo grado de conservación, (representadas como regiones blancas dentro de los bloques homólogos).


A
Cromosoma (1021)
Cromosoma (GR4)
Cromosoma (SM11)
Cromosoma 1 (AK83)
Cromosoma (BL225C)
Sinorhizobium meliloti 1021
Sinorhizobium meliloti GR4
Sinorhizobium meliloti SM11
Sinorhizobium meliloti AK83
Sinorhizobium meliloti BL225C
B
pSymA
pRmeGR4c
pSmeSM11c
Cromosoma 3 (AK83)
pSINMEB01
C
pSymB
pRmeGR4d
pSmeSM11d
Cromosoma 2 (AK83)
pSINMEB02

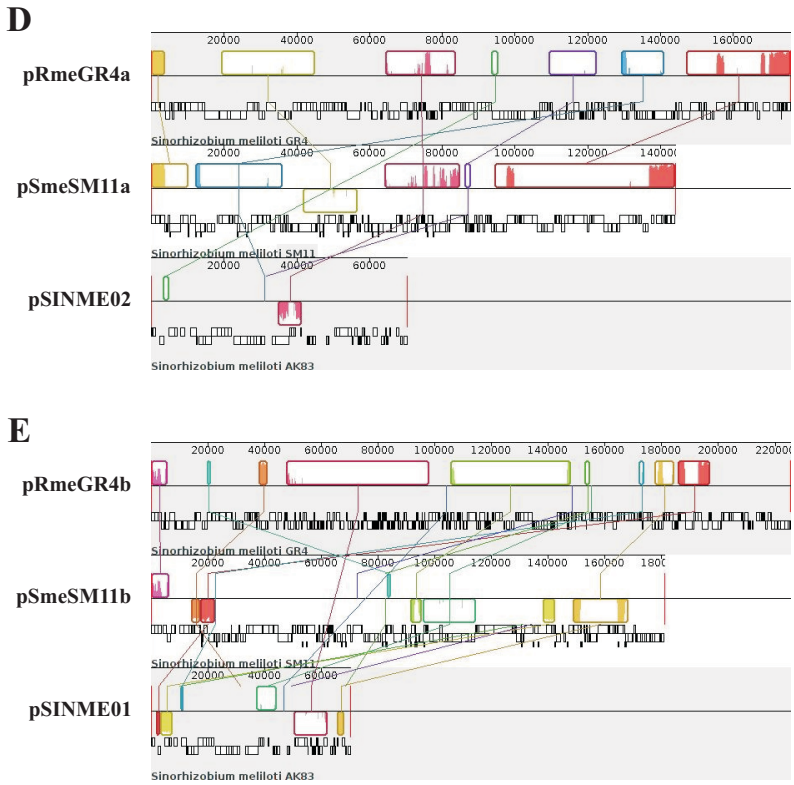


Figura R1.16: **Alineamiento del genoma de las cepas 1021, GR4, SM11, AK83 y BL225C.** En el alineamiento de los cromosomas (A), plásmidos tipo pSymA (B) y plásmidos tipo pSymB (C), se ha tomado como secuencia de referencia 1021, mientras que en el alineamiento de los plásmidos crípticos (D y E) se ha usado GR4. Las regiones de secuencia homóloga que comparten todas las cepas se muestran como bloques de un mismo color unidos por una línea. Las zonas blancas que hay en algunos bloques representan secuencias específicas de cada cepa en esa región. Los bloques que aparecen por encima o debajo de la línea central de un genoma se encuentran en la misma orientación o en orientación inversa en relación al genoma de referencia respectivamente.

Los datos obtenidos con el alineamiento del genoma completo de las cinco cepas analizadas permitió inferir el número de genes únicos presentes en cada una de ellas así como el número de genes que forman el pangenoma (11.362 genes) y el genoma común (4.778 genes) de la especie *S. meliloti* (figura R1.17). SM11 y GR4 son las que contienen un mayor número de genes únicos, 1.478 y 1.474 respectivamente. La cepa AK83 presenta 851 genes de este tipo, mientras que en 1021 y BL225C el número es menor (558 y 565 genes únicos respectivamente). Dentro del genoma de GR4, el cromosoma contiene 3.054 genes pertenecientes al genoma común de *S. meliloti* y 379 genes únicos; los genes presentes en pRmeGR4a y pRmeGR4b no se engloban dentro del genoma común de esta especie, y contienen 133 y 210 genes únicos respectivamente; en pRmeGR4c y pRmeGR4d encontramos 496 y 1.228 genes pertenecientes al genoma común de *S. meliloti*, y 523 y 229 genes únicos respectivamente.

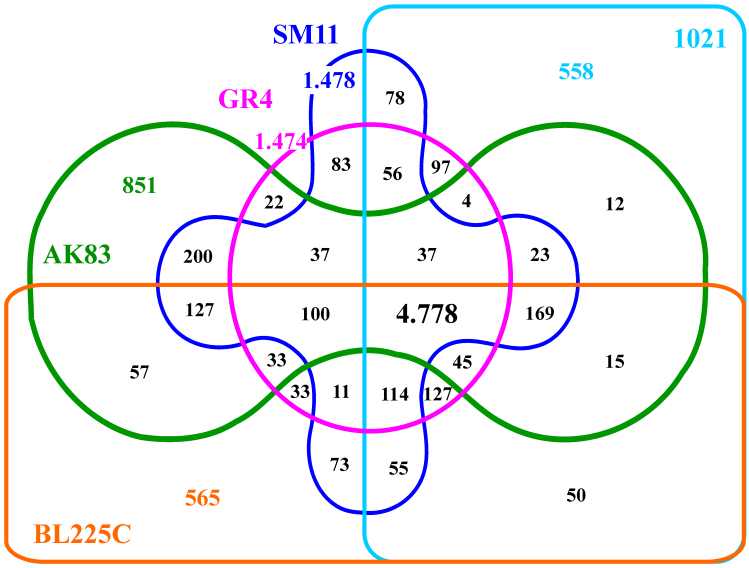


Figura R1.17: **Pangenoma de la especie *S. meliloti* a fecha 12-09-2012.** En el diagrama de Venn se representa el número de genes compartidos por las cepas completamente secuenciadas hasta esa fecha (1021, GR4, SM11, AK83 y BL225C). Se resalta el número de genes que forman parte del genoma común de esta especie y el número de genes únicos presentes en cada una de las cepas (representadas en diferente color).

R1.5 DISCUSIÓN

El avance en las nuevas tecnologías de secuenciación masiva ha puesto al alcance de la comunidad científica todo el desarrollo de una serie de plataformas que, bajo el común denominador del alto rendimiento en secuenciación, se diferencian en el modo de preparar la muestra, el proceso de secuenciación y el análisis de los datos (Metzker, 2010). La aplicación biológica de estas plataformas de secuenciación de segunda generación, como por ejemplo son la pirosecuenciación 454 (Margulies *et al*, 2005), ABI SOLiD (Shendure *et al*, 2005), e Illumina (Bentley *et al*, 2008), varía dependiendo del protocolo usado. Así, la tecnología 454 Titanium GS FLX de Roche es la más indicada para realizar la secuenciación y ensamblaje *de novo* de un genoma bacteriano (Metzker, 2010). Una de las principales ventajas de esta plataforma es que proporciona lecturas de mayor longitud (400 pb aproximadamente) que la mayoría de métodos de secuenciación (Lee & Tang, 2012), lo cual mejora el mapeo de regiones repetidas (Metzker, 2010) y produce ensamblajes significativamente más continuos en genomas procarióticos (Jiang *et al*, 2012). No obstante, ya existen plataformas de secuenciación de tercera generación, como por ejemplo Helicos (Harris *et al*, 2008), PacBio (Eid *et al*, 2009) o Ion Torrent (Rothberg *et al*, 2011), con las que se podrán realizar secuenciaciones de manera muy rápida y barata cuando, en unos años, alcancen su máximo potencial (Schadt *et al*, 2010).

En la secuenciación del aislado de *S. meliloti* GR4 llevada a cabo, el tamaño medio de todas las lecturas fue 304 pb (calculado como el total de bases secuenciadas entre el número total de lecturas; figura R1.1), incluyendo las lecturas obtenidas a partir de pirosecuenciación (Margulies *et al*, 2005) y *paired end* (Ng *et al*, 2006). Las lecturas pareadas, en el caso de GR4, provinieron de un *paired end* de 3 kb (Jarvie & Harkins, 2008), con una longitud media de los fragmentos generados de 2.685 pb (figura R1.1), que permitió construir 17 andamiajes. La técnica para la construcción de la genoteca *paired end* ha ido mejorando, y hoy en día existe la posibilidad de realizar un *paired end* de 8 kb que, respecto al de 3 kb, reduce significativamente el número de andamiajes generados (de 16 a 5 en un estudio de 100 genomas; Jiang *et al*,

2012). Además, una investigación sobre el tamaño de los huecos producidos tras el ensamblaje de 1.542 cromosomas determina que el 99'5 % de los huecos tiene una longitud menor de 8 kb, lo que sugiere que, para la mayoría de los genomas procarióticos, una carrera de *paired end* de 8 kb es suficiente para ordenar casi todos los *contigs* en un solo andamiaje (cuando el genoma se compone de un único cromosoma) o en varios, dependiendo del número de replicones (Li *et al*, 2012a).

En la secuenciación de GR4 se consiguió un alto grado de cobertura, un 77x, de los cuales un 17x se correspondió con las lecturas pareadas. Hay trabajos que demuestran que un grado de cobertura 10x (con lecturas de 400 pb), junto a una cobertura 10x de un *paired end* de 8 kb, es suficiente para producir ensamblajes altamente completos como con una cobertura 20x (Jiang *et al*, 2012; Li *et al*, 2012a). Aunque, a mayor grado de cobertura, menor tasa de error por nucleótido y mayor calidad en el ensamblaje (Li *et al*, 2012a). Además, gracias al elevado grado de cobertura conseguido en la secuenciación de GR4 se obtuvieron suficientes lecturas para construir los *contigs* y andamiajes correspondientes a los plásmidos crípticos, menos representados en la célula debido a su pérdida con el paso de las generaciones (Mercado-Blanco & Olivares, 1993).

Una vez construidos los andamiajes, el siguiente paso es la finalización del genoma (conocido como *finishing*), es decir, la obtención de su secuencia completa. A pesar de que éste es un proceso costoso en tiempo y dinero, lo deseable es que la secuencia de un genoma se encuentre lo más completa y con la mayor calidad posible. Los genomas completos presentan ventajas frente a los borradores de genoma, como por ejemplo, que con ellos se pueden realizar análisis de genómica comparada, o que pueden servir como genomas de referencia (Mardis *et al*, 2002). Aunque la secuenciación de nuevos genomas dentro de una misma especie está revelando que quizás el concepto de genoma de referencia no sea adecuado, puesto que el cromosoma de *S. meliloti* SM11 presenta una inversión de 1'68 Mb respecto al de 1021 flanqueada por dos copias de IS*Rm22* (Schneiker-Bekel *et al*, 2011). Por este motivo, se llevó a cabo un ensamblaje *de novo* del genoma de GR4, y su

cierre se abordó mediante estrategias en las que no se utilizó un genoma de referencia.

Los andamiajes se generan por la unión de los *contigs* y la adición de huecos entre ellos. Los elementos repetidos son la causa por la cual, tras el ensamblaje, no se produce una secuencia continua del replicón. El denominado índice del grado de cobertura (FCI) fue la base de la estrategia seguida para cerrar el genoma de GR4. Este parámetro permitió diferenciar los *contigs* de elementos repetidos de los *contigs* de secuencia única, y, además, estimar el número de copias que presenta cada uno de ellos dentro del genoma de GR4. Entre los *contigs* de secuencia única pudimos identificar los *contigs* correspondientes a los plásmidos crípticos por su bajo valor del FCI (tabla R1.1). Este resultado pone de manifiesto que, en la muestra de DNA total de GR4 que se utilizó como molde para la secuenciación, los plásmidos crípticos pRmeGR4a y pRmeGR4b se encontraban en una proporción molar respecto al cromosoma del 24 % y 36 % respectivamente. Esto apoya la hipótesis de que los plásmidos crípticos se pierden a través de las generaciones (Mercado-Blanco & Olivares, 1993). El cromosoma inicia la replicación una vez por ciclo celular, sin embargo, la replicación de los plásmidos en cada ciclo celular está sujeta al número de copias que contenga en ese momento la célula bacteriana (Venkova-Canova & Chattoraj, 2011). Este proceso de ajuste del número de copias de los plásmidos puede explicar el hecho de que el valor del FCI del cromosoma (FCI de 1'10) fuera superior a la media del FCI de los andamiajes que componen pRmeGR4c y pRmeGR4d (FCI de 1'00 y 0'95 respectivamente; tabla R1.1). Hay que señalar que el valor del FCI del *contig* 110 (1'33) no aumentó demasiado respecto al resto de *contigs* pertenecientes al cromosoma (FCI entre 1'02-1'20) por contener las lecturas de las tres copias del operón ribosomal. Una explicación puede ser su tamaño, ya que tiene una longitud de 87.006 pb, pero la zona repetida ocupa sólo 6.268 pb (figura R1.2).

La información que aportaba el cálculo del FCI (figura R1.3A) se vio complementada por la estrategia de las lecturas pareadas diseñada con objeto de completar el genoma de GR4 (figura R1.3B). En esta estrategia se usaron

las lecturas obtenidas a partir de la genoteca *paired end* de 3 kb para determinar el *contig* de elemento repetido que debía incluirse entre dos *contigs* de secuencia única. Así, pudimos cerrar el 83 % de los huecos generados en el borrador del genoma de GR4 y, además, resolver los SNPs presentes en algunos de los *contigs* de elementos repetidos (tabla R1.2), determinando la secuencia concreta del elemento repetido en cada hueco específico (figura R1.5). Existen varios programas que, para completar los huecos, utilizan genomas de referencia y siguen una estrategia basada en PCRs, como son Projector2 (van Hijum *et al*, 2005), OSlay (Richter *et al*, 2007), ABACAS (Assefa *et al*, 2009), CONTIGuator (Galardini *et al*, 2011a). Sin embargo, la estrategia planteada en este trabajo permitió resolver el genoma de GR4 mediante los datos de la propia secuenciación masiva y sin la necesidad de realizar un alto número de PCRs. Únicamente se requirió el diseño de oligonucleótidos para ordenar los andamiajes que componían el plásmido pRmeGR4c, mientras que en otros casos se llevaron a cabo PCRs sólo para confirmar la secuencia propuesta. Recientemente se ha desarrollado el programa GapFiller, que es capaz de completar automáticamente los huecos entre *contigs* de un mismo andamiaje utilizando las lecturas pareadas (Boetzer & Pirovano, 2012).

La figura R1.18 muestra todos los huecos que tuvieron que ser cerrados tras el ensamblaje *de novo* del genoma de la cepa *S. meliloti* GR4 para obtener su secuencia completa. En ella se observa que el hueco a partir del cual se construyó el plásmido pRmeGR4b es de mayor longitud que los huecos que dieron lugar al resto de replicones. En concreto, para cerrar este plásmido críptico tuvimos que introducir una secuencia de 32'6 kb, correspondiente a un puzzle de ISs más un *contig* de secuencia única que el ensamblador había incluido en otro andamiaje (figura R1.8B). Dentro de dicho plásmido esperábamos encontrar un elevado porcentaje de elementos repetidos (Molina-Sánchez, 2008); algunos previamente descritos son: IS*Rm3* (Soto *et al*, 1992a), IS*Rm4* (Soto *et al*, 1992b), IS*Rm2011-2* (Selbitschka *et al*, 1995), IS*Rm6* (Zekrí & Toro, 1996), IS*Rm4-1* e IS*Rm9* (Zekri *et al*, 1998), RmInt1 (intrón del grupo II; Martínez-Abarca *et al*, 1998), IS*Rm8* (Zekri & Toro, 1998), e IS*Rm7* (Selbitschka *et al*, 1999), todos ellos presentes en el plás-

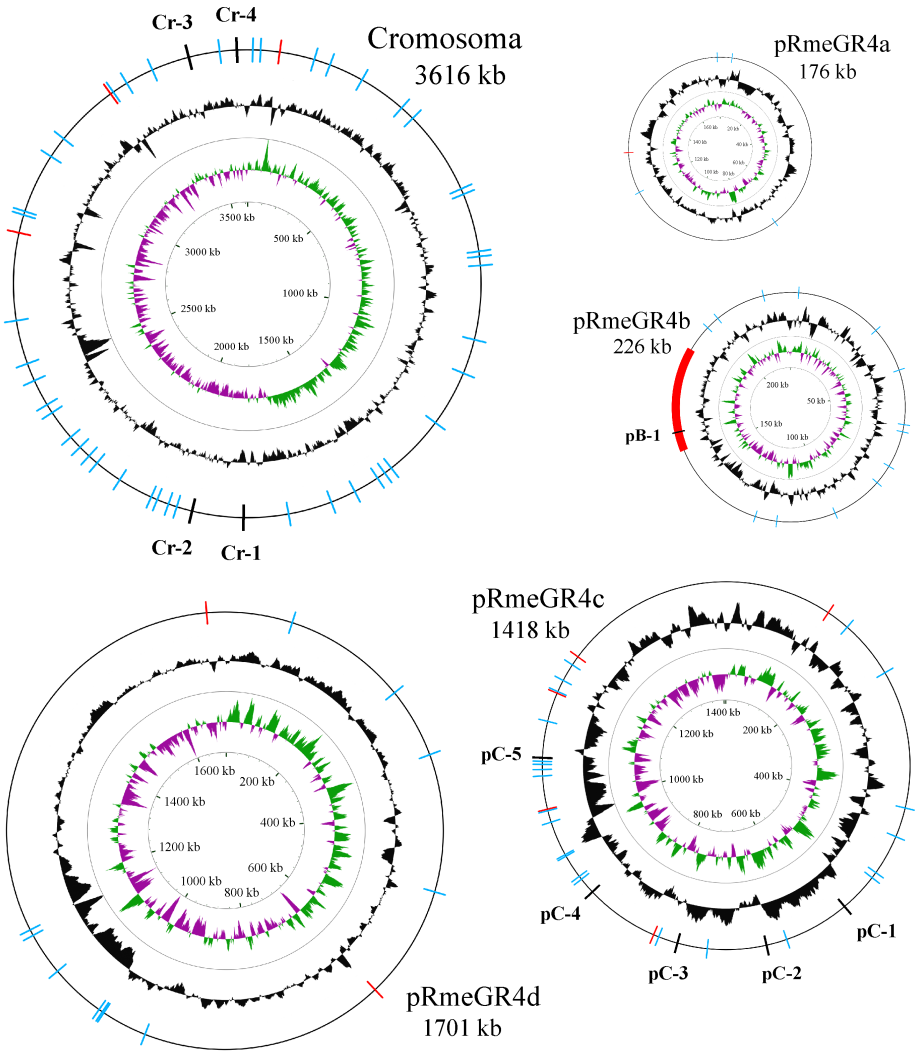


Figura R1.18: **Estructura del genoma de la cepa *S. meliloti* GR4, distribuida en cinco replicones.** Los plásmidos crípticos están representados en la misma escala relativa; los plásmidos simbióticos y el cromosoma están representados a un cuarto y un octavo de esa escala respectivamente. En el círculo exterior se muestran los huecos que han tenido que ser resueltos: en rojo, los huecos que han permitido la reconstrucción de los replicones, y en azul, los huecos entre *contigs* que componen un andamiaje. En este círculo, las líneas negras se corresponden con las copias del intrón del grupo II RmInt1 presentes en los huecos entre *contigs*, excepto la copia pB-1, que se encuentra en la unión del andamiaje que constituye el plásmido pRmeGR4b. Los círculos interiores indican el contenido G+C relativo (negro) y la desviación GC (verde/morado).

mido pRmNT140 (Toro, 1985). El tamaño de los fragmentos que han tenido que ser incluidos en todos los huecos entre andamiajes, salvo para cerrar pRmeGR4b, fue inferior a 7 kb, por lo que un *paired end* de 8 kb (en vez del *paired end* de 3 kb realizado) habría generado un andamiaje por cada replicón (Li *et al*, 2012a), evitando así el problema de la estructuración de los andamiajes dentro de los replicones. Incluso el *contig* 159 (inicialmente ensamblado en el andamiaje 16) hubiera sido ensamblado en el andamiaje 6 junto al *contig* 21, ya que la longitud real que separa estos dos *contigs* es 7.140 pb (inferior a las 8 kb del *paired end*).

La hibridación de DNA genómico de GR4 con sondas específicas para RmInt2 e IS*Rm17*, junto a la secuenciación Sanger llevada a cabo con varios productos de PCR, confirmaron la secuencia propuesta para las uniones entre andamiajes que contenían los *contigs* 10, 37 y 38. En la figura R1.12 se observa que la banda correspondiente a la unión entre la pareja de cebadores c3/c5, en pRmeGR4c, no aparece al hibridar con la sonda IS*Rm17*. Esto se debe a que, en ese hueco entre andamiajes, RmInt2 no se encuentra asociado a IS*Rm17* (estudio que ampliaremos en el capítulo 3 de esta Tesis Doctoral). La sonda IS*Rm17* hibrida entre las posiciones 754 y 1.012 de la IS (figura R1.12A), lo cual explica la doble intensidad mostrada por la banda correspondiente a la unión c6/c8, en pRmeGR4c, puesto que dan señal tanto la copia entera de esta IS como el *contig* 38 adicional al hibridar con dicha sonda respecto a la hibridación con la sonda RmInt2-3'. Este ensayo de hibridación genómica corroboró la hipótesis de que la unión d2/d4, en pRmeGR4d, sólo presenta una copia de IS*Rm17* y que, entre los dos *contigs* 10, no se encuentra ninguna secuencia adicional a la propuesta que hubiera incrementado el tamaño de la banda aparecida respecto al tamaño de la banda esperada (figura R1.11C). Esto se aprecia mejor en el carril correspondiente a la enzima *Sal*I debido a que genera tamaños de banda más pequeños que las otras enzimas. La presencia de dos *contigs* 10 determina que la banda correspondiente a esa unión d2/d4, al hibridar con la sonda RmInt2-3', aparezca con el doble de intensidad que las bandas de tamaños similares. Esas dos copias de intrón deben estar completas, ya que, además de que el tamaño de la banda del producto amplificado así lo indica, el análi-

sis de las lecturas ensambladas sobre el *contig* 10 con el visualizador Tablet no muestra lecturas en mitad del *contig* que pertenezcan a otra secuencia (como ha ocurrido con otros *contigs* repetidos interrumpidos).

La figura R1.18 muestra el análisis de la desviación GC, calculada como [(G – C)/(G + C)], llevado a cabo con cada uno de los replicones. Este parámetro estima el origen de replicación y su término en el cromosoma de una bacteria, sin embargo, no es capaz de detectar estas regiones en los plásmidos (Lobry, 1996; Grigoriev, 1998). El inicio del cromosoma de GR4 se determinó tomando como referencia la secuencia de inicio del cromosoma de la cepa 1021, sobre el que se han realizado análisis de la desviación GC (Capela *et al*, 2001) y se ha identificado experimentalmente un origen de replicación (Sibley *et al*, 2006). El resultado del análisis de la desviación GC llevado a cabo con el cromosoma de GR4 apoya la zona propuesta como inicio de este replicón, aunque posteriormente se corrigiera adelantando 72 nt para evitar problemas en la anotación del primer gen (ver Material y Métodos, apartado M.12.4.3). La zona de inicio de los cuatro plásmidos de GR4 se determinó en base a la secuencia inicial de los plásmidos correspondientes presentes en la cepa SM11, los cuales contienen el gen de inicio de la replicación *repA* (Schneiker-Bekel *et al*, 2011). La posterior anotación de los plásmidos de GR4 puso de manifiesto que el operón *repABC* se encuentra en el comienzo de todos ellos, al igual que ocurre en la mayoría de plásmidos dentro de la clase α-proteobacteria (Young *et al*, 2006).

R1.5.1 Características del genoma de *S. meliloti* GR4 y estudios de genómica comparada en la especie *S. meliloti*

Las características del genoma de GR4 tras su anotación automática (ver Material y Métodos, apartado M.12.4.2) y curación manual (ver Material y Métodos, apartado M.12.4.3) se detallan en la tabla R1.3. Las cepas pertenecientes a la especie *S. meliloti* que contienen plásmidos crípticos, además de los megaplásmidos simbióticos característicos de esta especie, presentan un genoma de tamaño superior a 7 Mb. SM11 contiene el genoma de ma-

yor tamaño de las cepas de *S. meliloti* secuenciadas hasta la fecha (7.499.157 pb, con dos plásmidos crípticos; Schneiker-Bekel *et al*, 2011). El tamaño del genoma de Rm41 (7.149.690 bp, con un plásmido críptico similar a pRmeGR4b; Weidner *et al*, 2013) y AK83 (7'14 Mb, con dos plásmidos crípticos; Galardini *et al*, 2011b) es similar al de la cepa GR4 (7.139.558 pb). 1021 y BL225C, sin plásmidos crípticos, presentan los genomas de menor tamaño, con 6.691.694 pb y 6'98 Mb respectivamente (Galibert *et al*, 2001; Galardini *et al*, 2011b). Respecto al contenido G+C, hay que señalar que, en todas estas cepas, tanto el cromosoma como el plásmido simbiótico tipo pSymB presentan un porcentaje G+C superior al 62 %, mientras que dicho porcentaje en el plásmido simbiótico tipo pSymA no alcanza el 61 %. Los plásmidos crípticos encontrados en alguna de ellas presentan un porcentaje G+C inferior al 60 %, lo cual sugiere que han ocurrido eventos de transferencia horizontal desde otras bacterias con un porcentaje G+C menor que éstas, como por ejemplo desde los plásmidos pRL7 o pRL8 de *R. leguminosarum* (Young *et al*, 2006). Una relación similar se observa en la región codificante: más del 85'8 % de la secuencia del cromosoma y del plásmido tipo pSymB es codificante (siendo superior en el caso del plásmido simbiótico), y en el plásmido tipo pSymA esta región no supera el 84 %. Hay que destacar que el plásmido críptico pRmeGR4b es el que presenta la proporción más baja de región codificante en todos los replicones analizados (78'9 %), quizás por la presencia de genes fragmentados debido al alto número de secuencias de inserción que contiene. El genoma de GR4 presenta 3 operones ribosomales, todos ellos en el cromosoma (figura R1.13), al igual que todas las cepas del género *S. meliloti* secuenciadas completamente hasta el momento. No obstante, el número y distribución de tRNAs en estos rizobios varía entre cepas, desde 47 en CCNWSX0020 hasta 55 en GR4 y BL225C.

La proporción de genes asignados a las diferentes categorías funcionales de los COGs en todo el genoma de GR4 (70'3 %) se vio disminuida respecto al alto porcentaje encontrado en el cromosoma (84'3 %) y plásmidos simbióticos (72'0 % en pRmeGR4c y 81'7 % en pRmeGR4d) debido a la presencia de los plásmidos crípticos (tabla R1.3). Esta disminución en el porcentaje

de genes asignados a COGs también se observa en otras cepas con plásmidos crípticos (AK83: 71'1 %) frente a cepas donde no están presentes este tipo de plásmidos (1021: 76'3 %, y BL225C: 75'4 %; Galardini *et al*, 2011b), lo cual sugiere que en los plásmidos crípticos se está acumulando un mayor número de elementos repetidos (Gil & Latorre, 2012). Además, esto también explicaría la baja proporción de región codificante presente en pRmeGR4b (78'9 %).

A pesar de que el porcentaje de genes relacionados con familias de proteínas presentes en la base de datos Pfam es mayor en todos los replicones de GR4 que el porcentaje de genes asociados a COGs (tabla R1.3), analizamos en detalle estos últimos por contener menor proporción de genes con función desconocida (9'1 % frente a 14'6 %; tabla R1.4). Dentro del genoma de GR4, la categoría funcional con el mayor porcentaje de genes asignados es replicación, recombinación y reparación (11'6 %; tabla R1.4), categoría relacionada con elementos genéticos móviles (MGEs). Generalmente, la densidad de estos MGEs en cromosomas bacterianos es inferior al 3 %, mientras que en los plásmidos está por encima de ese número (Siguier *et al*, 2006a). Aunque hay casos excepcionales como el de *Wolbachia*, con entre un 5 % y un 20 % dependiendo de la cepa (Leclercq *et al*, 2011). En pRmeGR4b, la mayoría de genes asociados a COGs se corresponden con MGEs (35'1 %; figura R1.14 y tabla R1.S3), un porcentaje superior al encontrado en pSmeSM11b (25'9 %). En pRmeGR4a encontramos un número menor de MGEs (9'7 %) que en los otros plásmidos crípticos, incluido su equivalente pSmeSM11a (14'0 %; Stiens *et al*, 2007). Respecto al resto de replicones, pRmeGR4c es el que contiene mayor proporción de MGEs (5'5 %), seguido del cromosoma (4'8 %) y pRmeGR4d (3'0 %). La relación entre replicones y número de MGEs en la cepa 1021 es igual que en GR4, aunque los porcentajes en GR4 son superiores a los encontrados en 1021 (pSymA: 3'6 %, cromosoma: 2'2 %, y pSymB: 0'9 %; Stiens *et al*, 2007). Una posible explicación es el mayor tamaño del genoma de GR4 frente al de 1021, puesto que, en procariotas, el tamaño del genoma está correlacionado positivamente con el número de ISs (Touchon & Rocha, 2007). En relación con la simbiosis

que crea la cepa GR4 con plantas de alfalfa, hay que destacar que este rizobio presenta un elevado porcentaje de genes involucrados en metabolismo y transporte de carbohidratos (7'96 %; tabla R1.4; Pini *et al*, 2011), sobre todo el plásmido simbiótico pRmeGR4d (17'85 %; tabla R1.3), cuyo porcentaje incluso supera al del plásmido pSymB de 1021 (12 %; Finan *et al*, 2001).

La evolución de la estructura genómica y la heterogeneidad que presenta entre bacterias se debe, en parte, a la transferencia horizontal de material genético. Estudios de genómica comparada ayudan a entender la historia evolutiva que han seguido los genomas. Por ello, realizamos una comparación genómica estructural con las cepas de la especie *S. meliloti* completamente secuenciadas disponibles en las bases de datos a principios del año 2012. Llevamos a cabo un alineamiento múltiple de estos genomas con el programa progressiveMauve, que permite alinear secuencias muy emparentadas que hayan sufrido reorganizaciones genómicas (Darling *et al*, 2010). El resultado muestra un alto grado de conservación en la estructura del cromosoma de las cepas de esta especie (figura R1.16A), al igual que en los plásmidos tipo pSymB (figura R1.16C). Sin embargo, los plásmidos tipo pSymA han sufrido un mayor número de reorganizaciones genómicas que ha causado una estructura en mosaico (figura R1.16A; Guo *et al*, 2007). El estudio de la estructura de los plásmidos crípticos presentes en las cepas GR4, SM11 y AK83 pone de manifiesto la existencia de pocas regiones homólogas entre ellos. No obstante, un análisis detallado de estos plásmidos podría revelar zonas de similitud con los replicones (cromosoma y plásmidos simbióticos) de otras cepas (Stiens *et al*, 2006, 2007; Kuhn *et al*, 2008; Galardini *et al*, 2011b).

La comparación llevada a cabo en este trabajo refleja, además, los problemas derivados de la anotación. La definición del origen y la orientación de cada replicón tras la secuenciación de un genoma es crítica para realizar estudios de genómica comparada. Hemos comentado anteriormente que la secuencia de los plásmidos de GR4 comienza con el operón *repABC* en sentido, como se ha descrito para la mayoría de plásmidos dentro de la clase α-proteobacteria (Young *et al*, 2006). Los plásmidos pSymA y pSymB,

pertenecientes a 1021, presentan la misma distribución de bloques homólogos que los plásmidos pRmeGR4c y pSmeSM11c, y pRmeGR4d y pSmeSM11d respectivamente, pero se encuentran anotados en orientación inversa. El caso del plásmido tipo pSymA en las cepas AK83 y BL225C es más complejo, ya que la definición de la secuencia de comienzo y la orientación ha sido diferente respecto al resto de plásmidos. Los plásmidos tipo pSymB de estas dos últimas cepas se han anotado en la misma orientación que pRmeGR4d y pSmeSM11d, sin embargo, la definición de la secuencia de inicio difiere entre ellos y con pRmeGR4d y pSmeSM11d. En cuanto al cromosoma, en 1021 y GR4 éste comienza con un bloque de secuencia homóloga que en SM11, AK83 y BL225C se encuentra al final. Esta disposición diferencial sugiere que la definición de la secuencia de inicio del cromosoma de esas últimas cepas es distinta a la de 1021 y GR4. Un análisis de la desviación GC (Lobry, 1996) realizado con el cromosoma de todas ellas confirmó que 1021 y GR4 presentan un cambio en la desviación GC al inicio y hacia la mitad del replicón, mientras que estos puntos en SM11, AK83 y BL225C están desplazados (figura R1.19).

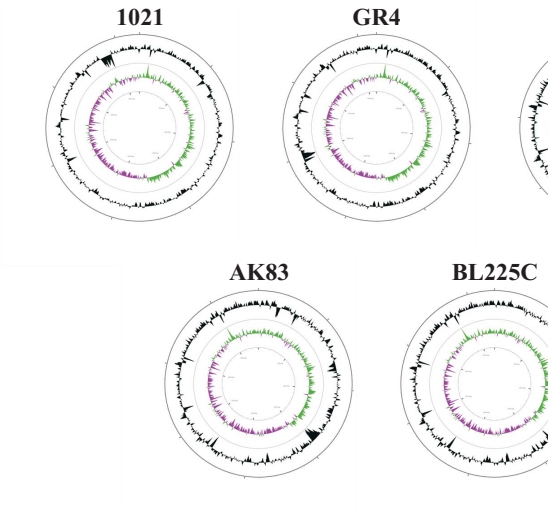


Figura R1.19: **Análisis de la desviación GC en el cromosoma de las cepas de *S. meliloti* 1021, GR4, SM11, AK83 y BL225C.** El círculo externo muestra el contenido G+C relativo (en negro) y el círculo interno indica la desviación GC (verde/morado).

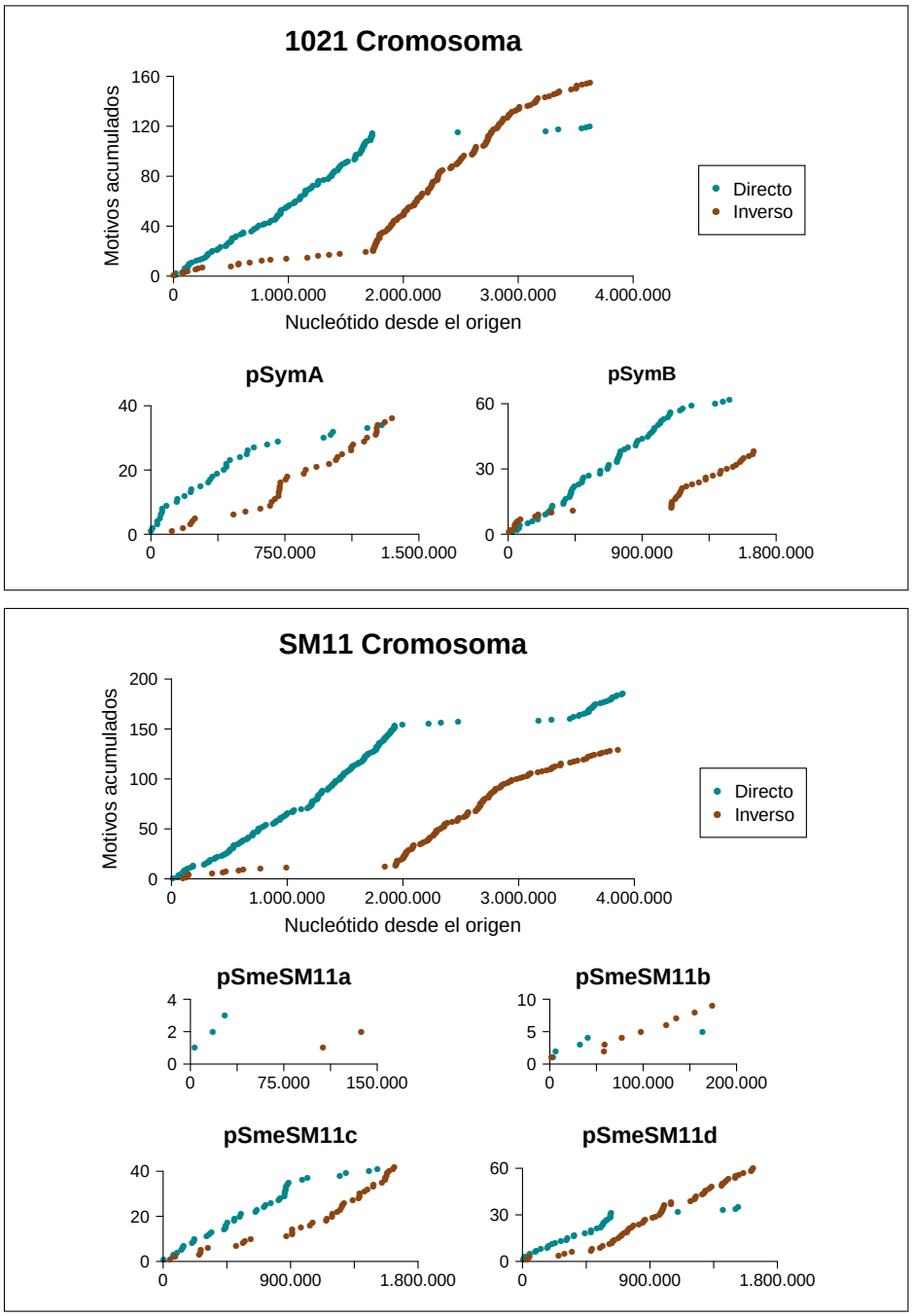
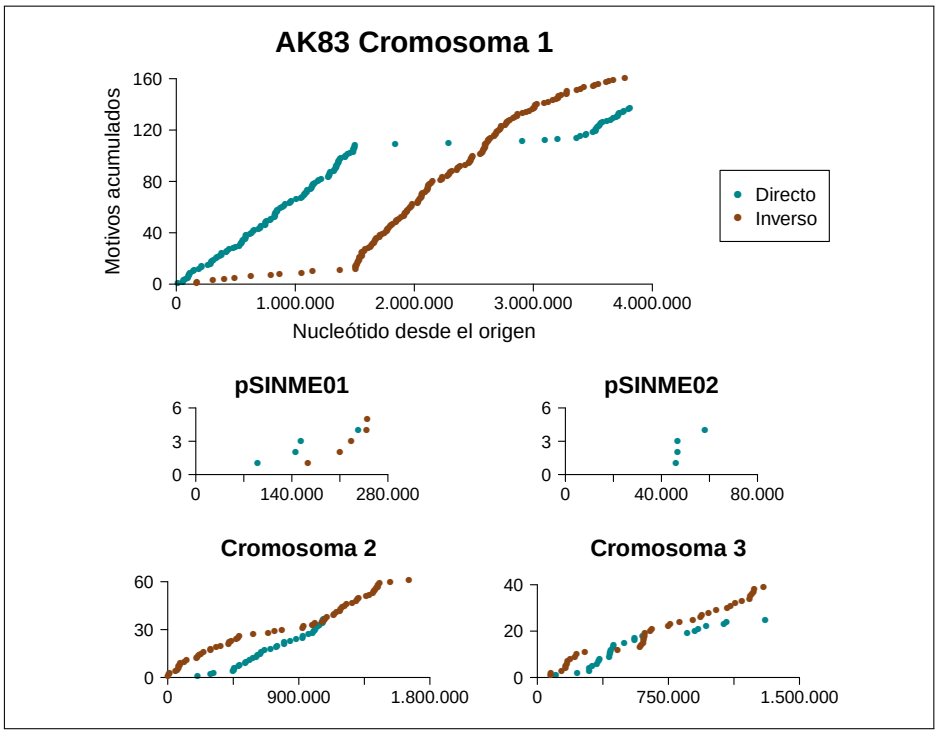
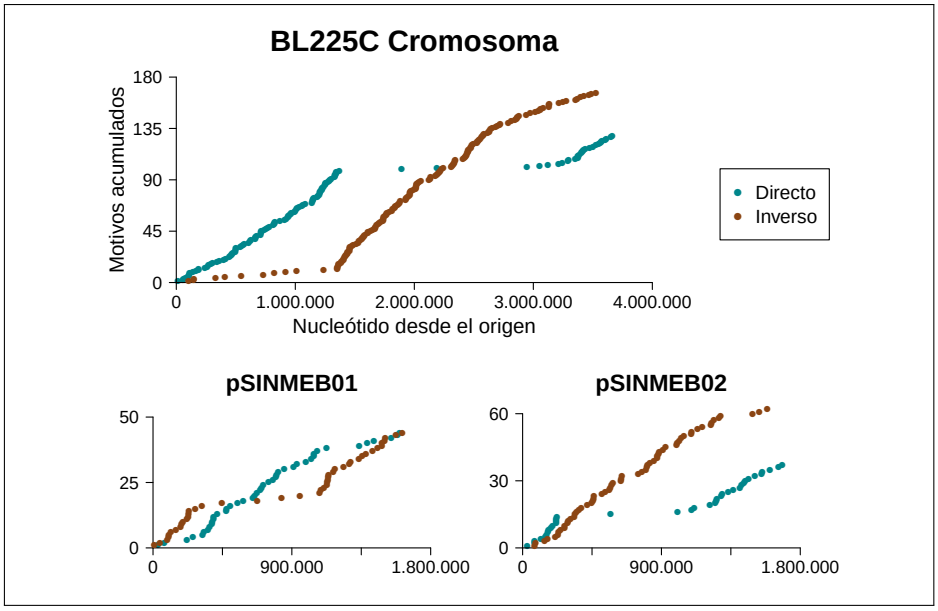


Figura R1.20: **Distribución del motivo octomérico GGGCAGGG en el genoma de varios rizobios.** Se muestra la acumulación del número ►





◄ de motivos KOPS en orientación directa e inversa en todos los replicones de las cepas 1021, SM11, AK83 y BL225C.

Empíricamente, sólo se ha demostrado el origen de replicación para unas pocas bacterias y arqueas. Para identificar el origen y término de replicación *in silico* se han descrito herramientas computacionales basadas en varios tipos de desviaciones y otras evidencias (Sernova & Gelfand, 2008). Con el fin de corroborar el origen de replicación propuesto para cada replicón de GR4 llevamos a cabo un análisis de la distribución del motivo octamérico KOPS (5'-GGGCAGGG-3'; Bigot *et al*, 2007). El genoma de GR4 mostró el patrón característico de este motivo (una acumulación de las copias del motivo en las regiones cercanas al origen y término de la replicación; figura R1.15), al igual que presentan otras bacterias como *R. leguminosarum* (Young et al., 2006), *Azorhizobium caulinodans* (Lee *et al*, 2008) o *E. coli* (Bigot *et al*, 2005). En la figura R1.20 se muestra el análisis de la distribución del motivo KOPS realizado en el genoma de las cepas 1021, SM11, AK83 y BL225C. 1021 presenta el patrón característico de ese motivo, al igual que GR4, mientras que en SM11, AK83 y BL225C se aprecia una distribución anómala hacia el final de la secuencia del cromosoma, coincidiendo con los resultados obtenidos tras la comparación genómica (figura R1.16) y tras el análisis de la desviación GC (figura R1.19). Por tanto, para definir el origen de replicación del cromosoma y plásmidos de un genoma se deben realizar diferentes análisis mediante herramientas computacionales que predicen la posible posición del *oriC* y *terC* (Sernova & Gelfand, 2008).

Una especie bacteriana puede describirse por su pangenoma, que incluye un genoma común (conjunto de genes presente en todos los aislados), un genoma accesorio (que contiene genes presentes en dos o más cepas) y genes únicos (específicos de una sola cepa; Medini *et al*, 2005). Tras la adición de los genes presentes en GR4 al pangenoma de *S. meliloti* descrito recientemente (Galardini *et al*, 2011b), éste se ha incrementado en 2.477 genes. Por consiguiente, se puede decir que el pangenoma de *S. meliloti* está abierto, al igual que el de otras especies bacterianas con varios representantes secuenciados, como es el caso de *Streptococcus agalactiae* (Medini *et al*, 2005). El análisis de 1021, AK83 y BL225C muestra un genoma común compuesto por 5.124 genes, que se reduce a 5.075 genes al añadir el genoma de SM11

(Galardini *et al*, 2011b), y a 4.778 genes al añadir el genoma de GR4. Los conceptos pangenoma y genoma común han sido ampliamente aceptados para evaluar la distribución de genes en especies bacterianas, sin embargo, sería necesario secuenciar la totalidad de cepas pertenecientes a una especie para conocer el valor real de esos parámetros. Por este motivo, Kislyuk *et al* (2011) han desarrollado un nuevo parámetro (*genomic fluidity*) que estima la diversidad génica entre genomas de una especie o de un grupo de organismos estrechamente relacionados. Futuros estudios con ésta y otras herramientas completarán los estudios de pangenoma y genoma común llevados a cabo en la especie *S. meliloti*.

Tabla R1.S1: FCI de todos los contigs que forman los andamiajes del genoma de GR4

Andamiaje	**Contig**	**Longitud**	**Lecturas**	**FCI**
	Contig 10	1906	4009	8.64
	Contig 11	13819	3692	1.10
	Contig 12	42999	10366	0.99
Andamiaje 4	Contig 13	1685	273	0.67
(FCI=1.07)	Contig 14	26951	6728	1.03
	Contig 15	2619	662	1.04
	Contig 16	18421	4518	1.01
	Contig 17	92591	21898	0.97
	Contig 21	31484	2822	0.37
	Contig 22	6003	698	0.48
	Contig 23	17243	1450	0.35
	Contig 24	15860	1321	0.34
	Contig 25	3896	401	0.42
Andamiaje 6	Contig 26	11219	947	0.35
(FCI=0.36)	Contig 27	1405	147	0.43
	Contig 28	15891	1263	0.33
	Contig 29	13422	1222	0.37
	Contig 30	26423	2192	0.34
	Contig 31	10553	909	0.35
	Contig 32	13404	1076	0.33

Continúa en la página siguiente

Tabla R1.S1 – Continuación de la página anterior

Andamiaje	**Contig**	**Longitud**	**Lecturas**	**FCI**
Andamiaje 6	Contig 33	3766	346	0.38
(FCI=0.36)	Contig 34	8650	743	0.35
Andamiaje 7 (FCI=1.04)	Contig 39	40708	10476	1.06
	Contig 40	6544	1804	1.13
	Contig 41	4414	1197	1.11
	Contig 42	3265	822	1.03
	Contig 43	1509	439	1.19
	Contig 44	44212	10935	1.02
	Contig 45	35702	8797	1.01
	Contig 46	2132	601	1.16
Andamiaje 8 (FCI=1.08)	Contig 47	7726	2237	1.19
	Contig 48	39732	10584	1.09
	Contig 49	69594	18401	1.09
	Contig 50	91580	23484	1.05
	Contig 51	76655	19736	1.06
	Contig 52	39210	10194	1.07
	Contig 53	39536	11213	1.16
	Contig 54	17845	4753	1.09
	Contig 55	47339	12121	1.05
Andamiaje 9 (FCI=0.24)	Contig 60	43736	2380	0.22
	Contig 61	3420	223	0.27
	Contig 62	63683	3694	0.24
	Contig 63	47227	2776	0.24
	Contig 64	11350	653	0.24

Continúa en la página siguiente

Tabla R1.S1 – Continuación de la página anterior

Andamiaje	**Contig**	**Longitud**	**Lecturas**	**FCI**
	Contig 75	215977	58676	1.12
	Contig 76	106815	29834	1.15
	Contig 77	42604	14033	1.35
	Contig 78	66987	18219	1.12
	Contig 79	21805	5666	1.07
	Contig 80	74581	19869	1.09
	Contig 81	32266	8199	1.04
	Contig 82	29896	8199	1.13
	Contig 83	26005	6888	1.09
	Contig 84	69262	18826	1.12
	Contig 85	87088	22969	1.08
Andamiaje 10	Contig 86	16229	4522	1.14
(FCI=1.11)	Contig 87	29504	7885	1.10
	Contig 88	21375	5780	1.11
	Contig 89	36102	9652	1.10
	Contig 90	125116	33135	1.09
	Contig 91	110012	29788	1.11
	Contig 92	107628	29426	1.12
	Contig 93	62160	16803	1.11
	Contig 94	64974	17633	1.11
	Contig 95	20512	6010	1.20
	Contig 96	19148	5540	1.19
	Contig 97	155049	41629	1.10
	Contig 98	138865	37243	1.10

Continúa en la página siguiente

Tabla R1.S1 – Continuación de la página anterior

Andamiaje	**Contig**	**Longitud**	**Lecturas**	**FCI**
	Contig 99	85410	23088	1.11
	Contig 100	182959	48381	1.09
	Contig 101	20410	5201	1.05
	Contig 102	11446	3101	1.11
	Contig 103	144659	39232	1.11
Andamiaje 10	Contig 104	14490	3875	1.10
(FCI=1.11)	Contig 105	210460	55407	1.08
	Contig 106	37258	9232	1.02
	Contig 107	119618	31133	1.07
	Contig 108	85897	21719	1.04
	Contig 109	39506	10265	1.07
	Contig 110	87006	28159	1.33
	Contig 113	138620	32187	0.95
	Contig 114	172945	40294	0.96
Andamiaje 11	Contig 115	86039	18922	0.90
(FCI=0.93)	Contig 116	155793	33982	0.90
	Contig 117	102068	22329	0.90
Andamiaje 12	Contig 120	344444	80512	0.96
	Contig 135	305775	68847	0.92
	Contig 136	57276	12344	0.88
Andamiaje 14	Contig 137	1683	398	0.97
(FCI=0.96)	Contig 138	2234	503	0.92
	Contig 139	67754	18844	1.14
	Contig 140	50996	12317	0.99

Continúa en la página siguiente

Tabla R1.S1 – Continuación de la página anterior

Andamiaje	**Contig**	**Longitud**	**Lecturas**	**FCI**
Andamiaje 14 (FCI=0.96)	Contig 141	7090	2381	1.38
	Contig 142	529788	123587	0.96
Andamiaje 15 (FCI=0.98)	Contig 148	7706	1920	1.02
	Contig 149	20847	4978	0.98
	Contig 150	10653	2477	0.95
Andamiaje 16 (FCI=0.96)	Contig 159	14618	1216	0.34
	Contig 160	1567	598	1.57
	Contig 161	21236	5186	1.00
	Contig 162	36909	8223	0.91
	Contig 163	70105	16361	0.96
	Contig 164	24885	5599	0.92
	Contig 165	79202	17600	0.91
	Contig 166	46303	10821	0.96
	Contig 167	6408	1731	1.11
	Contig 168	47785	12037	1.03
	Contig 169	35665	8624	0.99
	Contig 170	169357	40353	0.98
	Contig 171	72925	17358	0.98
	Contig 172	25580	6144	0.99
Andamiaje 17 (FCI=1.10)	Contig 184	153907	40747	1.09
	Contig 185	52588	14383	1.12
	Contig 186	148582	39360	1.09
	Contig 187	8415	2223	1.08
	Contig 188	39081	10816	1.14

<table>
<tr><td rowspan="6">Read</td><td>TotalNumberOfReads</td><td>the number of reads used in the assembly computation</td></tr>
<tr><td>TotalNumberOfBases</td><td>the number of read's bases used in the assembly computation</td></tr>
<tr><td>Assembled</td><td>the number of reads fully assembled into the contigs</td></tr>
<tr><td>Partial</td><td>the number of reads partially assembled into the contigs</td></tr>
<tr><td>Singleton</td><td>the number of reads that did not overlap with other reads</td></tr>
<tr><td>Repeat</td><td>the number of reads deemed to be from repeat regions</td></tr>
<tr><td rowspan="6">Large Contig</td><td>Contigs</td><td>the number of large contigs identified</td></tr>
<tr><td>Bases</td><td>the total number of bases in the large contigs</td></tr>
<tr><td>AvgContigSize</td><td>the average contig size</td></tr>
<tr><td>N50ContigSize</td><td>the N50 contig size

- An N50 contig size means that half of all bases reside in contigs of this size or longer</td></tr>
<tr><td>LargestContigSize</td><td>the size of the largest contig</td></tr>
<tr><td>Q40PlusBases</td><td>the number of bases called that have a quality score of 40 or above

- Quality score (Phred-equivalent) = -10Log10(A) where A is Probability of Error
<table>
<tr><th>Quality of Phred Score</th><th>Probability of incorrect base call</th><th>Base call accuracy</th></tr>
<tr><td>10</td><td>1 in 10</td><td>90%</td></tr>
<tr><td>20</td><td>1 in 100</td><td>99%</td></tr>
<tr><td>30</td><td>1 in 1000</td><td>99.9%</td></tr>
<tr><td>40</td><td>1 in 10000</td><td>99.99%</td></tr>
<tr><td>50</td><td>1 in 100000</td><td>99.999%</td></tr>
</table></td></tr>
<tr><td>%Q40</td><td>the percentage of bases called that have a quality score of 40 or above</td></tr>
<tr><td rowspan="2">All Contig</td><td>Contigs</td><td>the number of contigs identified</td></tr>
<tr><td>Bases</td><td>the total number of bases in the contigs</td></tr>
</table>

Figura R1.S1: **Descripción de la información sobre pirosecuenciación y ensamblaje de GR4 llevados a cabo.** Se muestra el significado de los términos de algunos datos proporcionados tras la secuenciación (recogidos en la figura R1.1).

Tabla R1.S2: FCI de todos los contigs repetidos que forman el genoma de GR4. Los contigs superiores a 2 kb fueron tratados por el ensamblador como un andamiaje (indicado entre paréntesis)

Contig	**Longitud**	**Lecturas**	**FCI**
Contig 1 (A1)	2528	1071	1.74
Contig 2 (A2)	2696	2319	3.53
Contig 3	499	273	2.25
Contig 4	467	305	2.68
Contig 5	784	105	0.55
Contig 6 (A3)	2488	1560	2.57
Contig 7	1355	404	1.22
Contig 8	1154	144	0.51
Contig 9	1112	125	0.46
Contig 10	1906	4009	8.64
Contig 18	1212	96	0.33
Contig 19	951	380	1.64
Contig 20 (A5)	2125	196	0.38
Contig 35	867	180	0.85
Contig 36	492	402	3.36
Contig 37	715	1074	6.17
Contig 38	914	1577	7.09
Contig 56	561	87	0.64
Contig 57	375	210	2.30
Contig 58	540	309	2.35
Contig 59	903	160	0.73
Contig 65	134	91	2.79

Continúa en la página siguiente

Tabla R1.S2 – Continuación de la página anterior

Contig	**Longitud**	**Lecturas**	**FCI**
Contig 68	475	2	0.02
Contig 69	474	5	0.04
Contig 71	507	297	2.41
Contig 72	123	2	0.07
Contig 74	507	2405	19.48
Contig 111	1341	2475	7.58
Contig 112	1197	3864	13.26
Contig 118	1175	818	2.86
Contig 119	540	1035	7.87
Contig 121	518	4	0.03
Contig 123	755	305	1.66
Contig 124	1269	482	1.56
Contig 125	1779	1077	2.49
Contig 127 (A13)	2120	2584	5.01
Contig 129	119	2	0.07
Contig 131	108	1	0.04
Contig 132	280	3	0.04
Contig 133	1426	542	1.56
Contig 143	414	96	0.95
Contig 146	112	17	0.62
Contig 147	649	48	0.30
Contig 151	166	9	0.22
Contig 152	101	26	1.06
Contig 153	818	541	2.72

Continúa en la página siguiente

Tabla R1.S2 – Continuación de la página anterior

Contig	**Longitud**	**Lecturas**	**FCI**
Contig 154	1882	5357	11.69
Contig 155	153	27	0.72
Contig 156	1314	2528	7.90
Contig 157	551	2356	17.56
Contig 158	1126	173	0.63
Contig 160	1567	598	1.57
Contig 173	498	177	1.46
Contig 176	584	291	2.05
Contig 177	855	564	2.71
Contig 183	1212	739	2.50
Contig 189	1522	1004	2.71
Contig 190	112	84	3.08
Contig 193	923	562	2.50
Contig 196	121	15	0.51

Tabla R1.S3: Porcentaje de genes de cada replicón de la cepa *S. meliloti* GR4 que se relaciona con cada una de las categorías funcionales de los COGs respecto al total de genes asociados con los COGs. Cr: cromosoma; GR4a: pRmeGR4a; GR4b: pRmeGR4b; GR4c: pRmeGR4c; GR4d: pRmeGR4d.

Código	**Categorías funcionales de los COGs**	**Cr**	**GR4a**	**GR4b**	**GR4c**	**GR4d**
A	RNA processing and modification	0'00	0'00	0'00	0'00	0'00
B	Chromatin structure and dynamics	0'03	0'00	0'00	0'00	0'00
C	Energy production and conversion	5'47	7'89	2'92	**9'83**	5'13
D	Cell cycle control, cell division, chromosome partitioning	0'79	3'51	1'17	0'34	0'70
E	Amino acid transport and metabolism	11'07	6'14	9'94	8'90	9'70
F	Nucleotide transport and metabolism	**2'55**	0'00	0'00	0'34	1'34
G	Carbohydrate transport and metabolism	7'58	4'39	1'75	8'22	**17'85**
H	Coenzyme transport and metabolism	4'03	1'75	4'09	2'46	2'88
I	Lipid transport and metabolism	3'65	1'75	0'58	3'31	3'65
J	Translation, ribosomal structure and biogenesis	**5'25**	0'88	0'58	0'76	1'26
K	Transcription	7'39	10'53	10'53	10'17	9'91

Continúa en la página siguiente

Tabla R1.S3 – Continuación de la página anterior

Código	Categorías funcionales de los COGs	Cr	GR4a	GR4b	GR4c	GR4d
L	Replication, recombination and repair	4′78	9′65	**35′09**	5′51	2′95
M	Cell wall/membrane/envelope biogenesis	4′53	4′39	2′34	3′14	**8′01**
N	Cell motility	1′67	0′00	2′34	1′19	0′35
O	Posttranslational modification, protein turnover, chaperones	3′96	1′75	2′92	2′54	1′41
P	Inorganic ion transport and metabolism	4′40	6′14	3′51	5′85	5′27
Q	Secondary metabolites biosynthesis, transport and catabolism	2′36	3′51	4′09	3′31	2′74
R	General function prediction only	12′17	9′65	4′68	13′05	11′52
S	Function unknown	11′45	11′40	4′68	8′81	9′21
T	Signal transduction mechanisms	3′71	7′02	3′51	**7′71**	4′43
U	Intracellular trafficking, secretion, and vesicular transport	2′04	8′77	5′26	3′05	0′28
V	Defense mechanisms	1′07	0′88	0′00	1′53	1′41
W	Extracellular structures	0′03	0′00	0′00	0′00	0′00
Y	Nuclear structure	0′00	0′00	0′00	0′00	0′00
Z	Cytoskeleton	0′00	0′00	0′00	0′00	0′00

CAPÍTULO 2

INTRONES DEL GRUPO II TIPO RmInt1. ANÁLISIS *IN SILICO* Y ESTUDIOS FUNCIONALES

Los intrones del grupo II se descubrieron a finales de los años 80 en genomas de organelas (mitocondrias y cloroplastos) pertenecientes a eucariotas inferiores y plantas superiores, donde interrumpen genes conservados (Michel *et al*, 1989). A principios de los años 90, Ferat y colaboradores identificaron por primera vez intrones del grupo II en proteobacterias y cianobacterias (Ferat & Michel, 1993; Ferat *et al*, 1994), siendo RmInt1 el primer intrón del grupo II descubierto en la familia Rhizobiaceae (Martínez-Abarca *et al*, 1998). Los avances en la tecnología de secuenciación masiva han permitido incrementar el número de genomas bacterianos secuenciados, lo que conlleva un aumento en el conocimiento de los elementos que componen los genomas, ya sean codificantes o no. Así se ha podido conocer más acerca de la distribución de los intrones del grupo II, encontrándose al menos un elemento de este tipo en aproximadamente uno de cada cuatro genomas bacterianos (Candales *et al*, 2011).

R2.1 INTRONES ENCONTRADOS EN LA CEPA *S. meliloti* GR4

Tras la secuenciación del genoma de GR4 realizamos una búsqueda de los genes anotados como reverso transcriptasas (RTs) en esta cepa, y encontramos RTs repetidas en el genoma y RTs de copia única (tabla R2.1). Como ya comentamos en el capítulo 1, GR4 contiene 10 copias del intrón RmInt1 y 7 copias de RmInt2, señalados como R1 y R2 respectivamente en la tabla R2.1. Entre las RTs de copia única descubrimos dos en el cromosoma: una se corresponde con el gen C770_GR4Chr2376, tiene una longitud de 505 aa

y también se encuentra en el genoma de otras cepas del género *Sinorhizobium* secuenciadas; y la otra se corresponde con el gen C770_GR4Chr2383, presenta una longitud de 569 aa y no muestra relación con los intrones del grupo II. En el plásmido pRmeGR4b encontramos dos RTs de copia única: una se corresponde con el gen C770_GR4pB012, tiene una longitud de 577

Tabla R2.1: RTs anotadas en *S. meliloti* GR4. Se indican los genes correspondientes a los intrones RmInt1 (R1) y RmInt2 (R2), y el porcentaje de identidad a nivel de aminoácido que muestran las RTs con respecto a la de RmInt1.

Replicón	**Nombre del gen**	**Longitud RT**	**% ID con RmInt1**	**Clase intrón**
Cromosoma	C770_GR4Chr1748 (R1)	419 aa	100	D
	C770_GR4Chr1869 (R1)	419 aa	100	D
	C770_GR4Chr2376	505 aa	28	G1
	C770_GR4Chr2383	569 aa	-	-
	C770_GR4Chr3316 (R1)	419 aa	100	D
	C770_GR4Chr3415 (R1)	419 aa	100	D
pRmeGR4b	C770_GR4pB012	577 aa	18	-
	C770_GR4pB077	pseudogen	-	E
	C770_GR4pB199 (R1)	419 aa	100	D
pRmeGR4c	C770_GR4pC0123 (R2)	419 aa	73	D
	C770_GR4pC0539 (R1)	419 aa	73	D
	C770_GR4pC0643 (R1)	419 aa	100	D
	C770_GR4pC0770 (R1)	419 aa	100	D
	C770_GR4pC0899 (R1)	419 aa	100	D
	C770_GR4pC1032 (R2)	419 aa	73	D
	C770_GR4pC1103 (R1)	419 aa	100	D
	C770_GR4pC1183 (R2)	419 aa	73	D
	C770_GR4pC1231 (R2)	419 aa	73	D
pRmeGR4d	C770_GR4pD0592 (R2)	419 aa	73	D
	C770_GR4pD0745	453 aa	27	C
	C770_GR4pD1533 (R2)	419 aa	73	D
	C770_GR4pD1537 (R2)	419 aa	73	D

aa y sólo se encuentran secuencias similares fuera de la familia Rhizobiaceae; y la otra, anotada como C770_GR4pB077, es un pseudogen. La última RT de copia única encontrada está en el plásmido pRmeGR4d, y se corresponde con el gen C770_GR4pD0745; contiene 453 aa, y también aparece distribuida entre otras cepas del género *Sinorhizobium* secuenciadas.

R2.2 DISTRIBUCIÓN DE LOS INTRONES DEL GRUPO II RELACIONADOS CON RmInt1

Con el fin de conocer los organismos que contienen elementos similares a RmInt1 se llevó a cabo una búsqueda en las bases de datos con la secuencia de 1.884 pb de este intrón (ver Material y Métodos, apartado M.12.1). El resultado de esta búsqueda pone de manifiesto que RmInt1 se encuentra presente en diversas cepas, y que los genomas contienen copias de este elemento tanto completas como parciales (tabla R2.2). Se observa que en la mayoría de las cepas que han sido secuenciadas RmInt1 está presente en más de un replicón y, a su vez, en más de una localización dentro de un mismo replicón. Las copias completas tienen un porcentaje de identidad con RmInt1 que varía entre el 100 % del intrón encontrado en *S. meliloti* Rm41 y el 88 % presentado por el intrón de la cepa *S. medicae* WSM419, al cual nos referiremos como SmedInt1. De las copias de intrón truncadas se observan fragmentos de muy distinta longitud e identidad, aunque llama la atención que 7 copias parciales presentan la misma longitud (desde el nucleótido 2 al 650) e identidades similares (entre el 87-93 %) en varias cepas.

Las copias de intrón completas se usaron para la construcción de árboles filogenéticos tanto a partir de la secuencia nucleotídica de la ribozima de éstos como de la secuencia de aminoácidos de la proteína que codifican. Junto a ellas se incluyó la secuencia de dos intrones presentes en la cepa *S. medicae* RMO09, obtenidos a partir de un ensayo de movilidad que se detalla en el apartado R2.3.2: uno similar a RmInt1 (RMO09-R1) y otro a SmedInt1 (RMO09-S1). Por tanto, el OTU nombrado como RMO09 en los árboles no representa al intrón incompleto de dicha cepa que aparece en la

Tabla R2.2: Resultado del BLASTn realizado con la secuencia del intrón del grupo II RmInt1 a fecha 22-01-13. En las cepas de *S. meliloti* se indica la equivalencia de los plásmidos a pSymA (A) y pSymB (B) de 1021.

Cepa bacteriana	**Localización genómica**	**Porcentaje cobertura**	**Porcentaje identidad**	**Nº copias intrón**	**Longitd secuencia**
S. meliloti Rm41	pSYMB (B)	100	100	1	1-1.884
S. meliloti GR4	Cromosoma	100	100	4	1-1.884
S. meliloti GR4	pRmeGR4b	100	99	1	1-1.884
S. meliloti GR4	pRmeGR4c (A)	100	99	5	1-1.884
S. meliloti 1021	pSymA	100	99	1	1-1.884
S. meliloti 1021	pSymB	100	99	2	1-1.884
S. meliloti 102F51	pSymB-B152 (B)	100	99	1	1-1.884
S. meliloti SM11	Cromosoma	100	99	2	1-1.884
S. meliloti SM11	pSmeSM11c (A)	100	99	3	1-1.884
S. meliloti SM11	pSmeSM11d (B)	100	99	1	1-1.884
S. meliloti BL225C	Cromosoma	100	99	2	1-1.884
S. meliloti BL225C	pSINMEB01 (A)	100	99	1	1-1.884
S. meliloti BL225C	pSINMEB02 (B)	100	99	2	1-1.884
S. fredii NGR234	pNGR234b	100	98	1	1-1.884*
E. adhaerens 5D19	ND	100	93	-	1-1.884
S. terangae ORS22	ND	100	89	-	1-1.884
S. medicae WSM419	Cromosoma	100	88	1	1-1.884
S. medicae WSM419	pSMED01	100	88	1	1-1.884
S. medicae WSM419	pSMED02	100	88	2	1-1.884
S. medicae RMO09	ND	91	99	-	134-1.850
S. medicae RMO09	ND	91	99	-	134-1.849
S. terangae ORS1009	ND	86	86	-	186-1.809
S. meliloti AK83	pSINME01	81	98	1	1-1.518
R. tropici CIAT899	pRtrCIAT899b	76	91	1	1-1.428
R. etli 8C-3	REB02	64	82	1	1-1.208
R. etli CFN42	p42d	64	75	1	1-1.208
R. etli CIAT652	pB	64	75	1	1-1.208
A. aromaticum EbN1	Cromosoma	57	75	2	1-1.082
A. aromaticum EbN1	Cromosoma	51	73	1	298-1.255
R. etli CFN42	p42a	45	91	1	1-856
R. etli CE3	ND	45	91	-	1-856
E. adhaerens R-6387	ND	38	94	-	1-714
R. etli Viking1	ND	38	94	-	1-714
S. medicae RMO02	ND	34	93	-	2-650
S. meliloti AK83	Cromosoma 3 (A)	34	90	2	2-650
S. meliloti C017	pHRC017	34	90	1	2-650
S. meliloti SM11	pSmeSM11c (A)	34	89	1	2-650
S. meliloti GR4	pRmeGR4b	34	88	1	2-650
S. medicae WSM419	pSMED02	34	87	1	2-650

* intrón originalmente interrumpido en el nucleótido 484 y reconstituido para formar la copia de intrón completa

tabla R2.2. En la construcción de esta filogenia también se incluyó la secuencia de 1.887 pb del nuevo intrón descubierto en GR4, RmInt2 (GR4-R2), que es un 74 % idéntico a RmInt1. Hay que señalar que, aunque haya cepas que contienen más de una copia de intrón tipo RmInt1, todas las copias de una misma cepa son idénticas entre sí, y por tanto, se usó sólo una de ellas para generar los árboles filogenéticos. Éste es el caso de las copias encontradas en BL225C, SM11 y WSM419; sin embargo, 1021 contiene tres copias de intrón un 99 % idénticas entre sí, y hemos seleccionado la copia que comparte más nucleótidos con la secuencia consenso obtenida a partir de esas tres copias (la insertada en IS*Rm2011-2* en pSymB). En la tabla R2.S1 se indica la localización de la copia de intrón de cada una de las cepas utilizada en la construcción de estos árboles.

Los intrones del grupo II más emparentados con RmInt1 presentan relaciones filogenéticas similares al analizar las secuencias de nucleótidos de la ribozima del intrón (figura R2.1A) y las secuencias de aminoácidos de la proteína codificada por ellos (figura R2.1B). La topología de los árboles se ve soportada por elevados valores de *bootstrap* (calculados por el método de ML) y de probabilidad posterior (calculado por el método de BI; ver Material y Métodos, apartado M.13). En la rama más basal encontramos a RmInt2, mientras que el resto de intrones se clasifican en dos clados bien diferenciados: uno compuesto por los intrones tipo RmInt1 propiamente dichos, donde encontramos el intrón originalmente descrito en la cepa GR4 (GR4-R1), y otro formado por los intrones que presentan un porcentaje de identidad inferior al 90 % con RmInt1. En este último clado aparece el intrón encontrado en *S. medicae* WSM419, SmedInt1, el cual tomaremos como representativo de este grupo para estudios posteriores. A pesar de la diferente posición que ocupa el intrón encontrado en las cepas ORS22, NGR234 y 5D19, un análisis mediante el test CADM global (ver Material y Métodos, apartado M.13) puso de manifiesto que ambos árboles son congruentes, con una fuerte significación (W=95 % y p=0’0001).

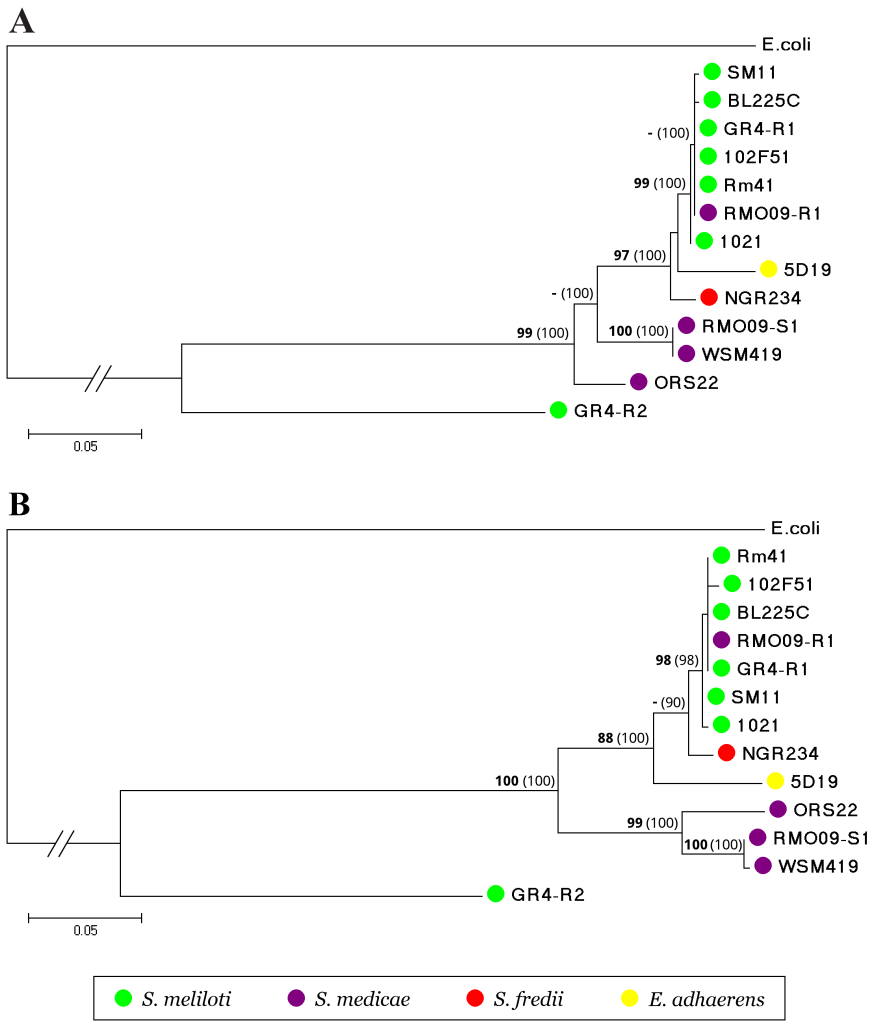


Figura R2.1: **Relación filogenética de los intrones del grupo II tipo RmInt1.** Se muestran los árboles obtenidos mediante el método de máxima verosimilitud. El soporte de las ramas indica los valores superiores a 75 de *bootstrap* (ML), en negrita, y de probabilidad posterior (BI), entre paréntesis. (A) Árbol generado a partir de la secuencia de nucleótidos de la ribozima de los intrones presentes en distintas cepas del grupo *Sinorhizobium/Ensifer*, mostradas en la tabla R2.2. (B) Árbol construido con la secuencia de aminoácidos de la proteína que codifican estos intrones. Como outgroup para enraizar los árboles se ha utilizado el intrón del grupo II IntB de la γ-proteobacteria *E. coli*. Ver Material y Métodos, apartado M.13, y tabla R2.S1 para más información sobre las secuencias utilizadas.

R2.2.1 Distribución de los nuevos intrones tipo RmInt1

Tras analizar la relación filogenética que muestran los intrones tipo RmInt1, observamos la aparición de dos nuevos clados fuera del grupo de RmInt1 propiamente dicho. Los intrones representativos de esos clados (RmInt2 y SmedInt1) han sido caracterizados en este trabajo (los resultados se muestran en el capítulo 3), y para conocer su distribución realizamos un análisis de BLASTn con ellos (tabla R2.3 y R2.4; ver Material y Métodos, apartado M.12.1). Hay que destacar que, hasta la fecha, en las bases de datos no se encuentran intrones con un 100 % de identidad a RmInt2 en ningún organismo salvo en GR4. Además, llevamos a cabo un análisis mediante PCR con cebadores específicos de RmInt2 sobre varios rizobios no secuenciados disponibles en el laboratorio, y tampoco encontramos copias de dicho intrón en ellos (figura R2.2). El intrón que más se parece a RmInt2 (con un 83 % de identidad) es el presente en el plásmido pRtrCIAT899b de la recientemente secuenciada *Rhizobium tropici* CIAT899 (tabla R2.3). El resto de intrones que aparecen al realizar el análisis de BLASTn con RmInt2 son SmedInt1 (en la cepa WSM419) y RmInt1 (en las distintas cepas donde

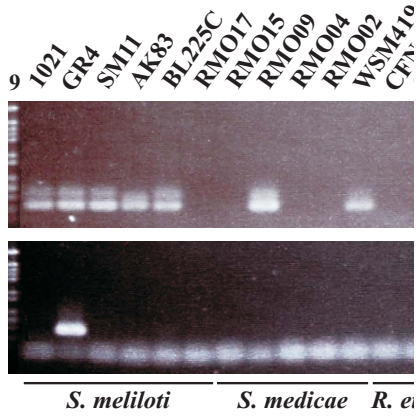


Figura R2.2: **Presencia de los intrones RmInt1, SmedInt1 y RmInt2 en diferentes rizobios.** Gel de agarosa en el que se muestra un ensayo de PCR realizado con dos pares de oligonucleótidos que permiten testar la presencia de RmInt1 o SmedInt1 por un lado, y de RmInt2 por otro, en varias cepas bacterianas del orden Rhizobiales.

Tabla R2.3: Resultado del BLASTn realizado con la secuencia del intrón del grupo II RmInt2 a fecha 22-01-13. En las cepas de *S. meliloti* se indica la equivalencia de los plásmidos a pSymA (A) y pSymB (B) de 1021.

Cepa bacteriana	**Localización genómica**	**Porcentaje cobertura**	**Porcentaje identidad**	**Nº copias intrón**	**Longitd secuencia**
S. meliloti GR4	pRmeGR4c (A)	100	100	4	1-1887
S. meliloti GR4	pRmeGR4d (B)	100	100	3	1-1887
R. tropici CIAT899	pRtrCIAT899b	97	83	1	54-1.887
S. medicae WSM419	Cromosoma	86	77	1	1-1623
S. medicae WSM419	pSMED01	86	77	1	1-1623
S. medicae WSM419	pSMED02	86	77	2	1-1623
S. meliloti GR4	Cromosoma	83	78	4	54-1.625
S. meliloti GR4	pRmeGR4b	83	78	1	54-1.625
S. meliloti GR4	pRmeGR4c (A)	83	78	5	54-1.625
S. meliloti Rm41	pSYMB (B)	83	78	1	54-1.625
S. meliloti 1021	pSymA	83	78	1	54-1.625
S. meliloti 1021	pSymB	83	78	2	54-1.625
S. meliloti 102F51	pSymB-B152 (B)	83	78	1	54-1.625
S. meliloti SM11	Cromosoma	83	78	2	54-1.625
S. meliloti SM11	pSmeSM11c (A)	83	78	3	54-1.625
S. meliloti SM11	pSmeSM11d (B)	83	78	1	54-1.625
S. meliloti BL225C	Cromosoma	83	78	2	54-1.625
S. meliloti BL225C	pSINMEB01 (A)	83	78	1	54-1.625
S. meliloti BL225C	pSINMEB02 (B)	83	78	2	54-1.625
S. terangae ORS22	ND	83	78	-	54-1.625
S. fredii NGR234	pNGR234b	83	76	1	54-1.625
S. meliloti AK83	pSINME01	77	78	1	54-1.525
R. etli 8C-3	REB02	42	81	1	53-860
R. etli CFN42	p42a	42	80	1	53-859
R. etli CE3	ND	42	80	-	53-859
R. etli CFN42	p42d	35	81	1	53-715
R. etli CIAT652	pB	35	80	1	53-715
E. adhaerens R-6387	ND	35	80	-	53-716
R. etli Viking1	ND	35	80	-	53-716

* intrón originalmente interrumpido en el nucleótido 484 y reconstituido para formar la copia de intrón completa

Tabla R2.4: Resultado del BLASTn realizado con la secuencia del intrón del grupo II SmedInt1 a fecha 22-01-13. En las cepas de *S. meliloti* se indica la equivalencia de los plásmidos a pSymA (A) y pSymB (B) de 1021.

Cepa bacteriana	Localización genómica	Porcentaje cobertura	Porcentaje identidad	Nº copias intrón	Longitd secuencia
S. medicae WSM419	Cromosoma	100	100	1	1-1.885
S. medicae WSM419	pSMED01	100	100	1	1-1.885
S. medicae WSM419	pSMED02	100	100	2	1-1.885
S. terangae ORS22	ND	100	94	-	1-1.885
S. meliloti GR4	Cromosoma	100	88	4	1-1.885
S. meliloti GR4	pRmeGR4b	100	88	1	1-1.885
S. meliloti GR4	pRmeGR4c (A)	100	88	5	1-1.885
S. meliloti Rm41	pSYMB (B)	100	88	1	1-1.885
S. meliloti 1021	pSymA	100	88	1	1-1.885
S. meliloti 1021	pSymB	100	88	2	1-1.885
S. meliloti 102F51	pSymB-B152 (B)	100	88	1	1-1.885
S. meliloti SM11	Cromosoma	100	88	2	1-1.885
S. meliloti SM11	pSmeSM11c (A)	100	88	3	1-1.885
S. meliloti SM11	pSmeSM11d (B)	100	88	1	1-1.885
S. meliloti BL225C	Cromosoma	100	88	2	1-1.885
S. meliloti BL225C	pSINMEB01 (A)	100	88	1	1-1.885
S. meliloti BL225C	pSINMEB02 (B)	100	88	2	1-1.885
S. fredii NGR234	pNGR234b	100	88	1	1-1.885
E. adhaerens 5D19	ND	100	87	-	1-1.885
S. medicae RMO09	ND	91	88	-	134-1.851
S. medicae RMO09	ND	91	88	-	134-1.850
S. terangae ORS1009	ND	86	92	-	195-1.810
R. tropici CIAT899	pRtrCIAT899b	86	76	1	1-1.620
S. meliloti AK83	pSINME01	80	88	1	1-1.519
R. etli 8C-3	REB02	65	85	1	1-1.220
R. etli CFN42	p42d	65	78	1	1-1.220
R. etli CIAT652	pB	65	78	1	1-1.220
A. aromaticum EbN1	Cromosoma	56	76	2	1-1.063
A. aromaticum EbN1	Cromosoma	53	74	1	279-1.281
R. etli CFN42	p42a	45	94	1	1-857
R. etli CE3	ND	45	94	-	1-857
E. adhaerens R-6387	ND	38	94	-	1-715
R. etli Viking1	ND	38	94	-	1-715
S. medicae RMO02	ND	34	91	-	2-651
S. meliloti GR4	pRmeGR4b	34	90	1	2-651
S. meliloti SM11	pSmeSM11c (A)	34	90	1	2-651
S. meliloti AK83	Cromosoma 3 (A)	34	90	2	2-651
S. meliloti C017	pHRC017	34	89	1	2-651
S. medicae WSM419	pSMED02	34	88	1	2-651

* intrón originalmente interrumpido en el nucleótido 484 y reconstituido para formar la copia de intrón completa

hemos visto que se localiza). Respecto al análisis de BLASTn llevado a cabo con SmedInt1, el intrón descrito en la cepa *S. terangae* ORS22 presenta la máxima identidad mostrada por un intrón no encontrado en WSM419 (un 94 %). El resto de intrones que aparecen son, como en el caso de RmInt2, las distintas copias de RmInt1 distribuidas por el genoma de otros rizobios (tabla R2.4).

R2.2.2 Distribución de las ISs diana asociadas a los tres intrones representativos: RmInt1, RmInt2 y SmedInt1

Con el fin de conocer el grado de dispersión de las ISs a las que se asocian los intrones estudiados, y estimar el número de dianas que pueden ser invadidas por estos intrones, llevamos a cabo un análisis de BLASTn con dichas ISs (ver Material y Métodos, apartado M.12.1): IS*Rm2011-2*, donde se inserta RmInt1 (Martínez-Abarca *et al*, 1998), IS*Rm17*, a la que se encuentra asociado RmInt2 (concretamente invade la repetición invertida, IR), e IS*Sme3* (una nueva IS descrita en este trabajo; capítulo 3), donde se inserta SmedInt1. El resultado de estos análisis reveló que tanto el número de copias por genoma como el número de cepas donde encontramos cada una de las ISs es diferente, siendo GR4, 1021, SM11, AK83, Rm41 y WSM419 los aislados que contienen copias de las tres ISs en su genoma (tabla R2.5, R2.6 y R2.7).

La IS que presenta mayor número de copias en los genomas analizados es IS*Rm2011-2*, con 62 copias completas (de entre 81-100 % de identidad) distribuidas en 9 cepas (tabla R2.5). Las cepas GR4, 1021 y AK83 son las que presentan más copias de esta IS, con 13, 12 y 12 copias respectivamente. Hay que destacar que, de todas las copias de IS*Rm2011-2* encontradas, sólo la copia truncada presente en BL225C no mantiene la diana de inserción de RmInt1. IS*Rm17* presenta 35 copias completas (de entre 81-100 % de identidad) repartidas en 14 cepas (tabla R2.6). Teniendo en cuenta que RmInt2 invade la IR de esta IS, realmente se encuentran 70 dianas disponibles para dicho intrón. El plásmido pSINMEB01 (de la cepa BL225C) es el que con-

tiene mayor número de copias de IS*Rm17*, con 6 copias completas y una truncada. De esta IS se encuentran 23 copias parciales en los genomas analizados, de las cuales 18 contienen al menos una diana para RmInt2 (una IR) y 5 han perdido las dos. Hay que señalar que la copia truncada de IS*Rm17* presente en el plásmido pRtrCIAT899b no se encuentra asociada a la copia

Tabla R2.5: Resultado del BLASTn realizado con la secuencia de IS*Rm2011-2* a fecha 22-01-13. En las cepas de *S. meliloti* se indica la equivalencia de los plásmidos a pSymA (A) y pSymB (B) de 1021.

Cepa bacteriana	**Localización genómica**	**Porcentaje cobertura**	**Porcentaje identidad**	**Nº copias IS**	**Longitd secuencia**
S. meliloti 1021	Cromosoma	100	100	6	1-1.053
S. meliloti 1021	pSymA	100	100	4	1-1.053
S. meliloti 1021	pSymB	100	100	2	1-1.053
S. meliloti GR4	Cromosoma	100	100	7	1-1.053
S. meliloti GR4	pRmeGR4b	100	100	2	1-1.053
S. meliloti GR4	pRmeGR4c (B)	100	100	4	1-1.053
S. meliloti SM11	Cromosoma	100	100	4	1-1.053
S. meliloti SM11	pSmeSM11c (A)	100	100	3	1-1.053
S. meliloti BL225C	Cromosoma	100	100	6	1-1.053
S. meliloti BL225C	pSINMEB01 (A)	100	100	1	1-1.053
S. meliloti BL225C	pSINMEB02 (B)	100	100	2	1-1.053
S. meliloti AK83	Cromosoma 1	100	100	1	1-1.053
S. meliloti AK83	Cromosoma 2 (B)	100	100	3	1-1.053
S. meliloti AK83	Cromosoma 3 (A)	100	100	7	1-1.053
S. meliloti AK83	pSINME01	100	99	1	1-1.053
S. meliloti Rm41	Cromosoma	100	99	1	1-1.053
S. meliloti Rm41	pSYMB (B)	100	99	1	1-1.053
S. meliloti Rm41	pSYMA (A)	100	99	2	1-1.053
S. meliloti 102F51	pSymB-B152 (B)	100	99	1	1-1.053
S. medicae WSM419	Cromosoma	100	99	1	1-1.053
S. medicae WSM419	pSMED02	100	99	1	1-1.053
R. leguminosarum bv. trifolii WSM2304	Cromosoma	100	81	1	1-1.053
R. leguminosarum bv. trifolii WSM2304	pRLG20	100	81	1	1-1.053
S. meliloti SM11	pSmeSM11d (B)	78	100	1	286-1.053
S. meliloti BL225C	pSINMEB01 (A)	31	94	1	730-1.053

Tabla R2.6: Resultado del BLASTn realizado con la secuencia de IS*Rm17* a fecha 22-01-13. En las cepas de *S. meliloti* se indica la equivalencia de los plásmidos a pSymA (A) y pSymB (B) de 1021.

Cepa bacteriana	**Localización genómica**	**Porcentaje cobertura**	**Porcentaje identidad**	**Nº copias IS**	**Longitd secuencia**
S. meliloti Rm41	Cromosoma	100	100	1	1-1.664
S. meliloti 1021	Cromosoma	100	100	3	1-1.664
S. meliloti 1021	pSymA	100	99	2	1-1.664
S. meliloti Rm41	pSYMB (B)	100	99	1	1-1.664
S. meliloti GR4	pRmeGR4c (A)	100	99	3	1-1.664
S. meliloti GR4	pRmeGR4d (B)	100	99	2	1-1.664
S. meliloti SM11	Cromosoma	100	99	2	1-1.664
S. meliloti AK83	Cromosoma 1	100	99	2	1-1.664
S. meliloti AK83	Cromosoma 2 (B)	100	99	2	1-1.664
S. meliloti AK83	Cromosoma 3 (A)	100	99	3	1-1.664
S. meliloti AK83	pSINME01	100	99	1	1-1.664
S. meliloti BL225C	Cromosoma	100	99	1	1-1.664
S. meliloti BL225C	pSINMEB01 (A)	100	99	6	1-1.664
S. meliloti C017	pHRC017	100	99	2	1-1.664
S. meliloti C017	pHRC017	100	83	1	1-1.664
R. leguminosarum bv. viciae 3841	Cromosoma	100	82	1	1-1.664
R. leguminosarum bv. viciae 3841	pRL7	100	82	1	1-1.664
S. medicae WSM419	pSMED02	100	81	1	1-1.664
S. fredii NGR234	pNGR234b	98	85	1	30-1.664
S. fredii NGR234	Cromosoma	98	82	3	30-1.664
S. medicae WSM419	pSMED03	97	81	1	57-1.664
M. loti R7A	ND	96	83	1	60-1.664
S. meliloti LPU88	ND	90	99	1	160-1.664
R. tropici CIAT899	pRtrCIAT899b	85	79	1	51-1.480
S. meliloti AK83	Cromosoma 2 (B)	79	99	1	346-1.664
S. fredii NGR234	pNGR234a	64	93	1	393-1.459
M. loti MAFF303099	Cromosoma	59	84	1	681-1.664
S. meliloti Rm41	Cromosoma	56	99	1	1-958
S. meliloti GR4	pRmeGR4c (A)	56	99	1	1-939
S. meliloti Rm41	Cromosoma	43	99	1	951-1.664
S. meliloti 1021	pSymB	40	98	1	1-672
M. loti MAFF303099	Cromosoma	38	83	1	51-685
S. meliloti SM11	pSmeSM11c (A)	24	97	1	1.042-1.473
S. meliloti GR4	pRmeGR4c (A)	24	92	1	1.042-1.473
S. meliloti C017	pHRC017	23	93	1	1-377
S. meliloti SM11	pSmeSM11b	22	99	1	1-365
S. meliloti BL225C	pSINMEB01 (A)	22	96	1	1.303-1.664
S. meliloti AK83	Cromosoma 2 (B)	21	99	1	1-351

Tabla R2.7: Resultado del BLASTn realizado con la secuencia de IS*Sme3* a fecha 22-01-13. En las cepas de *S. meliloti* se indica la equivalencia de los plásmidos a pSymA (A) y pSymB (B) de 1021.

Cepa bacteriana	**Localización genómica**	**Porcentaje cobertura**	**Porcentaje identidad**	**Nº copias IS**	**Longitd secuencia**
S. medicae WSM419	Cromosoma	100	100	1	1-3.149
S. medicae WSM419	pSMED02	100	100	2	1-3.149
S. medicae WSM419	pSMED03	100	100	1	1-3.149
S. meliloti AK83	Cromosoma 2 (B)	100	99	1	1-3.149
S. meliloti AK83	Cromosoma 3 (A)	100	95	2	1-3.149
S. meliloti GR4	pRmeGR4b	100	95	1	1-3.149
S. meliloti Rm41	Cromosoma	100	95	2	1-3.149
S. meliloti Rm41	pSYMA (A)	100	95	1	1-3.149
S. meliloti C017	pHRC017	100	95	1	1-3.149
S. meliloti SM11	pSmeSM11d (B)	100	95	1	1-3.149
S. meliloti SM11	pSmeSM11c (A)	100	94	2	1-3.149
O. anthropi ATCC 49188	Cromosoma 1	99	84	1	1-3.132
O. anthropi ATCC 49188	Cromosoma 2	99	83	1	1-3.132
S. meliloti SM11	pSmeSM11b	97	90	1	1-3.086
O. anthropi ATCC 49188	pOANT03	83	85	1	1-2.631
R. tropici CIAT899	pRtrCIAT899b	79	86	1	663-3.149
S. meliloti 1021	Cromosoma	50	89	1	1-1.601
S. meliloti SM11	pSmeSM11a	40	85	1	1-1.246
S. meliloti AK83	pSINME01	30	82	1	1-933
R. tropici CIAT899	pRtrCIAT899b	29	86	1	759-1.680
S. medicae WSM419	pSMED01	28	79	1	2.268-3.149
S. meliloti AK83	pSINME01	27	86	1	2.294-3.149

de intrón que contiene (un 83 % idéntico a RmInt2; tabla R2.3), sino que ambos elementos están a 118 kb de distancia. IS*Sme3* es, de las tres ISs analizadas, la menos extendida entre los genomas secuenciados, presentando 15 copias completas (de entre 94-100 % de identidad) en 9 cepas distintas (tabla R2.7). Además, encontramos 22 copias fragmentadas de esta IS, de las que 16 mantienen la diana para SmedInt1. Hay que destacar que detrás de la copia truncada de IS*Sme3* presente en el plásmido pRtrCIAT899b, que acaba en la posición 1.680 de la IS (sitio de inserción de SmedInt1), se encuentra una copia parcial del intrón (un fragmento del nucleótido 1 al 292 con un 96 % de identidad a SmedInt1).

R2.3 MOVILIDAD DE INTRONES DEL GRUPO II TIPO RmInt1 EN DIFERENTES RIZOBIOS

El hecho de que los intrones del grupo II relacionados con RmInt1 se localicen en distintas cepas del género *Sinorhizobium* corrobora que el genoma de este grupo bacteriano contiene la maquinaria necesaria para que dichos intrones puedan realizar su mecanismo de movilidad, y puedan así expandirse. Sin embargo, la eficiencia de movilidad que presenta RmInt1 varía en diferentes fondos genéticos (Fernández-López *et al*, 2005). Por este motivo, llevamos a cabo dos tipos de ensayos utilizando varios aislados disponibles en el laboratorio: uno con objeto de analizar la eficiencia de movilidad de RmInt1 expresado desde un plásmido, y otro para determinar la movilidad de copias de intrón genómicas de algunas cepas.

La clasificación bacteriana de ciertos aislados puede llegar a ser un tema controvertido entre los microbiólogos. Las especies del género *Sinorhizobium* están muy emparentadas, y el descubrimiento de nuevos aislados conlleva la identificación de éstos para incluirlos en la especie bacteriana correspondiente. Dourado *et al* (2009) han descrito unos oligonucleótidos

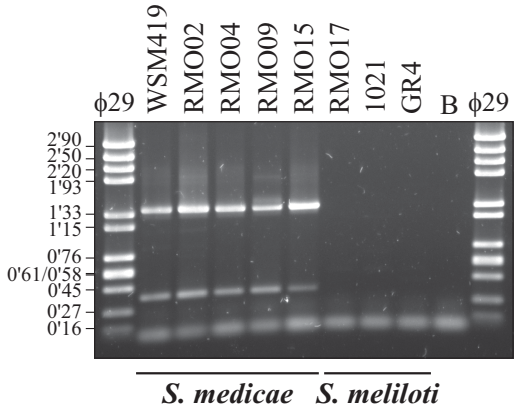


Figura R2.3: **Determinación de cepas pertenecientes a la especie bacteriana *S. medicae*.** Gel de agarosa en el que se muestra un ensayo de PCR llevado a cabo con una pareja de oligonucleótidos específica de *S. medicae* en las cepas del género *Sinorhizobium* indicadas.

específicos para la especie *S. medicae* (SMED01-f, 5'-GCGCGTAGTTCTG AAAGACC-3', y SMED01-r, (5'-G GTTCTGGACGATCGTGTTT-3'), que han permitido confirmar la identidad de varias cepas utilizadas en este trabajo (figura R2.3). La cepa WSM419 sirvió como control positivo en la reacción de PCR llevada a cabo, mientras que 1021 y GR4 se usaron como control negativo. El resultado de este ensayo corroboró que RMO17 pertenece a *S. meliloti*, y que las cepas RMO02, RMO04, RMO09 y RMO15 son de la especie *S. medicae*.

R2.3.1 Ensayo de movilidad de plásmido a plásmido

Con objeto de determinar la eficiencia de movilidad del intrón RmInt1 en rizobios que no contuvieran copias genómicas de éste, llevamos a cabo un ensayo de doble plásmido (ver Material y Métodos, apartado M.10) con cepas pertenecientes a las especies *S. medicae* y *R. etli*. Algunos de estos aislados contienen copias parciales, y por tanto inactivas, de RmInt1 (tabla R2.2). En este ensayo, donde también incluimos las cepas 1021 y RMO17 de *S. meliloti*, usamos como plásmido donador la construcción pKGEMA4 y como receptor los plásmidos con la diana IS*Rm2011-2* en ambas orientaciones, LAG (pJB0.6LAG) y LEAD (pJB0.6LEAD).

El porcentaje de movilidad que presentó RmInt1 en las diferentes cepas analizadas se muestra en la figura R2.4. En todas ellas el intrón se insertó más eficientemente sobre la diana en orientación LAG que sobre la diana en LEAD (figura R2.4C). Respecto al ensayo realizado sobre el plásmido pJB0.6LAG (figura R2.4A), en las cepas RMO15 y RMO04 RmInt1 presentó un porcentaje de movilidad del 100 %. En los aislados WSM419, RMO17, RMO02 y CIAT652 también se observó una elevada eficiencia de movilidad del intrón, con un 92 %, un 88 %, un 79 % y un 76 % respectivamente. El porcentaje de movilidad que mostró RmInt1 en las cepas CFN42 y 1021 fue menor, con un 56 % y un 21 % respectivamente. En cuanto al ensayo realizado sobre el plásmido pJB0.6LEAD (figura R2.4B), fue en las cepas RMO15 y RMO04 donde RmInt1 presentó el mayor porcentaje de movilidad (un 41 %

en ambos casos). En WSM419 y RMO02 se observó una eficiencia similar (37 % y 36 % respectivamente), que, a su vez, fue superior al porcentaje encontrado en CIAT652 (28 %), RMO17 (28 %) y CFN42 (15 %). En 1021 no se detectaron eventos de movilidad de RmInt1 hacia su diana en orientación LEAD.

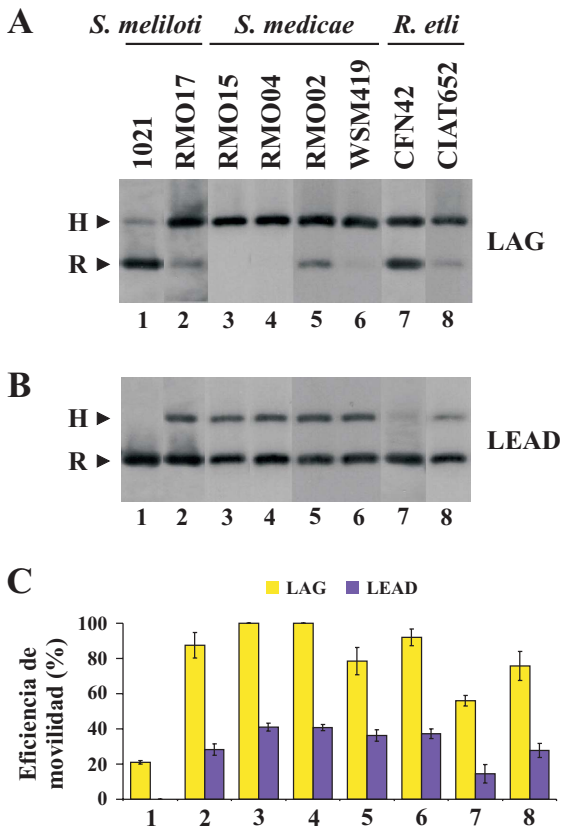


Figura R2.4: **Eficiencia de movilidad de RmInt1 en diferentes rizobios.** Las hibridaciones DNA-DNA muestran el resultado del ensayo de movilidad llevado a cabo con un plásmido donador del intrón RmInt1 en su forma derivada (pKGEMA4) y un plásmido receptor con la diana IS*Rm2011-2* en ambas orientaciones, LAG (A) y LEAD (B). Las flechas señalan el producto de *homing* (H) y el plásmido receptor sin invadir (R). (C) Eficiencia de movilidad de RmInt1 en las cepas indicadas (calculada como [H/(H+R)]*100), con su correspondiente error estándar.

R2.3.2 Ensayo de movilidad de genoma a plásmido

Estudios previos han puesto de manifiesto que la cepa *S. medicae* RMO09 contiene, al menos, tres copias del intrón RmInt1 (Fernández-López *et al*, 2005). Dos de ellas han sido secuenciadas y publicadas en las bases de datos (RMO09-1 y RMO09-2; tabla R2.2 y M.6). Con el fin de conocer la capacidad de movilidad de las copias de intrón genómicas presentes en cepas de la especie *S. medicae*, llevamos a cabo un ensayo (ver Material y Métodos, apartado M.10) con RMO09 y WSM419, usando la cepa RMO15 carente de copias genómicas de intrón como control negativo. Los plásmidos receptores utilizados en el ensayo contenían las dianas IS*Rm2011-2* (diana para RmInt1) e IS*Sme3* (diana para SmedInt1) en orientación LAG (figura R2.5). El resultado de este análisis puso de manifiesto la movilidad de las copias genómicas de SmedInt1 presentes en WSM419, y su especificidad hacia la diana donde se ha encontrado insertado (IS*Sme3*). Llama la atención el resultado mostrado por RMO09, cuyas copias genómicas de intrón presentan movilidad hacia los dos tipos de dianas. El aislamiento de formas H (plásmidos

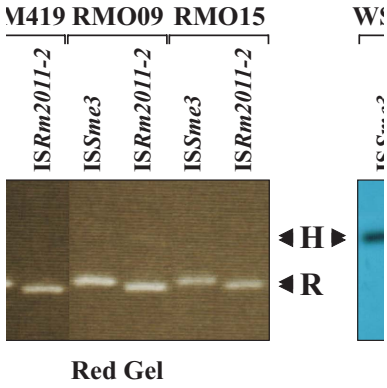


Figura R2.5: **Ensayo de movilidad de copias genómicas de intrones tipo RmInt1.** En cada carril se indica la diana que porta el plásmido receptor (R) introducido en las distintas cepas. La hibridación con la sonda RmInt1 muestra el producto de *homing* (H) que aparece en las cepas WSM419 y RMO09 (que contienen copias genómicas de intrón). RMO15 se usó como control negativo ya que su genoma no presenta copias de intrones tipo RmInt1.

donde ha ocurrido la invasión de la diana) procedentes de ambos plásmidos receptores, y su posterior secuenciación, reveló que la copia de intrón que había invadido la diana IS*Rm2011-2* era 100 % idéntica a RmInt1, y la que se insertó en IS*Sme3* tenía un 100 % de identidad con SmedInt1 (secuencias utilizadas en la construcción de los árboles filogenéticos mostrados al inicio del capítulo, apartado R2.2).

R2.4 FACTORES DEL HOSPEDADOR IMPLICADOS EN LA DISPERSIÓN DE INTRONES DEL GRUPO II (EFECTO DE LOS GENES *groEL*)

De los análisis mostrados en apartados anteriores, así como de estudios previos (Martínez-Abarca *et al*, 1998; Fernández-López *et al*, 2005; Biondi *et al*, 2011), se deduce que el movimiento de los intrones del grupo II se encuentra especialmente adaptado a bacterias de la familia Rhizobiaceae. A pesar de que estos elementos son autocatalíticos, se han descrito proteínas del hospedador que facilitan su escisión en la célula y, por tanto, su movilidad (Beauregard *et al*, 2008). Éste es el caso de las chaperonas, que son proteínas que contribuyen al correcto plegamiento de otras proteínas. GroEL es una chaperona que puede asistir el plegamiento de muchas proteínas de diferente tamaño, secuencia y estructura, y generalmente se encuentra en un operón junto a la cochaperona GroES. Esta chaperona está presente en casi todos los genomas bacterianos secuenciados, y un gran número de bacterias contienen dos o más genes con esta función (tabla R2.8; Lund, 2009). Hay que destacar que las α-proteobacterias pueden tener hasta 7 genes *groEL*, y, en concreto, en el genoma de *S. meliloti* 1021 encontramos cuatro operones *groESL* y un gen *groEL*. Para descubrir la posible redundancia funcional de estos genes, Bittner *et al* (2007) generaron una serie de mutantes de 1021 (tabla R2.9) que han sido usados en la presente Tesis Doctoral para evaluar la implicación de la chaperona GroEL en la escisión del intrón RmInt1.

Tabla R2.8: Distribución de los genes *groEL* entre 664 genomas bacterianos completamente secuenciados. Adaptada de Lund, 2009.

Filo bacteriano	**Número de homólogos del gen *groEL***								
	Total	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Actinobacteria	54	0	9	40	2	3	0	0	0
Alphaproteobacteria	90	0	55	18	8	2	5	1	1
Aquificae	1	0	1	0	0	0	0	0	0
Bacteroidetes-Chlorobi	23	0	22	1	0	0	0	0	0
Betaproteobacteria	60	0	35	17	6	0	1	1	0
Chlamydiae-Verrucomicrobia	13	0	0	0	13	0	0	0	0
Chloroflexi	7	0	4	3	0	0	0	0	0
Cyanobacteria	33	0	0	30	3	0	0	0	0
Deinococcus-Thermus	4	0	4	0	0	0	0	0	0
Deltaproteobacteria	19	0	12	6	1	0	0	0	0
Epsilonproteobacteria	20	0	20	0	0	0	0	0	0
Fibrobacter-Acidobacteria	2	0	2	0	0	0	0	0	0
Firmicutes	148	13	125	10	0	0	0	0	0
Fusobacteria	1	0	1	0	0	0	0	0	0
Planctomycetes	1	0	0	0	1	0	0	0	0
Gammaproteobacteria	172	0	158	12	2	0	0	0	0
Spirochaetes	14	0	14	0	0	0	0	0	0
Thermotogae	7	0	5	2	0	0	0	0	0
Total	669	13	467	139	36	5	6	2	1

Para llevar a cabo ese estudio, primeramente realizamos un ensayo de extensión a partir de cebador (ver Material y Métodos, apartado M.9) con RNA total extraído de 1021 y todos los mutantes groEL a los que previamente se les había introducido el plásmido donador de intrón pCm4. Para determinar el porcentaje de escisión de RmInt1 en los diferentes mutantes se tomó como referencia la escisión de éste en 1021 (cepa silvestre que contiene todos los genes *groEL*), considerándola una eficiencia del 100 % (figura R2.6). Respecto a los mutantes simples, RmInt1 mostró la mayor eficiencia de escisión en NI001 (carril 9 en la figura R2.6), un 68 %, aunque la escisión

Tabla R2.9: Relación de mutantes groEL de *S. meliloti* 1021 utilizados en este trabajo.

Nombre	Gen/Operón mutado	Resistencia	Referencia
AB249	*groEL1*	Sm, Nm	Bittner *et al* (2007)
AB247	*groEL2*	Sm, Nm	Bittner *et al* (2007)
AF14	*groESL3*	Sm, Tc	Bittner & Oke (2006)
VO3193	*groEL4*	Sm	Bittner & Oke (2006)
NI001	*groESL5*	Sm, Gm	Mitsui *et al* (2004)
AB221	*groEL1 groESL3*	Sm, Nm, Tc	Bittner *et al* (2007)
AB219	*groEL1 groESL5*	Sm, Nm, Gm	Bittner *et al* (2007)
AB257	*groEL1 groESL3 groEL4 groESL5*	Sm, Nm, Gm, Tc	Bittner *et al* (2007)
AB238	*groEL2 groESL3 groEL4 groESL5*	Sm, Nm, Gm, Tc	Bittner *et al* (2007)

de este intrón observada en VO3193, AB249 y AB247 (carriles 10, 5 y 4 respectivamente en la figura R2.6) no fue inferior al 55 % (62 %, 60 % y 57 % respectivamente). Entre estos mutantes simples hay que destacar que AF14, carente del gen *groEL3* mediante deleción del operón *groESL3* (carril 8 en la figura R2.6), mostró una disminución drástica de la escisión de RmInt1 respecto a 1021 (11 % de eficiencia). En el mutante doble AB221 (carril 3 en la figura R2.6) fue donde RmInt1 presentó mayor porcentaje de producto escindido (81 %), mientras que en AB219 (carril 2 en la figura R2.6) mostró un 49 %. En los dos cuádruples mutantes (AB238 y AB257; carriles 6 y 7 respectivamente en la figura R2.6) no se detectó escisión del intrón.

Con objeto de conocer la relación evolutiva que presenta esta chaperona en distintos rizobios, realizamos un análisis filogenético de los genes *groEL* y *groES* de las dos cepas del género *Sinorhizobium* secuenciadas y publicadas hasta el año 2010 (*S. meliloti* 1021, Galibert *et al*, 2001; y *S. medicae* WSM419, Reeve *et al*, 2010), y de la cepa *S. meliloti* GR4 (figura R2.7). WSM419 contiene tres operones *groESL* y dos genes *groEL* (*groEL3* y *groEL4*); GR4 presenta cinco operones *groESL* y el gen *groEL4* sin cochaperona groES. En los árboles, los genes de WSM419 y GR4 se han denominado con el nombre del gen al que corresponde en la anotación de cada cepa debido a que no se en-

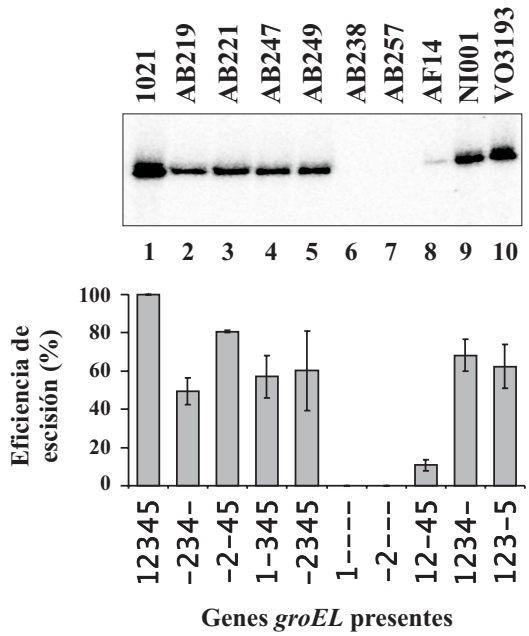


Figura R2.6: **Eficiencia de escisión de RmInt1 en los mutantes groEL de 1021.** El ensayo de extensión a partir de cebador se realizó introduciendo un plásmido donador que porta el intrón RmInt1 en su forma derivada (pCm4) en los mutantes indicados en cada carril. El gráfico de barras muestra la eficiencia de escisión del intrón (con su correspondiente error estándar), calculada respecto al resultado obtenido en 1021, frente a los genes *groEL* expresados en cada caso. El guión (-) indica los genes *groEL* mutados en la cepa.

cuentran identificadas las distintas copias de estos genes. El árbol generado con los genes *groEL* (figura R2.7A) presentó una topología diferente al árbol construido a partir de los genes *groES* (figura R2.7B). No obstante, en ambos casos se formaron tres clados soportados por elevados valores de *boostrap* y probabilidad posterior (ver Material y Métodos, apartado M.13): uno representado por los operones *groESL1* y *groESL2*, otro constituido por el operón *groESL3* y un tercero que engloba al operon *groESL5* y al gen *groEL4*. En la filogenia obtenida a partir de los genes *groEL*, el gen *groEL3* mostró la rama de mayor longitud, y el clado *groEL1-groEL2* se posicionó en la rama basal. Respecto a la filogenia de los genes *groES*, el gen *groES5* fue el que acumuló más cambios, y en la rama basal apareció el gen *groES3*.

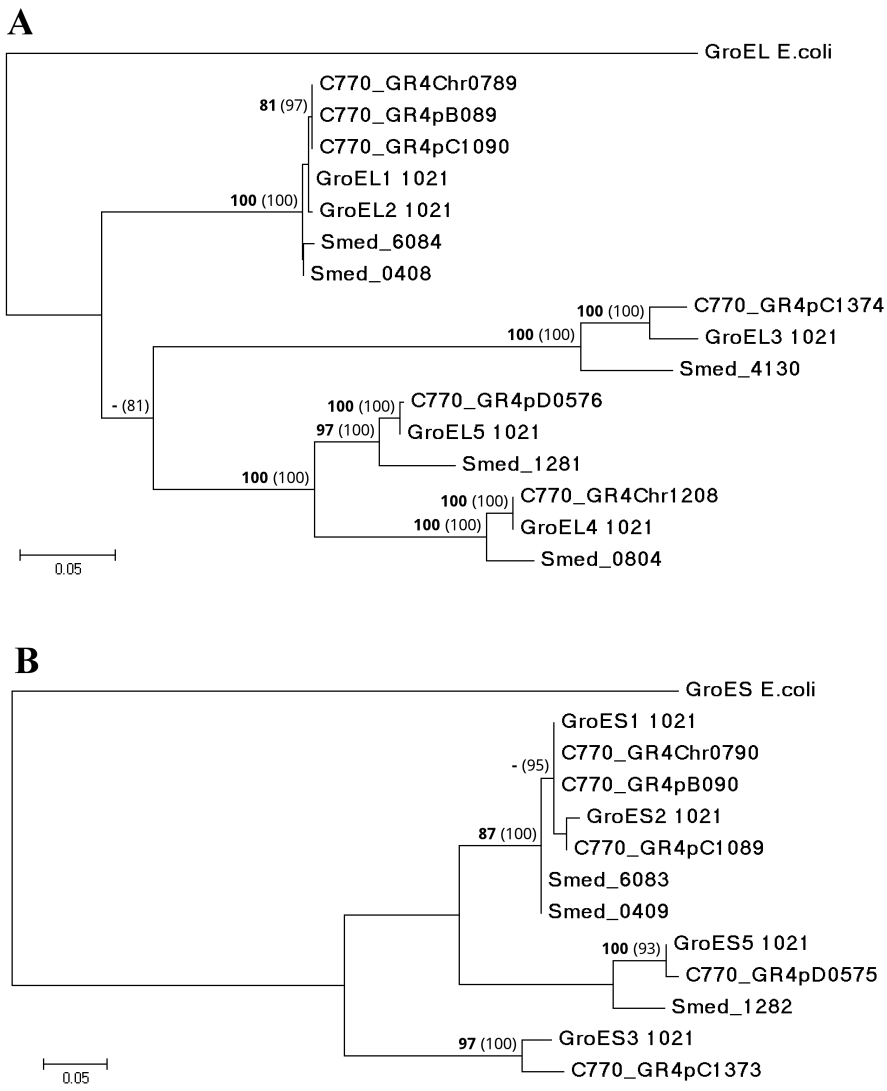


Figura R2.7: **Relación filogenética de los genes *groEL* y *groES* presentes en el genoma de 1021, GR4 y WSM419.** Se muestran los árboles obtenidos mediante el método de máxima verosimilitud. El soporte de las ramas indica los valores superiores a 75 de *bootstrap* (ML), en negrita, y de probabilidad posterior (BI), entre paréntesis. Los árboles se han generado a partir de la secuencia de aminoácidos de los genes *groEL* (A) y *groES* (B), utilizando como outgroup el gen *groEL* y *groES* de *E. coli* respectivamente.

Todos estos resultados nos llevaron a pensar que el operón *groESL3* podía estar jugando un papel importante en la escisión del intrón RmInt1. Por ello, a partir de un plásmido compatible con los mutantes y con el plásmido donador de intrón (pJBΔ129, con resistencia a Ap y Tc), expresamos en *trans* el operón *groESL3* (clonado en ambas orientaciones: sentido, pJBGroESL3s, y antisentido, pJBGroESL3as; ver Material y Métodos apartado M.2.3). Así, pretendíamos realizar un ensayo de complementación de los mutantes que presentaban una escisión del intrón RmInt1 menor del 20 %: el mutante simple del operón *groESL3* (AF14) y los dos cuádruples mutantes (AB238 y AB257). Además, podríamos comprobar si la orientación del operón dentro del plásmido tiene algún efecto en su expresión y en la complementación llevada a cabo. En una primera ronda de conjugación transferimos el plásmido donador de intrón pCm4 a dichos mutantes y a 1021; en una segunda ronda les introdujimos los plásmidos pJBΔ129 (como control negativo de la reacción), pJBGroESL3s y pJBGroESL3as, de manera independiente. Al analizar las colonias obtenidas de todas las muestras tras esta segunda ronda de conjugación, apreciamos cierto grado de inestabilidad del plásmido pCm4 y, por tanto, no pudimos llevar a cabo el ensayo de extensión a partir de cebador con estas muestras. Por este motivo, decidimos realizar el ensayo de complementación sólo con el mutante AF14, que sí permitía evaluar la escisión de RmInt1 desde el plásmido donador de intrón pKGEMA4 (con resistencia a Km; tabla R2.9). En la figura R2.8 se indica el porcentaje de escisión de RmInt1 en 1021 y AF14 dependiendo del plásmido pJB introducido. Como referencia para determinar dicho porcentaje se utilizó la muestra de 1021 complementada con el plásmido pJBΔ129 (carril 1 en la figura R2.8), en la cual consideramos un 100 % de eficiencia de escisión de RmInt1. Al introducir en la cepa 1021 un plásmido que contenía el operón *groESL3* en sentido (carril 2 en la figura R2.8), el porcentaje de escisión del intrón aumentó al 104 %; sin embargo, cuando dicho operón se encontraba en antisentido en el plásmido (carril 3 en la figura R2.8), RmInt1 presentó una eficiencia de escisión del 96 %. En el mutante AF14 se observó un patrón similar al de la cepa silvestre: con el plásmido carente del operón *groESL3* (carril 4 en la figura R2.8) RmInt1 mostró un

porcentaje de escisión del 67 %; al expresar dicho operón desde el plásmido pJBGroESL3s (carril 5 en la figura R2.8), el porcentaje se incrementó al 72 %; y desde pJBGroESL3as (carril 6 en la figura R2.8), la eficiencia de escisión de RmInt1 disminuyó al 60 %.

En el ensayo de extensión a partir de cebador realizado con el plásmido pKGEMA4 (figura R2.8) se observó una mayor eficiencia de escisión del intrón RmInt1 en el mutante AF14 que al realizar ese ensayo con el plásmido pCm4 (figura R2.6; 67 % frente a 11 %). Este resultado nos llevó a analizar

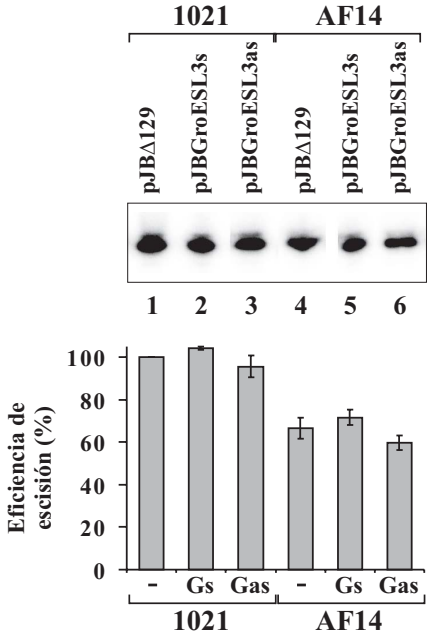


Figura R2.8: **Complementación del mutante AF14 (*groESL3*-).** Ensayo de extensión a partir de cebador realizado con un plásmido donador de la forma derivada de RmInt1 (pKGEMA4) y un plásmido que contiene el operón *groESL3* en ambas orientaciones, sentido (pJBGroESL3s) y antisentido (pJBGroESL3as), en 1021 y AF14. Como control negativo se introdujo en estas cepas el mismo plásmido que porta el operón *groESL3* pero sin éste (pJBΔ129). El gráfico de barras muestra la eficiencia de escisión del intrón (con su correspondiente error estándar) en los distintos fondos genéticos, tomando como referencia la escisión de RmInt1 en 1021 sin complementar.

la escisión de RmInt1 desde los distintos plásmidos donadores de intrón disponibles en el laboratorio tanto en 1021 como en AF14 (figura R2.9). El porcentaje de escisión de RmInt1 en cada muestra se determinó en relación a la eficiencia observada en la cepa silvestre con pKGEMA4 (donde consideramos una eficiencia del 100 %). Hay que destacar que la eficiencia de escisión del intrón desde los plásmidos pGm4 y pCm4 disminuyó al 40 % y al 52 % respectivamente en 1021 (carriles 3 y 5 respectivamente en la figura R2.9). En el mutante AF14 se observó una mayor disminución en la cantidad de producto escindido desde los tres plásmidos respecto a 1021 con pKGEMA4 (66 %, 24 % y 7 % desde pKGEMA4, pGm4 y pCm4 respectivamente; carriles 2, 4 y 6 respectivamente en la figura R2.9). El porcentaje de escisión de RmInt1 en AF14, calculado respecto a 1021 cuando se expresa desde el mismo plásmido dondador, corroboró la diferencia observada pre-

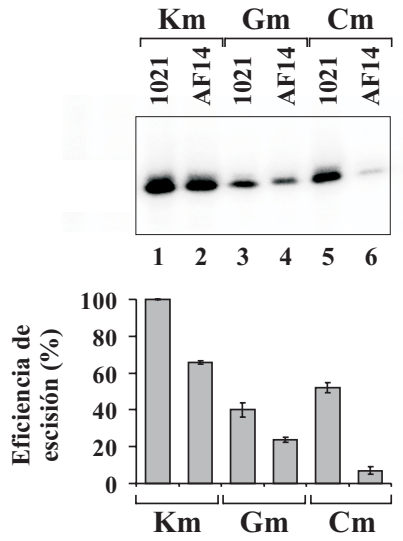


Figura R2.9: **Eficiencia de escisión de RmInt1 desde diferentes plásmidos donadores de intrón en 1021 y AF14.** Ensayo de extensión a partir de cebador realizado con tres plásmidos donadores de la forma derivada de RmInt1 que se diferencian por el gen de resistencia que portan: Km, pKGEMA4; Gm, pGm4; y Cm, pCm4. El gráfico de barras muestra la eficiencia de escisión del intrón (con su correspondiente error estándar) en todas las muestras respecto a 1021 con pKGEMA4.

viamente: desde pKGEMA4 mostró una eficiencia de escisión del 66 % y desde pCm4 un 13 %.

R2.5 DISCUSIÓN

La búsqueda de todas las RTs anotadas en el genoma de GR4 ha revelado la presencia de diversos intrones del grupo II. Así, aparecen diez copias de RmInt1, un intrón del grupo II perteneciente a la clase D (Toro, 2003) descubierto originalmente en el plásmido pRmeGR4b de *S. meliloti* GR4 y denominado Sr.me.I1 en la base de datos para los intrones del grupo II bacterianos (Candales *et al*, 2011). En esta cepa hemos encontrado, a su vez, siete copias de otro intrón del grupo II al que hemos nombrado RmInt2 y que también pertenece a la clase D (intrón denominado D171 en Toro & Martínez-Abarca, 2013). Dos de las RTs de copia única descubiertas en GR4 están relacionadas con otras presentes en 1021 (Toro *et al*, 2003): C770_GR4Chr2376 es un 80 % idéntica a SMa1875, una RT de la clase G1 (Toro & Martínez-Abarca, 2013) denominada Sr.me.I2 en la base de datos para los intrones del grupo II bacterianos (Candales *et al*, 2011), y C770_GR4pD0745 presenta un 98 % de identidad con SM_b21167, una RT perteneciente a la clase C (Toro, 2003). La RT de la clase E anotada como pseudogen en GR4 (C770_GR4pB077) fue descubierta por Toro *et al* (2002) en el plásmido pRmeGR4b. Tras la corrección de los cambios en la pauta de lectura, esta RT tiene una longitud de 469 aa y se relaciona con intrones presentes en *Bradyrhizobium*.

RmInt1 fue el primer intrón del grupo II descrito en la familia Rhizobiaceae (Martínez-Abarca *et al*, 1998). Mediante búsqueda en las bases de datos, hemos comprobado que la presencia de este intrón entre rizobios es muy extendida, y no sólo entre las bacterias que han sido secuenciadas (tabla R2.2), sino que trabajos previos han puesto de manifiesto que intrones tipo RmInt1 se encuentran en algunas colecciones en el 90 % de los aislados de nódulos de alfalfa (Muñoz *et al*, 2001; Biondi *et al*, 2011). No todos los intrones anotados en las bases de datos están completos, sino que encontramos copias parciales, la gran mayoría truncadas en su extremo 3′ (tabla R2.2). Esto

es indicativo de que, como ya se había observado previamente, RmInt1 tiende a evolucionar hacia formas inactivas mediante fragmentación, con pérdida de la zona 3' incluyendo la proteína (Fernández-López *et al*, 2005). Con las copias completas inferimos la relación filogenética que presentan los intrones encontrados en distintas especies y cepas bacterianas (figura R2.1). Se diferencian tres clados, de los que hemos tomado como representativos para estudios funcionales posteriores los intrones RmInt1, RmInt2 y SmedInt1 (denominado Sr.md.I1 en la base de datos para los intrones del grupo II bacterianos, Candales *et al*, 2011). Además, la elevada congruencia entre los árboles generados con la ribozima y la proteína (95 %) corrobora, para este subgrupo en particular, la coevolución entre el RNA del intrón y la proteína que codifican ya descrita previamente para los intrones del grupo II de manera más generalizada (Toor *et al*, 2001; Simon *et al*, 2009).

Un análisis de la distribución que presentan los tres intrones estudiados en este trabajo y las ISs que invaden pone de manifiesto que RmInt1 es el más disperso por la especie *S. meliloti*, donde IS*Rm2011-2* está ampliamente distribuida (tabla R2.10). En esta especie bacteriana encontramos copias de IS*Rm10-1* (tabla R2.11), una IS relacionada con IS*Rm2011-2* y que también puede ser invadida por RmInt1 (Martínez-Abarca & Toro, 2000). La expansión diferencial que presentan las ISs analizadas puede tener su explicación en la longitud del DNA diana que invaden. IS*Rm2011-2* es la que muestra mayor número de copias y, al igual que IS*Rm10-1*, reconoce una diana de 2 nucleótidos. IS*Rm17* se inserta en una diana compuesta por 6-7 nucleótidos, e IS*Sme3*, la menos extendida, en una diana de 8 nucleótidos (Siguier *et al*, 2006b). Esta expansión de las ISs a las que se asocian los intrones es necesaria para la dispersión de éstos, sin embargo, no parece ser suficiente para que se mantengan en la población. Los intrones son elementos deletéreos para las bacterias, las cuales son eliminadas de las poblaciones por la selección purificadora cuando los portan. No obstante, hay evidencias de que algunos intrones del grupo II no unen los exones cuando se escinden, impidiendo así la proliferación de la IS en la que se encuentran insertados y minimizando los daños al hospedador (Dai & Zimmerly, 2002; Chillón *et al*, 2011).

Tabla R2.10: Porcentaje de ocupación (O) de los tres intrones en las cepas totalmente secuenciadas donde se han encontrado copias de las ISs que invaden. Se muestra el número de copias completas de cada intrón (I) presentes en las distintas cepas. Como diana potencial para ser invadida (dianas disponibles, D) se han considerado todas las copias de IS que contuvieran la secuencia de reconocimiento para la inserción del intrón (en el caso de copias completas de IS*Rm17* se han contabilizado dos dianas). El guión (-) indica las cepas donde no se han encontrado copias de intrón ni de la IS con la que se asocia.

Cepa bacteriana	**IS*Rm2011-2***			**IR-IS*Rm17***			**IS*Sme3***		
	I	D	O (%)	I	D	O (%)	I	D	O (%)
S. meliloti 1021	3*	10	17	0	11	0	-	-	-
S. meliloti GR4	10*	4	69	7	5	58	0	1	0
S. meliloti SM11	6*	3	63	0	5	0	0	4	0
S. meliloti AK83	0	12	0	0	18	0	0	3	0
S. meliloti BL225C	5	4	56	0	15	0	-	-	-
S. meliloti Rm41	1*	4	0	0	6	0	0	3	0
S. medicae WSM419	0	2	0	0	3	0	4	1	80
S. fredii NGR234	-	-	-	0	5	0	-	-	-
R. leguminosarum bv. trifolii WSM2304	0	2	0	-	-	-	-	-	-
R. leguminosarum bv. viciae 3841	-	-	-	0	4	0	-	-	-
M. loti MAFF303099	-	-	-	0	1	0	-	-	-
O. anthropi ATCC 49188	-	-	-	-	-	-	0	3	0

* una copia del intrón RmInt1 se encuentra insertada en IS*Rm10-1*

La supervivencia y evolución de cualquier elemento móvil en bacterias está íntimamente ligado a su habilidad de dispersión a través de transferencia horizontal, como se ha sugerido para intrones del grupo II e ISs (Cerveau *et al*, 2011). La dinámica que siguen los elementos móviles se basa en esa transferencia horizontal, con la que comienzan el ciclo característico de extinción-recolonización dentro de una población bacteriana. La ausencia de genomas saturados de intrones (tabla R2.10) sugiere que éstos siguen un modelo de dinámica de extinción por selección (Sel-DE), en el que los genomas altamente colonizados son eliminados a través de la selección purificadora (Leclercq & Cordaux, 2012). La amplia distribución de las ISs diana de RmInt2 y SmedInt1, y la ausencia de ISs invadidas (tabla R2.10), indi-

Tabla R2.11: Resultado del BLASTn realizado con la secuencia de IS*Rm10-1* a fecha 22-01-13. Entre paréntesis se indica el número de copias invadidas por el intrón RmInt1 (I) y la equivalencia de los plásmidos de las cepas de *S. meliloti* a pSymA (A) y pSymB (B) de 1021.

Cepa bacteriana	**Localización genómica**	**Porcentaje cobertura**	**Porcentaje identidad**	**Nº copias intrón**	**Longitd secuencia**
S. meliloti 1021	pSymB	100	100	1 (I)	1-1.042
S. meliloti GR4	pRmeGR4c (A)	100	100	1 (I)	1-1.042
S. meliloti SM11	pSmeSM11c (A)	100	100	2 (I)	1-1.042
S. meliloti SM11	Cromosoma	100	99	1	1-1.042
S. meliloti BL225C	pSINMEB01 (A)	100	99	1	1-1.042
S. meliloti AK83	Cromosoma 3 (A)	100	99	2	1-1.042
S. meliloti Rm41	pSYMB (B)	100	99	1 (I)	1-1.042
R. etli 8C-3	REB01	100	86	1	1-1.042
R. etli CFN42	p42d	100	86	1	1-1.042
R. etli CIAT652	pB	100	86	2	1-1.042
R. etli CFN42	p42d	52	88	1*	1-538
R. etli CIAT652	pB	52	88	1*	1-538

* tras el fragmento de IS se encuentra el intrón RmInt1

ca que estos intrones pueden haberse adquirido recientemente (encontrándose en un estado de recolonización del género *Sinorhizobium*) o pueden estar en la fase de extinción del ciclo (figura R2.10). RmInt1, por el contrario, presenta una gran dispersión por las cepas de *S. meliloti*, indicando que está en un estado de proliferación dentro del ciclo de extinción-recolonización. Además, en varias cepas se encuentran copias fragmentadas de este intrón (tabla R2.2), lo cual sugiere que RmInt1 estaría en un nuevo estado dentro del modelo Sel-DE propuesto por Leclercq & Cordaux (2012). Sería un estado intermedio, en el que hay genomas sólo con copias completas o fragmentadas del intrón, y otros que contienen al mismo tiempo copias completas y truncadas (figura R2.10).

Estudios previos han puesto de manifiesto la movilidad de RmInt1 en distintos hospedadores, inicialmente dentro de la especie *S. meliloti* (Martínez-Abarca & Toro, 2000) y algunos aislados de *S. medicae* (Martínez-Abarca *et al*,

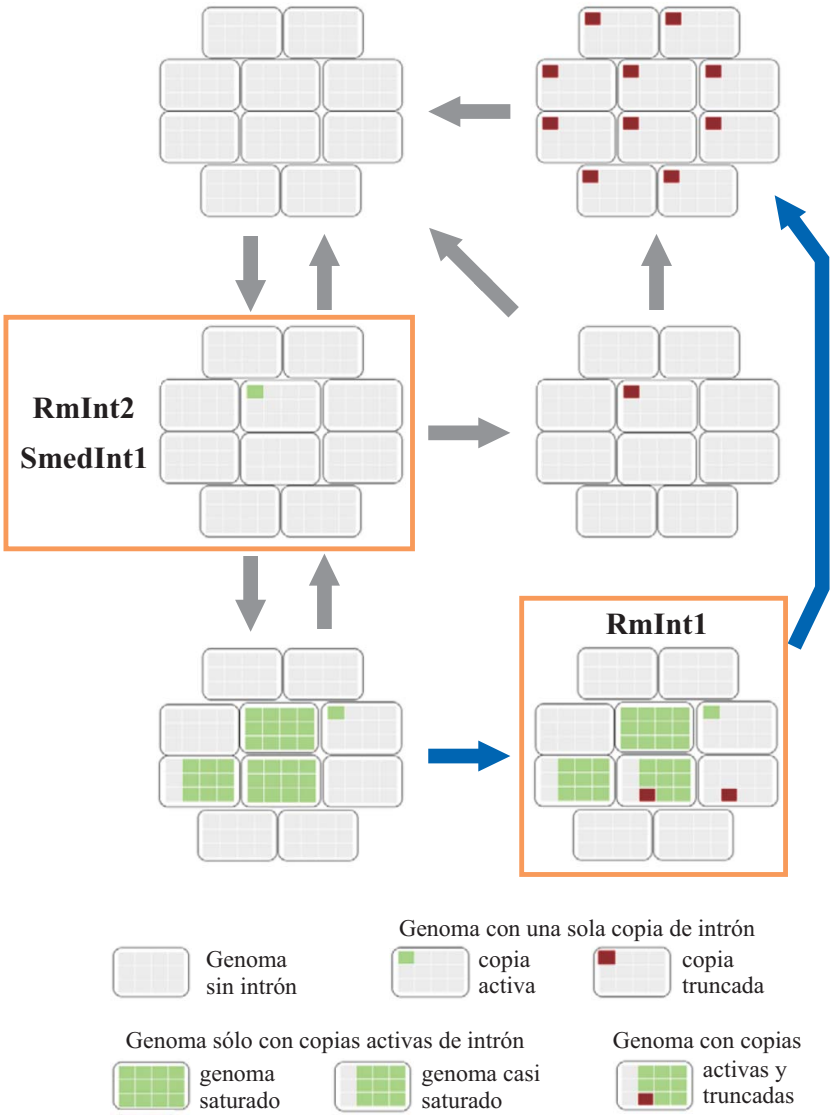


Figura R2.10: **Modelo de la dinámica de extinción-recolonización de los intrones del grupo II mediante selección purificadora (Sel-DE).** Enmarcados se señalan los estados del ciclo en el que se encuentran RmInt1, RmInt2 y SmedInt1. Las flechas azules indican las transiciones entre el estado actual, pasado y futuro de RmInt1. Adaptado de Leclercq y Cordaux, 2012.

2000), y posteriormente en diferentes cepas de la familia Rhizobiaceae tales como *S. terangae*, *Rhizobium tropici*, *R. leguminosarum bv. phaseoli*, *Agrobacterium rhizogenes* y *A. tumefaciens* (Fernández-López *et al*, 2005). Recientemente se ha demostrado la movilidad de este intrón en especies más alejadas filogenéticamente (γ-proteobacterias) como *Pseudomonas putida* (Nisa Martínez, 2011) y *E. coli* (García-Rodríguez *et al*, 2011). Sin embargo, aunque se haya probado su movilidad, estos estudios no se han hecho siempre desde las mismas construcciones donadoras de intrón y de un manera comparativa. En este trabajo hemos evaluado, bajo las mismas condiciones de ensayo, la movilidad de RmInt1 en tres especies estrechamente relacionadas y que incluyen cepas con genomas secuenciados: 1021 (Galibert *et al*, 2001) dentro de la especie *S. meliloti*, WSM419 (Reeve *et al*, 2010) dentro de *S. medicae*, y CFN42 (González *et al*, 2006) y CIAT652 (González *et al*, 2010) dentro de *R. etli*. RmInt1 es capaz de moverse en todos estos fondos genéticos, aunque presenta una eficiencia distinta dependiendo de la cepa y la orientación de la diana (LAG o LEAD; figura R2.4). Como era de esperar, las mayores eficiencias se observaron hacia la diana LAG, puesto que este intrón, para su invasión, usa predominantemente la hebra que sirve de molde para la cadena retardada (LAG) en la horquilla de replicación (Martínez-Abarca *et al*, 2004). A su vez, los resultados presentados en este trabajo corroboran los datos previamente publicados sobre 1021, donde la eficiencia de movilidad de RmInt1 está claramente disminuida (Toro *et al*, 2003). Además, estos datos nos pueden servir como referencia para encontrar, mediante genómica comparada, genes comúnes a todas las cepas analizadas que pudieran estar implicados en la movilidad de RmInt1.

Para determinar de manera más precisa los factores del hospedador que facilitan o bloquean la movilidad de este tipo de elementos, y cómo lo hacen, se necesitan llevar a cabo más análisis funcionales. Estudios proteómicos llevados a cabo en el grupo pusieron de manifiesto que uno de los genes *groEL* era una de las proteínas predominantes asociada a fracciones de RNPs de RmInt1 (Villadas et al., sin publicar). El conjunto de mutantes para los genes *groEL* generados en la cepa 1021 (tabla R2.9; Bittner *et al*,

2007) permitió analizar la posible implicación de estos genes en la escisión de RmInt1. Inicialmente, analizamos la escisión de RmInt1 en esa colección de mutantes y en la cepa silvestre desde el plásmido pCm4. Este ensayo puso de manifiesto que el efecto de los genes *groEL* es cuantitativo además de cualitativo, es decir, la escisión de RmInt1 requiere la expresión de más de un gen *groEL* puesto que no es capaz de escindirse en los cuádruples mutantes (figura R2.6). En todos los mutantes simples RmInt1 presenta una eficiencia de escisión superior al 49 %, salvo en el mutante AF14, donde muestra un 11 %, lo cual sugiere que el gen *groEL3* podría estar influyendo en su escisión. Además, el análisis filogenético entre las distintas copias de los genes *groEL* reveló que los genes *groEL3* formaban un clado diferencial (figura R2.7), lo que podría estar indicando una adaptación de este gen al correcto plegamiento de proteínas no esenciales para la bacteria como es la IEP de RmInt1. Sin embargo, estudios posteriores pusieron de manifiesto que no se reestablecen los niveles de escisión del silvestre cuando se complementa el mutante AF14 con el operón *groESL3* de GR4 en *trans* (figura R2.8).

El ensayo de complementación de los mutantes no se pudo llevar a cabo con el plásmido pCm4 debido a su pérdida en algunas muestras después de dos rondas de conjugación. No obstante, realizamos un ensayo de qRT-PCR (figura R2.11) para comprobar la estabilidad de dicho plásmido en las muestras utilizadas en el primer ensayo de extensión a partir de cebador (figura R2.6). Este análisis reveló diferencias en la cantidad de transcrito primario de la IEP que contenía cada muestra, pero esas diferencias no explican la ausencia de escisión de RmInt1 en los cuádruples mutantes ni su bajo nivel en AF14. Para los ensayos de complementación decidimos usar el operón *groESL3* de la cepa GR4 puesto que en ella se describió originalmente el intrón RmInt1 (Martínez-Abarca *et al*, 1998) y contiene 7 copias más de éste que la cepa 1021 (Toro *et al*, 2003). Además, los genes *groEL3* de ambas cepas presentan un 94 % de identidad (teniendo las proteínas codificadas por ellos un 96 % de identidad y un 98 % de similitud), lo que hace pensar que deben ser equivalentes. Sin embargo, no hemos analizado la

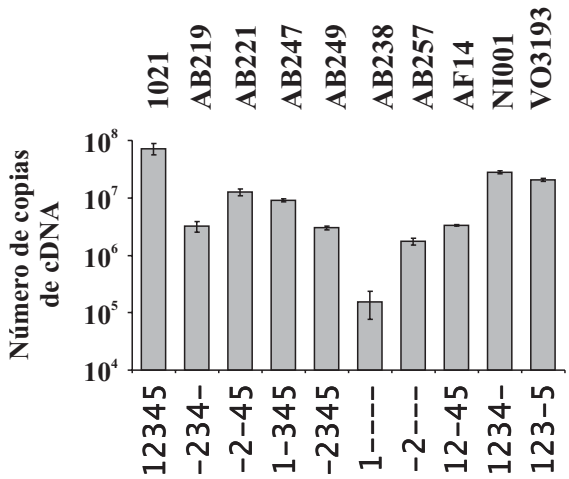


Figura R2.11: **RNA total extraído de células de *S. meliloti* 1021 y los mutantes groEL con el plásmido pCm4.** Los niveles de mRNA del intrón se determinaron mediante una cuantificación absoluta. Cada medida es la media de al menos tres muestras independientes, con su correspondiente error estándar. El guión (-) indica los genes *groEL* mutados en la cepa.

complementación del mutante AF14 con el operón *groESL3* de 1021, por lo que no podemos descartar que la falta de complementación esté causada por el uso del operón *groESL3* de GR4 en el fondo genético de 1021.

Todos estos datos sugieren que el gen *groEL3* no es imprescindible para la escisión de RmInt1, y puede que la disminución de ésta observada en el mutante AF14 se deba a otra variación genómica ocurrida durante el proceso de obtención de dicho mutante. Para determinar la alteración genómica de AF14 que afecta a la escisión de RmInt1 se podría llevar a cabo una secuenciación del genoma de este mutante y comparar su secuencia con la cepa de referencia 1021. Así se descubriría una de las regiones génicas implicadas en dicho proceso y se podrían realizar estudios de complementación con esa zona para corroborar su influencia en la escisión de RmInt1.

INFORMACIÓN SUPLEMENTARIA

Tabla R2.S1: Intrones utilizados en la construcción de los árboles filogenéticos (figura R2.1).

Nombre del intrón	**Replicón**	**Coordenadas del intrón completo**[a]	**Protein_ID**[b]
1021	pSymB	1.576.500-1.574.617	1.575.954-1.574.695
102F51	pSymB-B152	5.240-3.357	4.694-3.435
5D19	-	AY248839.1	AAP83798.1
BL225C	Cromosoma	667.127-669.010	AEG03566.1
E. coli	-	X77508	CAA54637.1
GR4-R1	-	Y11597.2	CAA72334.3
GR4-R2	pRmeGR4c	132.907-134.793	YP_007192462.1
NGR234	pNGR234b	796.157-801.848	800.511-801.770
ORS22	-	AY608908.1	AAU95643.1
Rm41	pSYMB	1.129.686-1.127.803	YP_006843383.1
RMO09-R1[c]	-	-	-
RMO09-S1[c]	-	-	-
SM11	Cromosoma	3.677.924-3.679.807	YP_005722110.1
WSM419	Cromosoma	1.305.475-1.307.359	ABR60073.1

[a] si la secuencia del intrón está publicada se indica su número de acceso
[b] si las proteínas no tienen asignado un número de acceso se indican sus coordenadas
[c] secuencias obtenidas en este trabajo mediante ensayos de movilidad (ver apartado R2.3.2)

CAPÍTULO 3

CARACTERIZACIÓN DE DOS NUEVOS INTRONES DEL GRUPO II TIPO RmInt1 Y DEL DNA DIANA CON EL QUE SE ENCUENTRAN ASOCIADOS.

Como hemos estudiado en el capítulo anterior, se encuentran intrones del grupo II relacionados con RmInt1 en muchas cepas dentro de la familia Rhizobiaceae. En especial, están ampliamente distribuidos en el género *Sinorhizobium*, donde se describió originalmente el intrón RmInt1 (en el plásmido pRmeGR4b de la cepa *S. meliloti* GR4; Martínez-Abarca *et al*, 1998). En el capítulo anterior distinguimos tres intrones que fueron considerados como representativos de los tres clados observados en los árboles filogenéticos (figura R2.1). En este capítulo llevaremos a cabo un análisis detallado de los nuevos intrones RmInt2 y SmedInt1 (Sr.md.I1, Candales *et al*, 2011), así como su caracterización funcional.

R3.1 CONTEXTO GÉNICO DE LOS DIFERENTES INTRONES

Los tres intrones objeto de estudio en este capítulo (RmInt1, RmInt2 y SmedInt1) están relacionados filogenéticamente, sin embargo, difieren en su DNA diana, siendo en todos los casos una secuencia de inserción (IS). En la figura R3.1 se muestra una representación a escala de estos intrones y sus regiones flanqueantes.

El intrón RmInt1 se encuentra asociado a IS*Rm2011-2* (Martínez-Abarca *et al*, 1998), una secuencia de inserción que pertenece a la familia IS*630*-Tc1/IS*3*. Dicha IS se caracteriza por presentar repeticiones invertidas terminales imperfectas de 19 nt, así como por producir una transposasa a partir

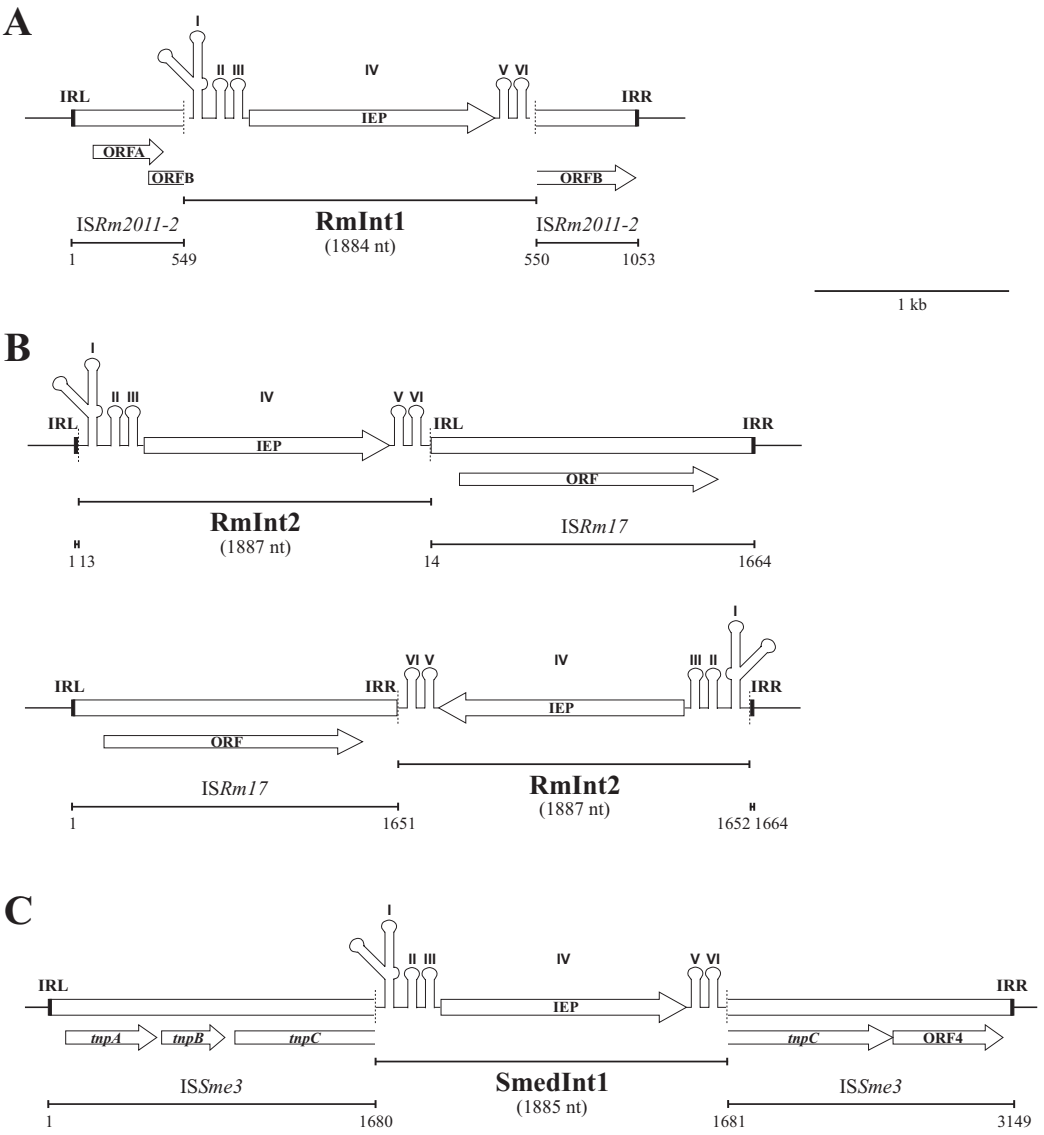


Figura R3.1: **Contexto genético de los tres intrones del grupo II objeto de estudio.** Representación a escala de RmInt1 (A), RmInt2 (B) y SmedInt1 (C), y sus regiones flanqueantes. Se muestran los seis dominios que forman este tipo de intrones y la proteína codificada por ellos presente en el DIV. También se indican los ORFs que componen la IS asociada a cada uno de los intrones, así como las repeticiones terminales invertidas izquierda (IRL) y derecha (IRR) de estas ISs. Las flechas señalan la orientación de los ORFs.

de dos ORFs solapados gracias a un cambio en el marco de lectura debido a un deslizamiento ribosomal entre el ORFA y el ORFB de la IS (Selbitschka *et al*, 1995). La inserción de RmInt1 ocurre entre los nucleótidos 549 y 550 de la IS, interrumpiendo la traducción de la transposasa en su ORFB debido a la presencia de un codón de parada en el nucleótido 33 del intrón (Martínez-Abarca *et al*, 1998).

El intrón RmInt2 tiene como diana de inserción la repetición invertida (IR) de IS*Rm17*. Las repeticiones invertidas de dicha IS constan de 16 pb y difieren en el nucleótido 7, siendo en la IRL una A y en la IRR una G. RmInt2 es capaz de invadir ambas IRs, insertándose entre las posiciones 13 y 14 de la IS cuando invade la IRL o entre los nucleótidos 1.651 y 1.652 si la IR invadida es la derecha (IRR). Con esta estrategia RmInt2 se asegura dos

Tabla R3.1: Secuencia de las repeticiones directas de IS*Sme3* en distintas localizaciones (correspondientes a copias completas de dicha IS).

Cepa bacteriana	**Localización**	**DR**
S. medicae WSM419	Cromosoma-1	GACTTGCC
	pSMED02-1	ATCCAGTT
	pSMED02-2	GTACAGCG
	pSMED03-1	ATTCTTCG
S. meliloti GR4	pRmeGR4b-1	ATTCAATT
S. meliloti SM11	pSmeSM11c-1	CCTCGAAT
	pSmeSM11c-2	GAGGTATG
	pSmeSM11d-1	CATTGATC
S. meliloti AK83	Cromosoma 2-1	GCCTTCTG
	Cromosoma 3-1	AGTCAACG
	Cromosoma 3-2	GCTATCTT
S. meliloti Rm41	Cromosoma-1	GTTTGATC
	Cromosoma-2	GTTCGGTT
	pSYMA-1	GGTCATAG
S. meliloti CO17	pHRC017	GATTGTCG

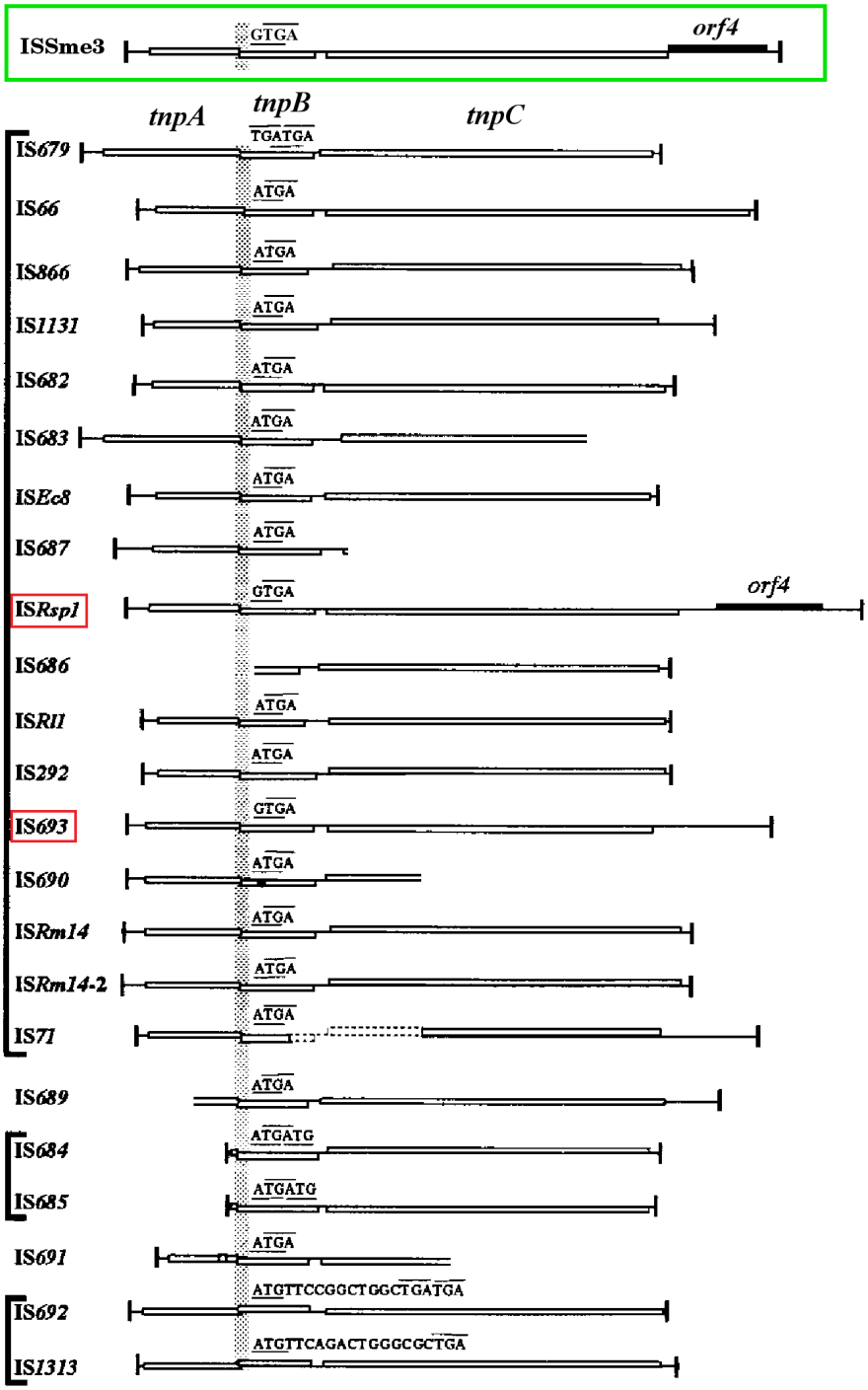
sitios de inserción en cada copia de IS*Rm17*. Otras características de esta IS son que está flanqueada por repeticiones directas (DR) de 6-7 nt, contiene un único ORF que produce una transposasa de 444 aa, y es un elemento no clasificado, es decir, no se incluye en ninguna familia de ISs (Siguier *et al*, 2006b).

Por otro lado, SmedInt1 se encuentra asociado a una IS que no está presente en las bases de datos y que se enmarca dentro de la familia IS*66*. El contexto génico de SmedInt1 en su genoma de referencia *S. medicae* WSM419 muestra la repetición de un casete de genes de unas 3 kb. Gracias a esta anotación hemos descrito una nueva IS de 3.149 pb de longitud, denominada IS*Sme3*, que contiene cuatro ORFs correspondientes a los cuatro genes del casete repetido que flanquean al intrón. Esta IS*Sme3* presenta repeticiones invertidas imperfectas de 22 nt, que difieren en las posiciones 9, 14 y 17, y repeticiones directas de 8 nt en sus dos extremos. Estas DRs son diferentes en cada una de las localizaciones de dicha IS tanto dentro del genoma de WSM419 como en otros rizobios donde se encuentran copias completas de IS*Sme3* (tabla R3.1). El primer ORF, *tnpA*, codifica una proteína de 156 aa anotada en WSM419 como una transposasa de la familia IS*3*/IS*911*; la pro-

---

Figura R3.2: **Estructura de IS*Sme3* y varios miembros de la familia IS*66*.** Representación a escala de la nueva IS*Sme3* descrita en este trabajo (dentro del recuadro verde) y los elementos pertenecientes a la familia IS*66* descritos hasta 2001. Las líneas verticales presentes en los extremos de cada elemento indican las IRs. Los rectángulos blancos muestran los tres ORFs característicos de esta familia (*tnpA*, *tnpB* y *tnpC*). *tnpA* se encuentra en el marco de lectura 0, y se dispone en el centro de la línea que representa la secuencia de la IS; el marco de lectura -1 y +1 se representa debajo y encima de esta línea respectivamente. La región vertical sombreada que cubre a todos los miembros indica el solapamiento entre *tnpA* y *tnpB*. El codón de terminación de *tnpA* en este solapamiento tiene una línea encima, y el codón de iniciación de *tnpB* se encuentra subrayado. Un rectángulo negro indica el ORF adicional aguas abajo de *tnpC* presente en IS*Rsp1* e IS*Sme3*. Los recuadros rojos señalan las dos ISs más relacionadas con IS*Sme3*, IS*Rsp1* estructuralmente e IS*693* respecto a secuencia. Adaptado de Han *et al*, 2001.

(Página siguiente) ►


ISSme3
GTGA
orf4
tnpA
tnpB
tnpC
IS679
TGATGA
IS66
ATGA
IS866
ATGA
IS1131
ATGA
IS682
ATGA
IS683
ATGA
ISEc8
ATGA
IS687
ATGA
ISRsp1
GTGA
orf4
IS686
ISRl1
ATGA
IS292
ATGA
IS693
GTGA
IS690
ATGA
ISRm14
ATGA
ISRm14-2
ATGA
IS71
ATGA
IS689
ATGA
IS684
ATGATG
IS685
ATGATG
IS691
ATGA
IS692
ATGTTCCGGCTGGCTGATGA
IS1313
ATGTTCAGACTGGGCGCTGA

teína codificada por *tnpB* tiene una longitud de 118 aa y está anotada como ORF2 de IS*66*; *tnpC* codifica una proteína de 524 aa anotada como ORF3 de IS*66*; y la proteína de 187 aa codificada por el ORF4 se encuentra anotada como una proteína yecA, de la cual se desconoce su función. Cuando SmedInt1 invade su diana, interrumpe *tnpC* entre sus nucleótidos 727 y 728 (correspondiente a las posiciones 1.680-1.681 de la IS completa).

Analizando los elementos que se encuadran dentro de la familia IS*66* podemos decir que, estructuralmente, la nueva IS descrita en este trabajo se asemeja a IS*Rsp1*, puesto que ambas presentan cuatro ORFs que además tienen una longitud similar (figura R3.2). Sin embargo, la IS a la que más se asemeja en cuanto a secuencia es IS*693*. Frente a ésta presenta un 63 % de identidad a nivel de nucleótido, y ambas superan las 3 kb de longitud. A nivel de proteína, *tnpB* es el ORF más conservado, con un 78 % de identidad y un 88 % de similitud, seguido de *tnpC* (67 % de identidad y 81 % de similitud) y *tnpA* (55 % de identidad y 73 % de similitud). El ORF4 no presenta una identidad superior al 20 % con ningún otro ORF, y ha sido considerado por la base de datos ISfinder como un gen pasajero.

R3.2 LOCALIZACIÓN DE LOS INTRONES EN EL GENOMA DE SUS CEPAS DE ORIGEN

Tras analizar detalladamente las regiones que flanquean cada uno de los intrones estudiados, haremos una descripción de la localización de las copias genómicas de intrón, así como de las ISs asociadas, presentes en las cepas donde se encontraron originalmente. RmInt1 presenta una tendencia de invasión a dianas localizadas en la secuencia que sirve como molde de la cadena retrasada en la horquilla de replicación (Martínez-Abarca *et al*, 2004). Por ello, es relevante diferenciar cada diana dependiendo de su posición dentro del replicón, pudiendo localizarse en la hebra usada como molde para la síntesis de la cadena retrasada (orientación LAG) o líder (orientación LEAD) en la horquilla de replicación.

Como se mencionó en el capítulo 1, la cepa bacteriana *S. meliloti* GR4 presenta diez copias del intrón RmInt1 con un 99 % de identidad, nueve de las cuales están insertadas en IS*Rm2011-2* y la restante en la única copia de IS*Rm10-1* que encontramos en esta cepa, situada en el plásmido pRmeGR4c (figura R3.3A). En este genoma, las copias de RmInt1 están distribuidas: cuatro en el cromosoma, todas en orientación LAG, una en pRmeGR4b en LEAD y cinco en pRmeGR4c, tres de las cuales están en orientación LAG y dos en LEAD. Además de estas copias completas, hay una copia parcial (fragmento de 653 nt que tiene un 88 % de identidad con la región de RmInt1 comprendida entre los nucleótidos 2 y 650) en el plásmido pRmeGR4b en LAG. A pesar del alto número de copias de este intrón presentes en GR4, encontramos cuatro dianas (IS*Rm2011-2*) adicionales en LEAD no invadidas, tres en el cromosoma y una en pRmeGR4b.

Con respecto a RmInt2, *S. meliloti* GR4 contiene siete copias 100 % idénticas, todas ellas insertadas en la IR de IS*Rm17*; sin embargo, una de las copias (marcada con un asterisco en la figura R3.3A) carece de la transposasa, encontrándose la IR de IS*Rm17* dentro de la región intergénica entre *nifE* y *nifX*. En cuanto a la distribución de las copias de este intrón en el genoma de GR4, encontramos cuatro de ellas en pRmeGR4c, todas en orientación LAG, y tres en pRmeGR4d, de las cuales dos están en LAG y una en LEAD. No existen más copias de IS*Rm17* en el genoma de este rizobio, pero puesto que toda IS*Rm17* tiene dos IRs, todavía quedan cuatro dianas de RmInt2 sin invadir, tres en pRmeGR4c y una en pRmeGR4d, todas en LEAD. Además, delante de la segunda copia de IS*Rm17* presente en pRmeGR4c encontramos otra copia de dicha IS pero truncada (fragmento que contiene desde el nucleótido 1 hasta el 939 de IS*Rm17*). Por tanto, existe una IRL adicional no invadida en orientación LEAD.

El genoma de *S. medicae* WSM419 contiene cuatro copias 100 % idénticas de SmedInt1, tres de ellas insertadas en IS*Sme3* y la restante (marcada con un asterisco en la figura R3.3B) interrumpiendo un gen que codifica una proteína de función desconocida (gen Smed_4905). La distribución genómica de las copias de SmedInt1 es: una copia en el cromosoma en orientación

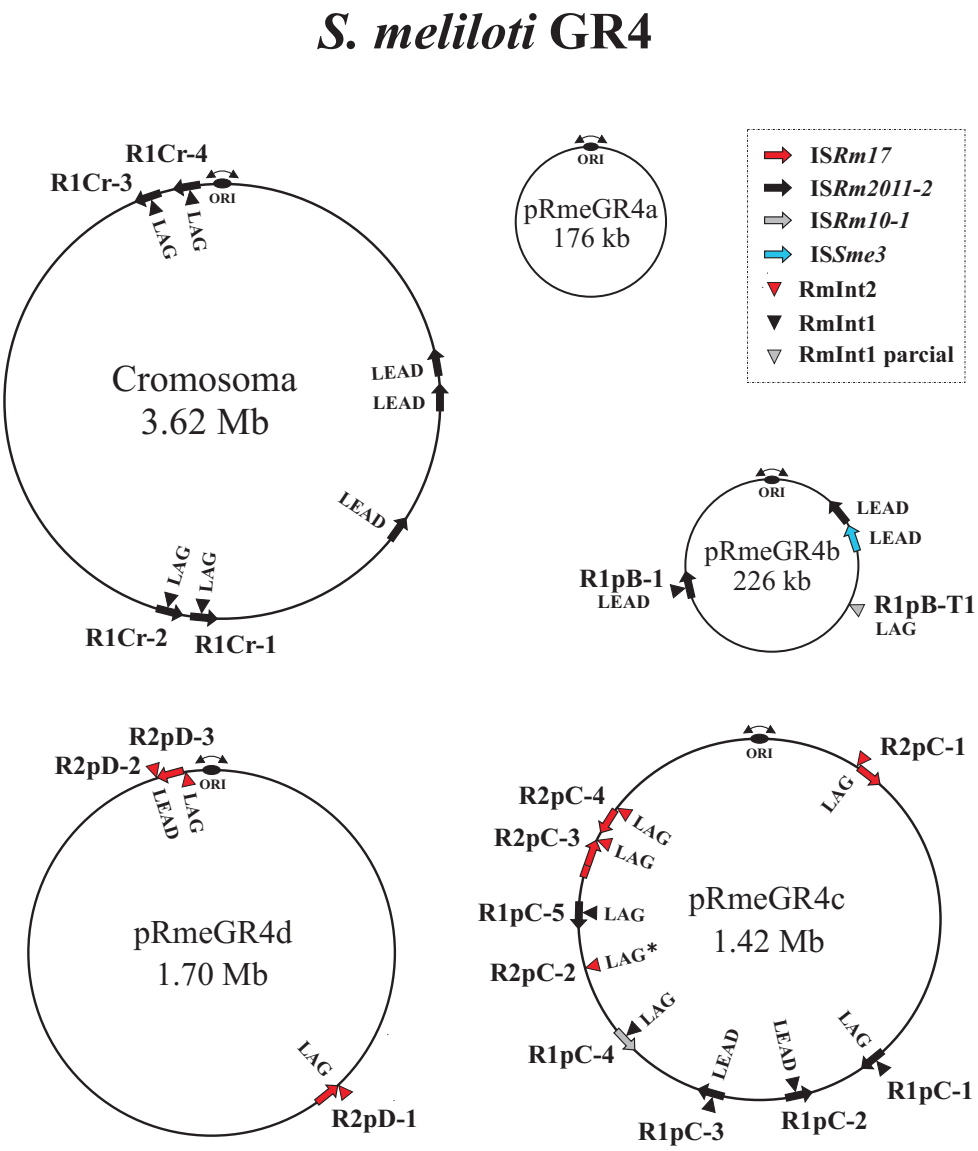
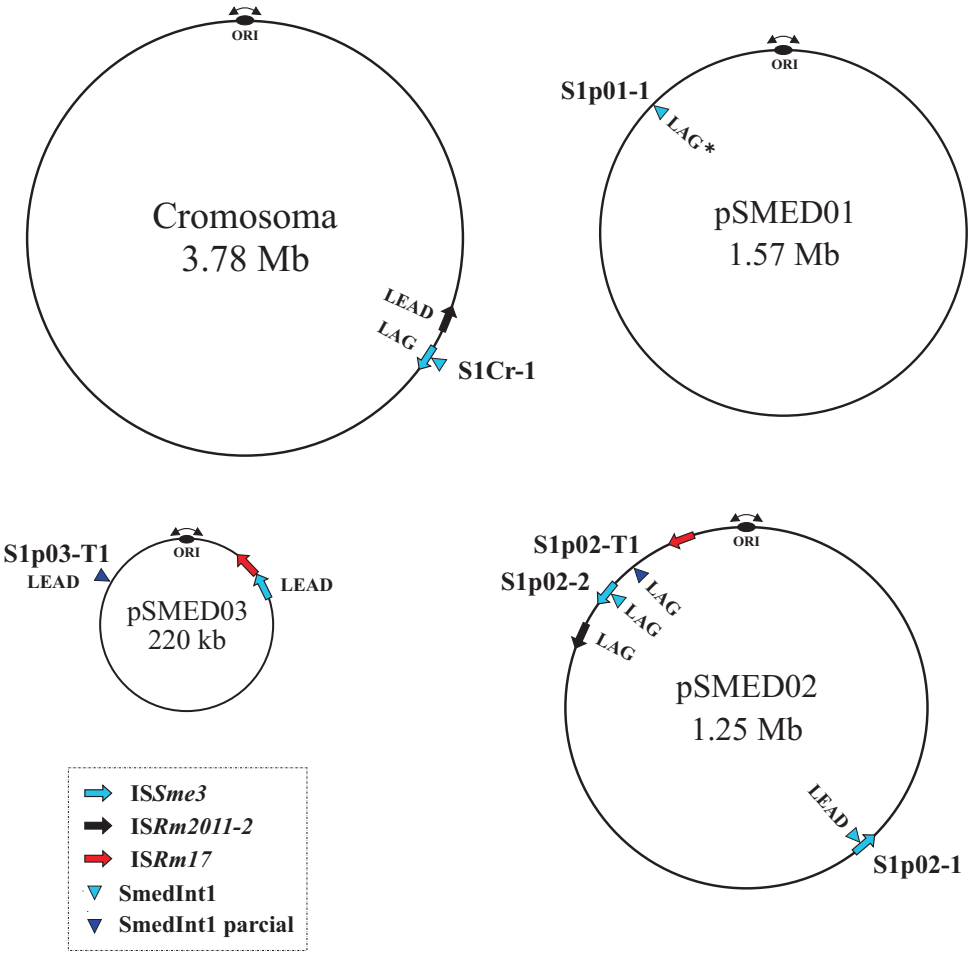


Figura R3.3: **Distribución genómica de los tres intrones y dianas estudiados en sus cepas de origen.** Los diagramas (no representados a escala) muestran la localización y orientación de estos elementos móviles respecto al origen de replicación en cada uno de los replicones de las cepas *S. meliloti* GR4 (A) y *S. medicae* WSM419 (B). RmInt1 e IS*Rm2011-2* aparecen en negro, RmInt2 e IS*Rm17* en rojo, y SmedInt1 e IS*Sme3* ►

S. medicae WSM419



◄ en celeste. Los intrones que se encuentran en la cadena sentido o antisentido están representados fuera o dentro del círculo que forma el replicón respectivamente. Las dianas situadas en la hebra que sirve como molde para la síntesis de la hebra retardada o continua en la horquilla de replicación se indican como LAG o LEAD respectivamente. Marcadas con un asterisco se señalan una copia de RmInt2 y otra de SmedInt1, que están insertadas en la secuencia diana dentro de sus respectivas ISs, pero les falta la IS completa.

LAG, una en el plásmido pSMED01 en LAG y dos en pSMED02, de las cuales una copia está en LAG y otra en LEAD. También encontramos dos copias parciales de este intrón, una en pSMED02 en LAG (fragmento de 639 nt un 88 % idéntico a la región de SmedInt1 comprendida entre los nucleótidos 2 y 651) y otra en pSMED03 en LEAD, formada por dos pequeños fragmentos unidos y con identidad a distintas zonas del intrón (uno de 404 nt que tiene un 93 % de identidad con la zona situada entre las posiciones 49 y 455 de SmedInt1; y otro de 287 nt un 81 % idéntico a la región de dicho intrón situada entre los nucleótidos 1.068 y 1.354). Llama la atención que esta última copia parcial encontrada en pSMED03 se parece más a RmInt1 que a SmedInt1, conteniendo un primer fragmento de 417 nt con una identidad del 99 % con la zona comprendida entre las posiciones 37 y 455 de RmInt1, y un segundo fragmento de 287 nt un 98 % idéntico a la región de dicho intrón situada entre los nucleótidos 1.067 y 1.353. Además, esta cepa contiene otra copia de IS*Sme3* sin invadir en pSMED03 en LEAD.

Como hemos visto en el capítulo anterior, la presencia de las ISs en otros genomas no va siempre acompañada de sus respectivos intrones. Así, la cepa GR4 presenta una copia de IS*Sme3* en orientación LEAD en pRmeGR4b que no se encuentra invadida por el intrón SmedInt1. WSM419 contiene dos copias de IS*Rm2011-2* libres de RmInt1, una en el cromosoma en LEAD y otra en pSMED02 en LAG, y dos copias de IS*Rm17*, una en pSMED02 y otra en pSMED03, con ninguna de las IRs invadidas por RmInt2.

R3.3 CARACTERÍSTICAS ESTRUCTURALES DE LOS INTRONES DEL GRUPO II

RmInt1, al igual que el resto de intrones del grupo II, consiste en seis dominios estructurales entre los que aparecen interacciones terciarias. El dominio IV presenta un ORF que codifica una proteína conocida como IEP (proteína codificada por el intrón, del inglés *Intron Encoded Protein*), que contiene, a su vez, varios dominios: el dominio reverso transcriptasa (RT), el

dominio madurasa (X) y una región C-terminal sin dominio endonucleasa (Martínez-Abarca *et al*, 1998).

Analizando la estructura primaria de los tres intrones objeto de estudio observamos una serie de mutaciones, que aparecen tanto en zonas de bucle, que no afectan al plegamiento de la molécula, como en zonas de apareamiento, donde se observan cambios compensados que permiten la formación de la estructura secundaria. En la figura R3.4 se muestra la estructura secundaria predicha para RmInt1, indicándose, a su vez, las mutaciones ocurridas en SmedInt1 (en celeste) y RmInt2 (en rojo). Se observa que todos los cambios acontecidos en SmedInt1 son sustituciones salvo una inserción aguas abajo de la EBS2; sin embargo, RmInt2 presenta, además de gran cantidad de sustituciones, varias inserciones y deleciones a lo largo de su secuencia respecto a la de RmInt1. Esto puede llevar a un cambio en la longitud de los tallos y/o bucles, como ocurre por ejemplo en el dominio VI, donde las dos deleciones que encontramos en RmInt2 producen la formación de un solo tallo sin bucles y acabado en cuatro nucleótidos desapareados (figura R3.5). La mayor parte de las interacciones terciarias que aparecen entre los dominios no se ven afectadas por las mutaciones ocurridas en los nuevos intrones. Los cambios más significativos los encontramos en las EBSs: la EBS1 de RmInt2 presenta cuatro nucleótidos diferentes; la EBS2 ha cambiado en ambos intrones, siendo más importante la deleción encontrada en RmInt2, puesto que mueve la posición de dicha EBS dentro de la secuencia del intrón; y la EBS3 es distinta sólo en RmInt2. Estos cambios, así como los ocurridos en las IBSs, serán expuestos de manera más detallada el siguiente apartado.

Con respecto a la secuencia de aminoácidos de las proteínas codificadas por estos intrones (figura R3.6), se observa que, al igual que ocurre con la ribozima, la IEP con más variaciones es la codificada por RmInt2. Los cambios se concentran hacia el final de la proteína, mientras que el dominio RT es el más conservado. Varias regiones del dominio RT y gran parte del dominio madurasa presentan aminoácidos con propiedades muy distintas a los que aparecen en RmInt1, por tanto, la actividad de esos do-

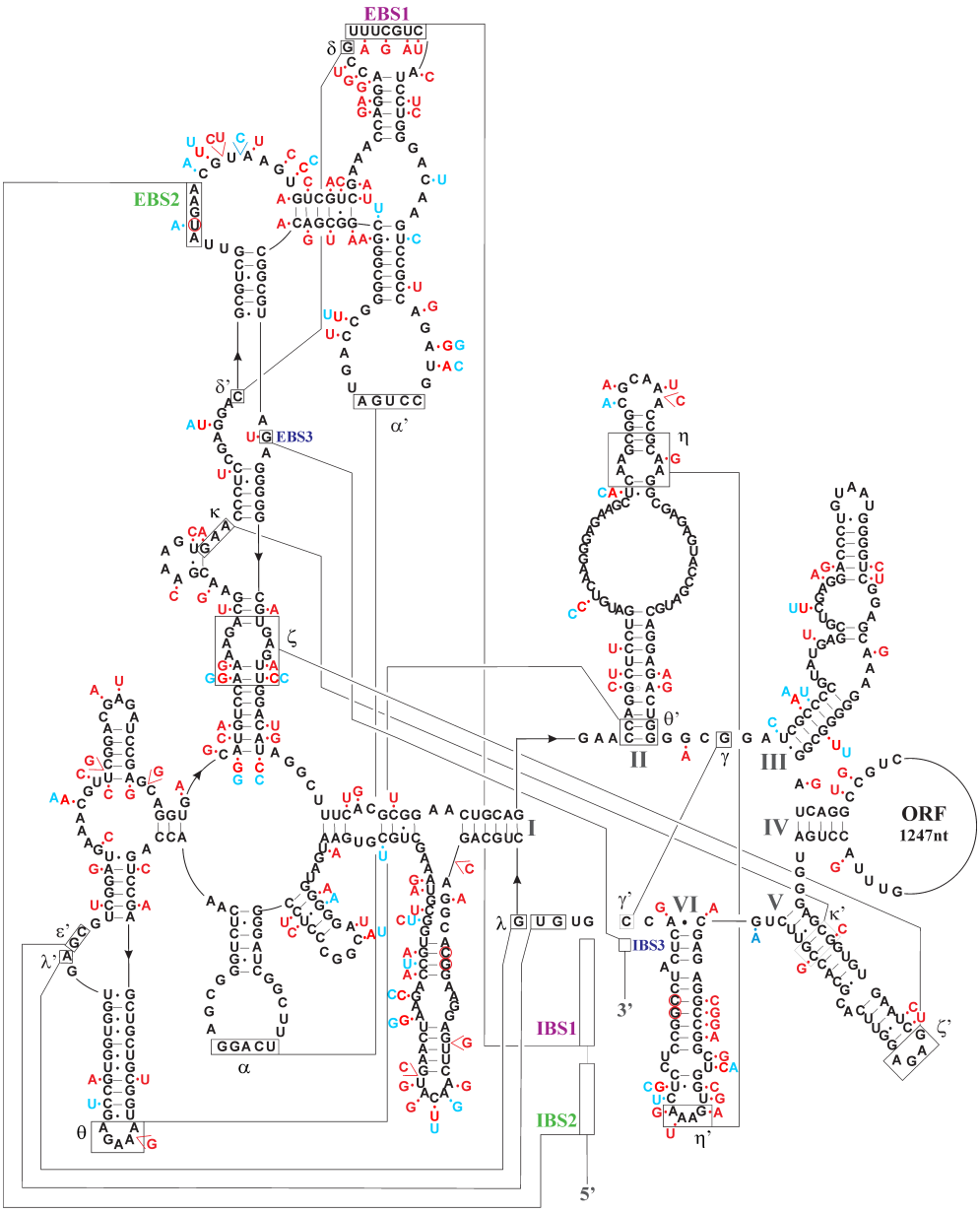


Figura R3.4: **Estructura secundaria de los intrones del grupo IIB.** Estructura propuesta para RmInt1 donde se señalan las mutaciones ocurridas en la secuencia de RmInt2 (en rojo) y SmedInt1 (en celeste). Este tipo de intrones están compuestos por seis dominios estructurales principales, que se indican con números romanos en gris. Las cajas, unidas por líneas, determinan las interacciones terciarias entre los dominios del intrón. En morado, verde y azul se representan las interacciones entre las IBSs-EBSs 1, 2 y 3 respectivamente.

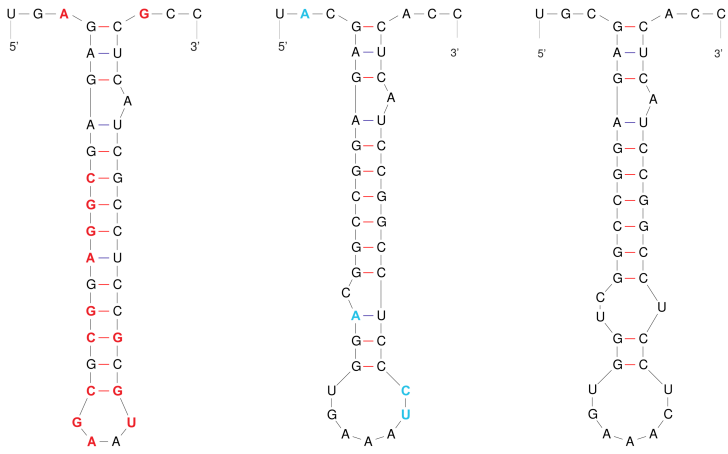


Figura R3.5: **Comparación de la estructura secundaria predicha del DVI de los tres intrones estudiados.** Se indican los cambios ocurridos en RmInt2 (en rojo) y SmedInt1 (en celeste) con respecto a la secuencia de RmInt1.

minios, al igual que las zonas de estructura secundaria predichas en ellos, pueden verse afectadas. No obstante, entre las proteínas se conserva el motivo catalítico para la actividad reverso transciptasa (RYADD) que reside en la región 5 del dominio RT, y los dos residuos de tirosina (YY) dentro del dominio madurasa. El dominio C-terminal en RmInt2 es prácticamente distinto al que presentan los otros dos intrones, sin embargo, la cadena beta que se forma en dicho dominio se mantiene en todos ellos.

R3.4 CARACTERÍSTICAS DE LAS INTERACCIONES INTRÓN-DIANA

Se ha demostrado que para la movilidad *in vivo* del intrón RmInt1 se requiere como mínimo un DNA diana que se extienda desde la posición -20 del exón 1 hasta la posición +5 del exón 2, llamada diana -20/+5. En una diana de esta longitud encontramos los diferentes elementos que forman parte del reconocimiento del intrón. Los apareamientos de bases entre las IBSs y EBSs (1, 2 y 3) de RmInt1 han sido bien caracterizados, encontrándose un nucleótido conector entre las secuencias IBS 1 y 2, llamado *linker*,

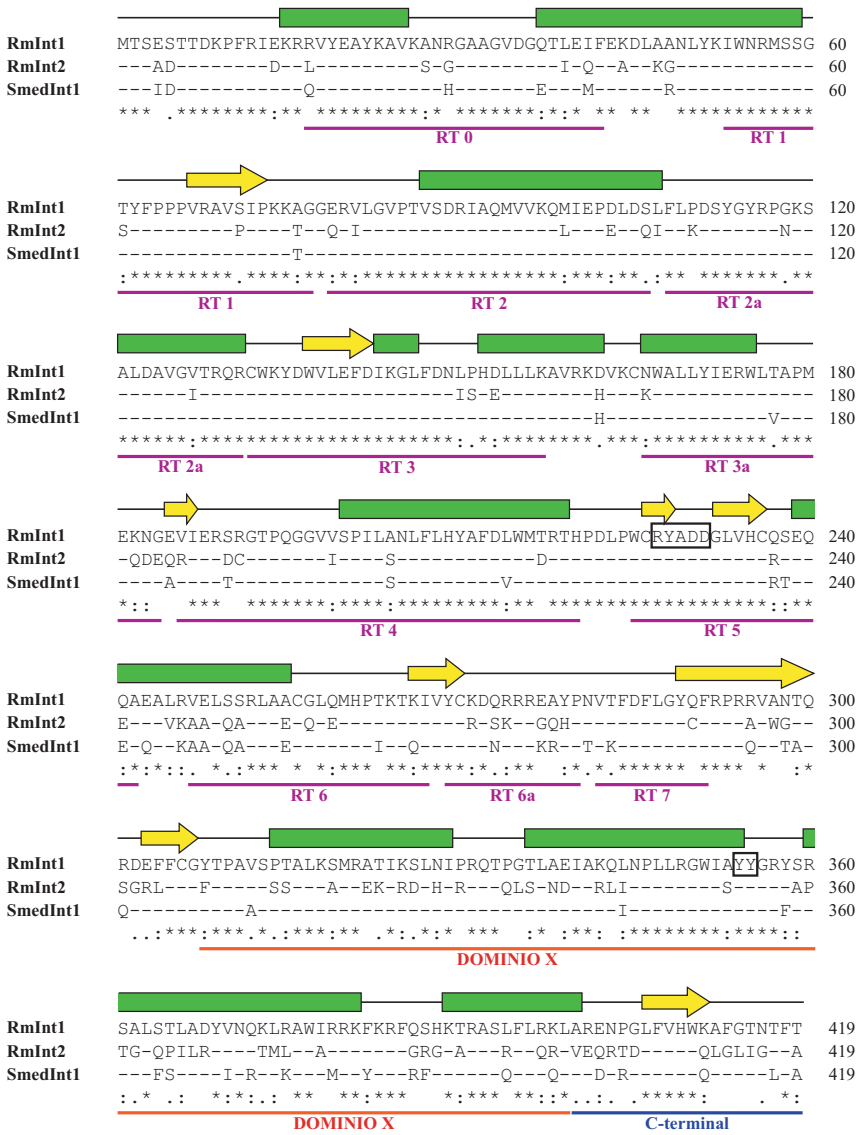


Figura R3.6: **Alineamiento de la secuencia de las proteínas codificadas por RmInt1, RmInt2 y SmedInt1.** Los aminoácidos idénticos a los presentes en la secuencia de RmInt1 están representados por guiones (-), mientras que las variaciones se indican con el aminoácido por el cual ha cambiado. Enmarcados se señalan los residuos críticos del dominio RT5 y del dominio madurasa. Bajo la secuencia se indican los límites de los dominios identificados: en morado, de los dominios RT0 al RT7, en naranja, el dominio madurasa, y en azul, la extensión C-terminal. Sobre la secuencia se muestra la predicción de la estructura secundaria (Jpred): en rectángulos verdes las hélices alfa y en flechas amarillas las cadenas beta.

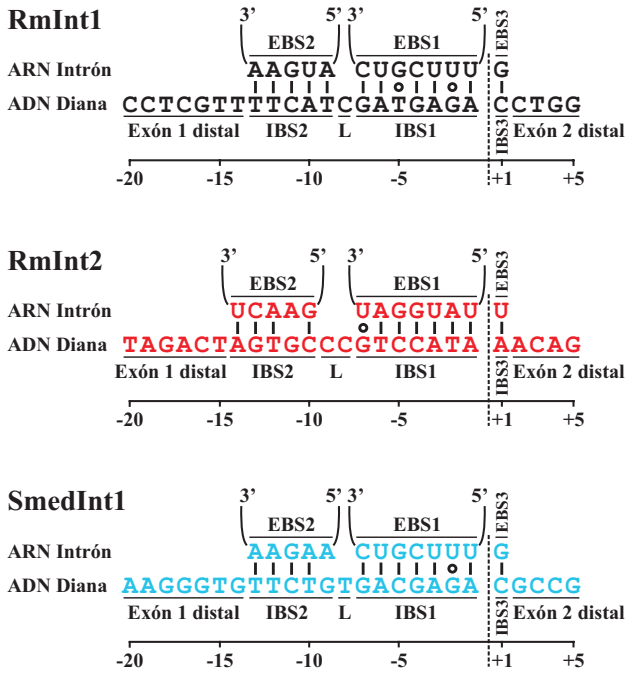


Figura R3.7: **Comparación de la zona de reconocimiento de los intrones RmInt1, RmInt2 y SmedInt1, y las interacciones que presentan con sus respectivas dianas.** El diagrama muestra una diana de longitud -20/+5, donde se representan los diferentes elementos que la componen, así como los apareamientos de bases IBS-EBS determinadas para RmInt1 (Jiménez-Zurdo *et al*, 2003) y predichas para RmInt2 y SmedInt1. La línea de puntos vertical indica el sitio de inserción del intrón.

que no tiene una posición de interacción aparente en el RNA del intrón (Jiménez-Zurdo et al., 2003). En base a este conocimiento y a la estructura del RNA (figura R3.4), hemos podido determinar dichos apareamientos en RmInt2 y SmedInt1. Para estos nuevos intrones, las interacciones IBS-EBS 1 y 3 presentan un apareamiento Watson-Crick de bases total, sin embargo, en la interacción IBS2-EBS2 el apareamiento parece no ser total (figura R3.7).

En las EBSs de SmedInt1 no existe ninguna variación con respecto a las de RmInt1, salvo en el segundo nucleótido de la EBS2. Las IBSs presentes

Tabla R3.2: Secuencia de la diana de SmedInt1 según su localización genómica. Las IBS 1 y 2, en el exón 1, y la IBS3, en el exón 2, aparecen subrayadas. En negrita se muestran los nucleótidos que difieren de la secuencia de IS*Sme3*.

Localización genómica	**Secuencia invadida**	**Exón 1**	**Exón 2**
Cromosoma	IS*Sme3*	AAGGGTGTTCTGTGACGAGA	CGCCG
pSMED01-1	Smed_4905	**GGAA**GTGTTC**ATG**GAT**G**AA**A**	CGC**GT**
pSMED02-1	IS*Sme3*	AAGGGTGTTCTGTGACGAGA	CGCCG
pSMED02-2	IS*Sme3*	AAGGGTGTTCTGTGACGAGA	CGCCG

en la diana de este intrón (la secuencia de inserción IS*Sme3*) tampoco han sufrido muchos cambios: el tercer nucleótido de la IBS1 (posición -5) ahora es una C, que sigue apareando con la G que hay en el intrón; la IBS2 presenta dos cambios, uno en la posición -10 que compensa la mutación ocurrida en el intrón, y otro en la -9 que no da lugar a un correcto apareamiento de bases; en la IBS3 se encuentra el mismo nucleótido que para RmInt1. El *linker* ha mutado, pero no debe afectar al movimiento del intrón puesto que no interviene en el reconocimiento del mismo. Hay que señalar que la copia de SmedInt1 que se encuentra interrumpiendo el gen Smed_4905 está flanqueada por unos exones distintos a la diana que presenta la secuencia de inserción IS*Sme3* (tabla R3.2). En la diana contenida en ese gen, la IBS3 no ha cambiado, y aunque la IBS1 presente dos cambios, éstos son compensados y permiten la unión con la EBS1. La IBS2 contiene dos nucleótidos diferentes, que dejan una base desapareada al interaccionar con la EBS2. Sin embargo, ese desapareamiento se encuentra en una posición distinta al que se genera cuando el intrón se une a la diana IS*Sme3*.

Las EBSs de RmInt2, así como las IBSs presentes su diana (la IR de IS*Rm17*), sí han cambiado considerablemente con respecto a lo observado en RmInt1. A pesar de ello, sigue habiendo apareamiento de bases en todos los casos salvo en la interacción IBS2-EBS2, donde, en la posición -11, no encontramos complementariedad de bases. Debido a la deleción en RmInt2 del U presente en la EBS2 de RmInt1 (figura R3.4), hemos tenido que analizar

Tabla R3.3: Secuencia de la diana de RmInt2 según su localización genómica. La zona distal de los exones 1 y 2 aparece subrayada. En negrita se muestran los nucleótidos que varían entre secuencias. Las IRs invadidas por el intrón RmInt2 se indican con la letra I entre paréntesis.

Localización genómica	**IR-IS*Rm17***	**Exón 1**	**Exón 2**
pRmeGR4c-1	IRL (I)	**TAGT**CTAGTGCCC**A**TCCATA	AAC**AG**
	IRR	**TAGA**CTAGTGCCC**A**TCCATA	AAC**GC**
pRmeGR4c*	- (I)	**TACG**CTAGTGCCC**G**TCCATA	AAC**GC**
pRmeGR4c-2^	IRL	**TAGT**CTAGTGCCC**G**TCCATA	AAC**AG**
pRmeGR4c-2	IRL	**TAGA**CTAGTGCCC**A**TCCATA	AAC**AG**
	IRR (I)	**TAGA**CTAGTGCCC**G**TCCATA	AAC**GC**
pRmeGR4c-3	IRL (I)	**TAGA**CTAGTGCCC**A**TCCATA	AAC**AG**
	IRR	**TAGT**CTAGTGCCC**A**TCCATA	AAC**GC**
pRmeGR4d-1	IRL	**TAGA**CTAGTGCCC**A**TCCATA	AAC**AG**
	IRR (I)	**TAGT**CTAGTGCCC**G**TCCATA	AAC**GC**
pRmeGR4d-2	IRL (I)	**TTAG**CTAGTGCCC**A**TCCATA	AAC**AG**
	IRR (I)	**CTAG**CTAGTGCCC**G**TCCATA	AAC**GC**

* secuencia diana que carece de IS*Rm17* completa

^ copia truncada de IS*Rm17* (contiene desde el nucleótido 1 hasta el 939) que se encuentra delante de la segunda copia de dicha IS presente en pRmeGR4c

en detalle los nucleótidos próximos a esta zona para encontrar la posición óptima de la EBS2 en RmInt2. La secuencia mostrada en la figura R3.7 es la que mejor predice la posible IBS2, aunque presente ese desapareamiento de bases. Esta propuesta supone un desplazamiento en secuencia de la IBS2 con respecto a RmInt1, y causa, a su vez, que el *linker* esté formado por dos nucleótidos en vez de por uno. Además, hay que destacar que la posición -7 es variable y puede ser una G o una A (tabla R3.3), aunque ambos nucleótidos interaccionan con la T del intrón.

Si observamos las regiones distales del DNA diana, tanto del exón 1 (posiciones -20 a -15) como del exón 2 (posiciones +2 a +5), encontramos que son completamente distintas entre intrones. En el caso de RmInt2, la zona distal del exón 1 está formada por la DR de IS*Rm17*, que varía según la localización genómica de dicha IS. Además, el intrón reconoce la secuencia

diana desde el extremo 5' hacia el 3', por consiguiente, la DR se lee en sentido en una de las IRs y en antisentido en la otra. Esto causa que los últimos cuatro nucleótidos del exón 1 sean distintos entres secuencias (tabla R3.3). Para obtener una diana de longitud -20/+5 hemos determinado su secuencia en base a la repetición de cada nucleótido en las distintas posiciones. Así, el nucleótido mayoritario en la posición -20 es la T, en la -19 es la A, en la -18 la G y en la -17 la A (TAGA). La elección de los dos nucleótidos diferentes del exón 2 se basó en el número de lecturas que, tras la secuenciación de GR4, se emparejaban con esa región. Aunque existe el mismo número de secuencias diana que presentan el final AG que GC, el número de lecturas solapadas sobre esa zona era ligeramente mayor para el final AG. Por último, se decidió que la posición variable -7 fuera una G debido a que la IR que se encuentra más veces invadida en GR4 presenta dicho nucleótido.

Las dianas analizadas en este apartado (figura R3.7) son las que se introdujeron en un vector para crear los plásmidos receptores (ver Material y Métodos, apartado M.2.2) usados en los ensayos de movilidad que posteriormente se realizaron con los intrones.

R3.5 ESCISIÓN Y MOVILIDAD DE LOS DISTINTOS INTRONES

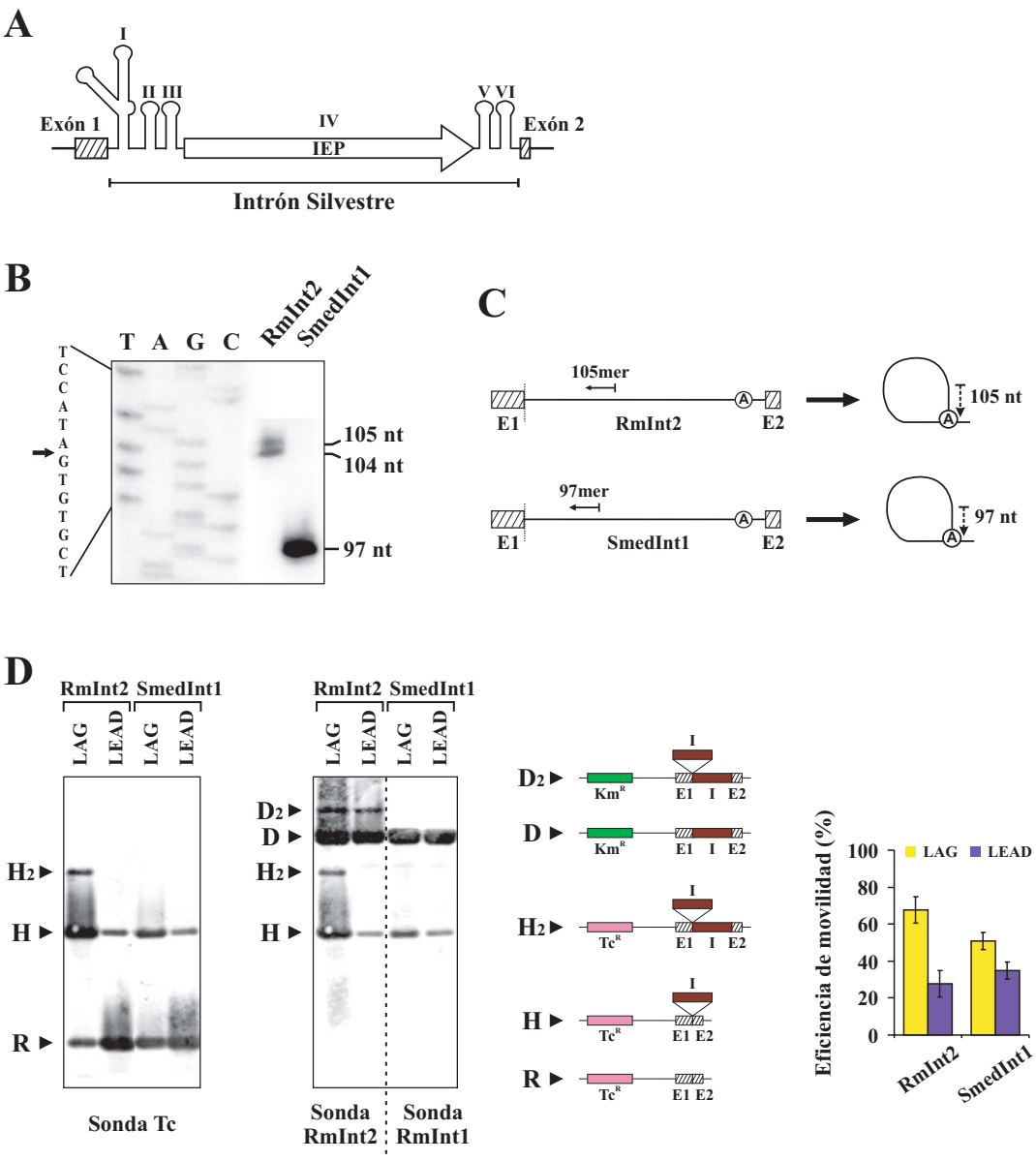
R3.5.1 Ensayos con formas de intrón silvestre

Debido a todos los cambios ocurridos en la secuencia tanto de la ribozima como de la proteína de los intrones RmInt2 y SmedInt1, nos preguntamos si su capacidad de escisión y movilidad eran comparables a las de RmInt1. Con el fin de averiguar si los nuevos intrones se escindían, llevamos a cabo un ensayo de extensión a partir de cebador (ver Material y Métodos, apartado M.9). En el ensayo se analizaron, mediante dos oligonucleótidos distintos, muestras procedentes de células de RMO17 que expresaban cada uno de los intrones silvestres (figura R3.8A). Tras realizar una electroforesis en gel de poliacrilamida desnaturalizante al 6 % con dichas muestras, se observaron varios productos de extensión (figura R3.8B). Con SmedInt1

apareció una banda de 97 nt, mientras que el oligonucleótido usado con RmInt2, complementario a una zona más alejada del inicio del intrón (figura R3.81C), generó dos bandas: una de 105 nt y otra de 104 nt. Las bandas de 97 nt y 105 nt corresponden al producto de escisión de ambos intrones. La banda de 104 nt observada en la muestra de RmInt2 todavía no ha sido caracterizada.

Tras comprobar que los nuevos intrones mantenían la capacidad de escisión, nos preguntamos si éstos podrían insertarse en la diana donde han sido encontrados a pesar de que la interacción IBS2-EBS2 presenta desapareamiento de bases en ambos intrones. Para ello planteamos un ensayo de movilidad que sigue un sistema de dos plásmidos (ver Material y Métodos, apartado M.10), uno donador que contiene cada uno de los intrones silvestres (figura R3.8A), y otro receptor que presenta la diana de inserción para esos intrones en ambas orientaciones, LAG y LEAD (figura M.3B y M.3C). El resultado de los eventos de movilidad (figura R3.8D) se analizó mediante hibridación DNA-DNA con una sonda específica del plásmido receptor (sonda Tc; tabla M.5). La digestión del DNA plasmídico con *Sal*I permite diferenciar el producto de *homing* (H), de 3'3 kb, del plásmido receptor sin invadir (R), de 1'4 kb, y por tanto, permite calcular la eficiencia de invasión aplicando la fórmula: H/(H+R). Para comprobar que realmente la banda del producto de *homing* es el resultado de un evento de invasión y no una reacción cruzada (inespecificidad de la sonda), se realizaron nuevas hibridaciones de la membrana con una sonda específica de cada intrón. Así se puso de manifiesto la banda correspondiente al producto de *homing* y al plásmido donador (D), que con *Sal*I se linealiza (7'2 kb). En las muestras de RmInt2 encontramos dos bandas adicionales (que aparecen al hibridar tanto con la sonda Tc como con la sonda RmInt2), una de aproximadamente 5'2 kb (H_2) y otra algo inferior a los 9'4 kb (D_2). Ambas tienen el tamaño correspondiente a un plásmido con intrón (producto de *homing* o plásmido donador) que ha sufrido una nueva invasión del intrón.

El porcentaje de movilidad de cada muestra es el resultado de la media de la eficiencia de invasión, con su correspondiente error estándar, de al menos


A
I
II III
IV
V VI
Exón 1
IEP
Exón 2
Intrón Silvestre
B
T A G C
RmInt2
SmedInt1
T C C A T A G T G T G C T
105 nt
104 nt
97 nt
C
105mer
E1
RmInt2
E2
105 nt
97mer
E1
SmedInt1
E2
97 nt
D
RmInt2 SmedInt1
LAG LEAD LAG LEAD
H2
H
R
Sonda Tc
RmInt2 SmedInt1
LAG LEAD LAG LEAD
D2
D
H2
H
Sonda RmInt2
Sonda RmInt1
D2
Km^R
E1 I E2
D
Km^R
E1 I E2
H2
Tc^R
E1 I E2
H
Tc^R
E1 E2
R
Tc^R
E1 E2
Eficiencia de movilidad (%)
100
80
60
40
20
0
LAG
LEAD
RmInt2
SmedInt1

cuatro colonias individuales de transconjugantes de RMO17 multiplicado por 100. Este porcentaje es mayor en RmInt2 (68 %) que en SmedInt1 (51 %) cuando la invasión se produce en la diana situada en orientación LAG. Sin embargo, SmedInt1 presenta mayor eficiencia de movilidad sobre la diana en orientación LEAD (35 %) que RmInt2 (28 %). No obstante, en ambos intrones los eventos de invasión ocurren con mayor frecuencia sobre la diana localizada en la hebra que sirve de molde para la síntesis de la cadena retrasada en la horquilla de replicación.

R3.5.2 Ensayos con formas de intrón derivadas (ΔORF)

En trabajos previos se ha descrito que la eficiencia de movilidad del intrón RmInt1 se mejora generando una construcción donde la IEP se encuentra delante de la ribozima (ΔORF), plásmido denominado pKGEMA4. Expresado desde este plásmido, RmInt1 presenta una eficiencia de movili-

---

Figura R3.8: **Ensayos de escisión y movilidad de RmInt2 y SmedInt1.** (A) Diagrama de la construcción donadora usada en estos ensayos. (B) Para la reacción de extensión a partir de cebador se extrajo RNA total de células de RMO17 que portan un plásmido con el intrón silvestre. Los carriles T, A, G, C representan la escalera de nucleótidos producida para determinar el tamaño de banda correcto del intrón RmInt2 escindido, indicado con una flecha. (C) Representación de la reacción de extensión con el cebador usado para cada intrón sobre una molécula de éste no escindida. La línea de puntos indica el nucleótido de parada de dicha reacción, y a la derecha se muestra el producto de escisión esperado. También se destaca la A desapareada del dominio VI. (D) El ensayo de movilidad se llevó a cabo en células de RMO17 que contienen un plásmido donador y otro receptor (con la diana específica para cada intrón en ambas orientaciones LAG y LEAD). Las hibridaciones DNA-DNA muestran cinco tipos de moléculas, representadas esquemáticamente a la derecha de éstas, siendo: R, plásmido receptor sin invadir; H, producto de *homing*; H2, producto de *homing* con una segunda invasión; D, plásmido donador; D2, plásmido donador que ha sufrido una invasión. La eficiencia de *homing* de cada intrón (calculada como [H/(H+R)]*100), con su correspondiente error estándar, se muestra en el gráfico de barras.

◀ (Página anterior)

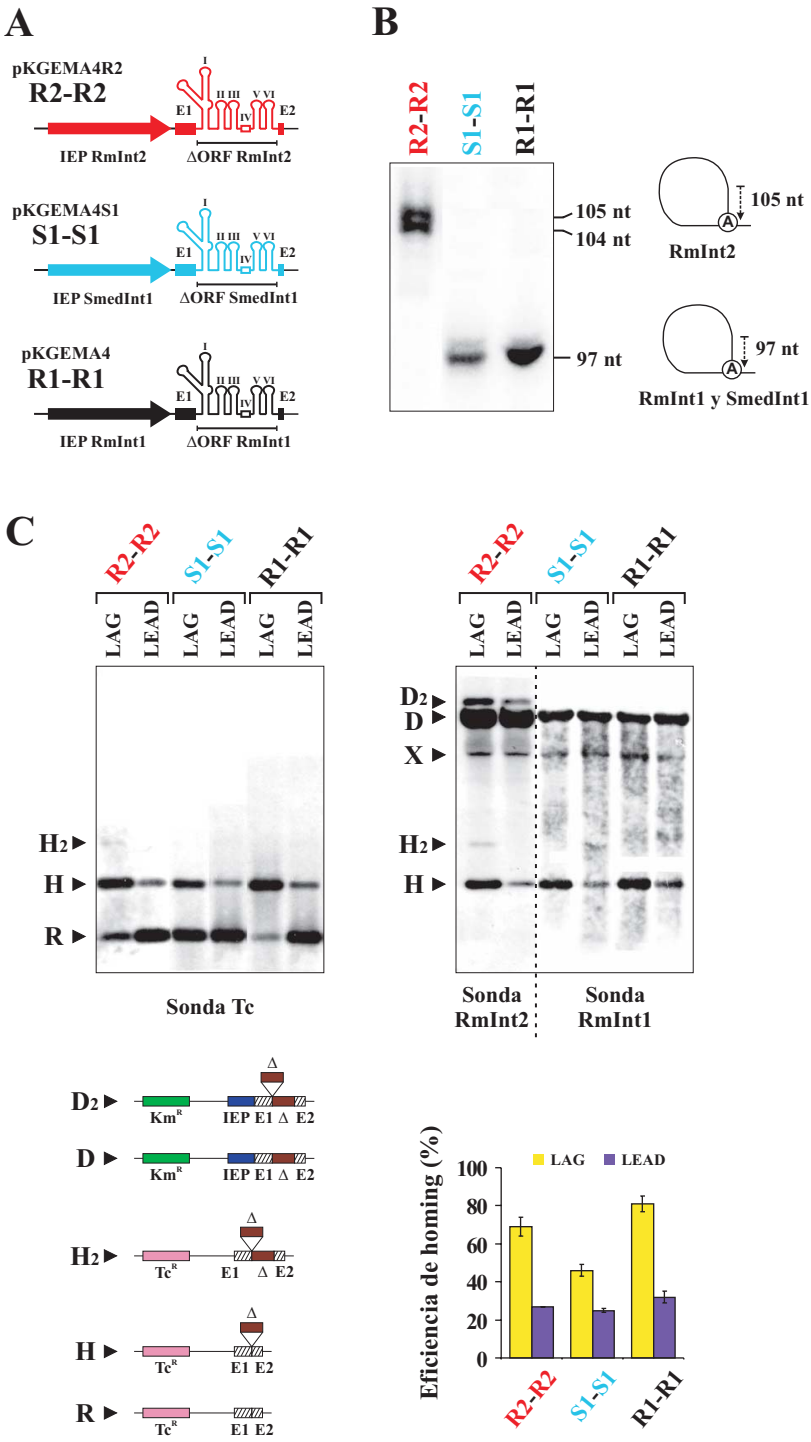
dad dos veces superior a la del intrón en su forma silvestre (Nisa-Martínez *et al*, 2007). Basándonos en este plásmido, se generaron construcciones similares para los nuevos intrones estudiados (ver Material y Métodos, apartado M.2.1; figura R3.9A).

El análisis llevado a cabo mediante extensión a partir de cebador (ver Material y Métodos, apartado M.9) reveló que los intrones RmInt2 y SmedInt1 en su forma derivada ΔORF, al igual que en su forma silvestre, eran capaces de escindirse, dando lugar a los mismos productos de escisión (figura R3.9C). En las muestras de RmInt1 y SmedInt1, donde el ensayo se realizó con el oligonucleótido 97mer, apareció una banda de 97 nt. En el caso de RmInt2, donde se usó el oligonucleótido 105mer, se generó una banda de 105 nt (correspondiente al producto de escisión) y otra de 104 nt (aún no caracterizada).

Con este tipo de construcciones también llevamos a cabo ensayos de movilidad (ver Material y Métodos, apartado M.10) en células de RMO17 donde se introdujeron uno de estos plásmidos donadores y el plásmido receptor con la diana de cada ribozima en ambas orientaciones, LAG y LEAD.

Figura R3.9: **Ensayos de escisión y movilidad de la forma derivada ΔORF de los tres intrones estudiados.** (A) Representación de las construcciones donadoras usadas en estos ensayos, donde la IEP se encuentra aguas arriba de la ribozima. (B) Resultado de la extensión a partir de cebador llevada a cabo con RNA total extraído de células de RMO17 que portan cada una de las construcciones donadoras. A su derecha se muestra el producto de escisión esperado para cada intrón. (C) El ensayo de movilidad se realizó en células de RMO17 que llevan un plásmido donador y otro receptor (con la diana específica para cada intrón en ambas orientaciones LAG y LEAD). Las hibridaciones DNA-DNA muestran seis tipos de moléculas, representadas esquemáticamente debajo de éstas, siendo: R, plásmido receptor sin invadir; H, producto de *homing*; H2, producto de *homing* con una segunda invasión; x, banda inespecífica; D, plásmido donador; D2, plásmido donador que ha sufrido una invasión. La eficiencia de *homing* de cada intrón (calculada como [H/(H+R)]*100), con su correspondiente error estándar, se muestra en el gráfico de barras.

(Página siguiente) ►


A
pKGEMA4R2
R2-R2
IEP RmInt2
ΔORF RmInt2
pKGEMA4S1
S1-S1
IEP SmedInt1
ΔORF SmedInt1
pKGEMA4
R1-R1
IEP RmInt1
ΔORF RmInt1
B
R2-R2
S1-S1
R1-R1
105 nt
104 nt
97 nt
105 nt
RmInt2
97 nt
RmInt1 y SmedInt1
C
LAG
LEAD
Sonda Tc
Sonda RmInt2
Sonda RmInt1
Eficiencia de homing (%)

El resultado se muestra en la figura R3.9D, donde, tras una hibridación DNA-DNA con sondas específicas del plásmido receptor y de los intrones, aparecieron las bandas características de este tipo de ensayos: una banda correspondiente al plásmido receptor sin invadir (R) y al producto de *homing* (H) cuando hibridamos con la sonda Tc, y una banda correspondiente al producto de *homing* y al plásmido donador (D) cuando se hibridó con la sonda del intrón. Esto ocurrió en las muestras de RmInt1 y SmedInt1, sin embargo, en las de RmInt2 encontramos bandas adicionales correspondientes a plásmidos que, conteniendo una copia de intrón, sufrieron un nuevo evento de invasión (bandas H_2 y D_2).

Respecto a la movilidad de estas formas de intrón derivadas, RmInt1 mostró la mayor eficiencia (81 %) cuando la invasión tiene lugar sobre la diana situada en orientación LAG. El porcentaje de movilidad de RmInt2 no se alejó mucho (69 %), sin embargo, la eficiencia de SmedInt1 fue casi la mitad que la de RmInt1 (46 %). En el resultado del ensayo sobre la diana

Tabla R3.4: Eficiencias de escisión y movilidad sobre la diana en ambas orientaciones, LAG y LEAD, de todas las construcciones con intrones quiméricos y formas ΔORF.

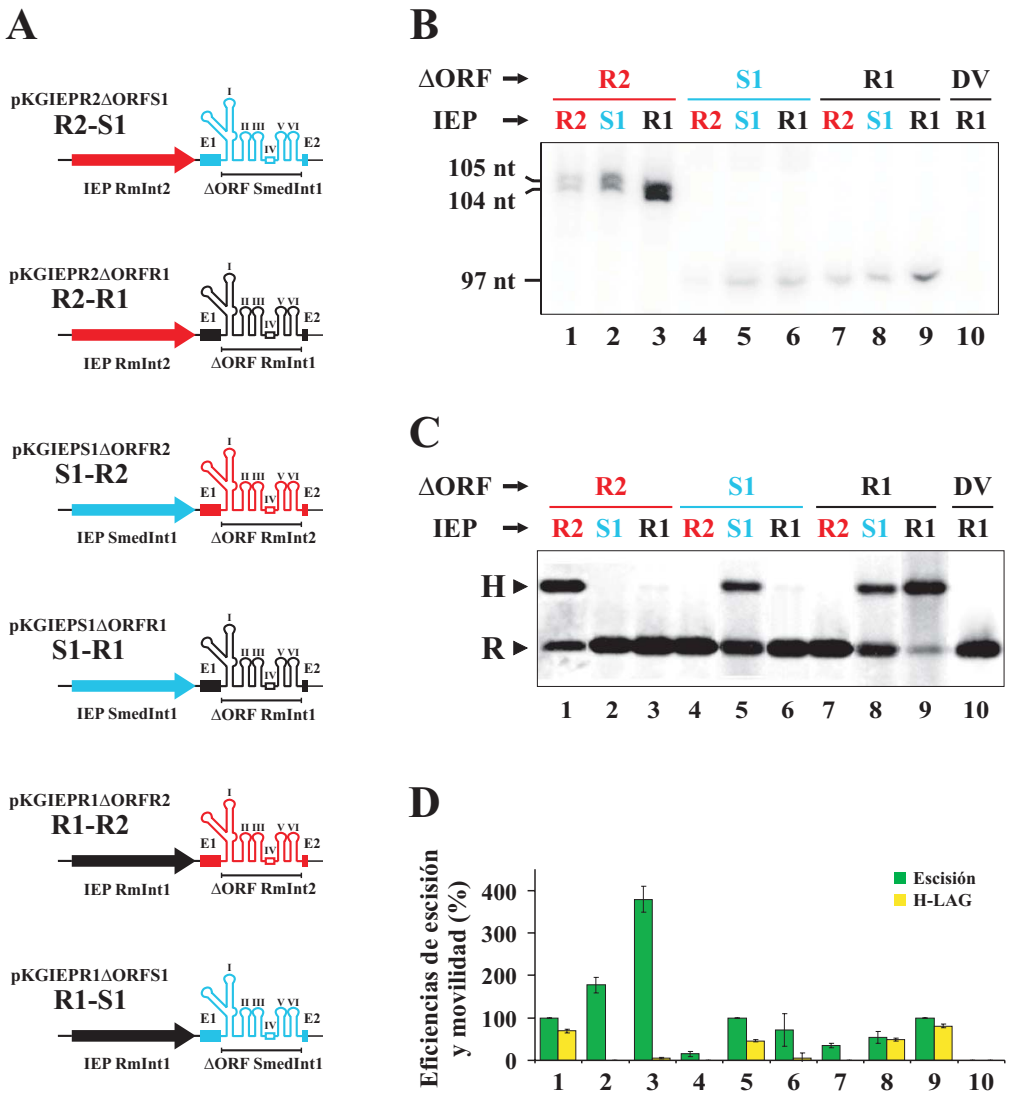
Plásmido	**% Escisión**	***% Homing***	
		LAG	**LEAD**
pKGIEPR2ΔORFR2	100±0	69±3	27±0
pKGIEPS1ΔORFR2	178±13	0±0	0±0
pKGIEPR1ΔORFR2	379±22	5±1	0±0
pKGIEPR2ΔORFS1	15±4	0±0	0±0
pKGIEPS1ΔORFS1	100±0	46±2	25±1
pKGIEPR1ΔORFS1	71±27	6±6	0±0
pKGIEPR2ΔORFR1	34±4	0±0	0±0
pKGIEPS1ΔORFR1	54±10	49±2	8±5
pKGIEPR1ΔORFR1	100±0	81±2	32±2
pKGIEPdVΔORFR1	0±0	0±0	0±0

en orientación LEAD se observó una relación similar, donde RmInt1 sigue siendo el intrón con mayor porcentaje de movilidad (32 %). No obstante, el rango de eficiencia de invasión de los tres intrones en este caso fue inferior al obtenido con la diana en LAG, presentando RmInt2 un 27 % y SmedInt1 un 25 % (tabla R3.4).

R3.5.3 Ensayos con formas de intrón quiméricas.

En base a estudios filogenéticos se ha sugerido una coevolución entre la ribozima y la proteína de los intrones del grupo II (Toor *et al*, 2001). Sin embargo, no hay datos empíricos mediante el uso de intrones quiméricos que lo demuestren. Una vez analizada la escisión y movilidad de las nuevas construcciones donadoras de intrón, donde la IEP se encuentra aguas arriba de la ribozima, y sabiendo que los tres intrones estudiados están muy relacionados evolutivamente, nos preguntamos si las proteínas y las ribozimas de estos intrones podrían intercambiarse. Para resolver esta cuestión, generamos una serie de plásmidos donadores siguiendo la misma estrategia que para la construcción de la forma derivada ΔORF de los intrones a partir de pKGEMA4 (ver Material y Métodos, apartado M.2.1). Estos nuevos plásmidos donadores, que contienen intrones quiméricos, se representan en la figura R3.10A.

Primeramente analizamos la escisión de estos intrones quiméricos mediante extensión a partir de cebador (ver Material y Métodos, apartado M.9). En las muestras que contenían plásmidos con el ΔORF de RmInt2 se usó, para la extensión, el oligonucleótido 105mer, mientras que para las muestras que presentaban construcciones con las ribozimas de RmInt1 y SmedInt1 se utilizó el oligonucleótido 97mer. Como hemos comentado para construcciones anteriores, en las muestras con ribozima R1 y S1 apareció una banda de 97 nt, y con ribozima R2 una banda de 105 nt y otra de 104 nt. Las bandas correspondientes al producto de escisión de esas ribozimas (bandas de 97 nt y 105 nt) son las que se usaron para cuantificar el proceso de escisión en cada una de las construcciones con intrones quiméricos.


A
pKGIEPR2ΔORFS1
R2-S1
IEP RmInt2
ΔORF SmedInt1
pKGIEPR2ΔORFR1
R2-R1
IEP RmInt2
ΔORF RmInt1
pKGIEPS1ΔORFR2
S1-R2
IEP SmedInt1
ΔORF RmInt2
pKGIEPS1ΔORFR1
S1-R1
IEP SmedInt1
ΔORF RmInt1
pKGIEPR1ΔORFR2
R1-R2
IEP RmInt1
ΔORF RmInt2
pKGIEPR1ΔORFS1
R1-S1
IEP RmInt1
ΔORF SmedInt1
E1
E2
I
II III
IV
V VI
B
ΔORF
IEP
R2
S1
R1
DV
105 nt
104 nt
97 nt
1 2 3 4 5 6 7 8 9 10
C
H
R
D
Eficiencias de escisión y movilidad (%)
Escisión
H-LAG
0
100
200
300
400

El porcentaje de escisión de estas nuevas construcciones es la media de al menos tres muestras independientes, y se calculó en relación a cada plásmido ΔORF de referencia, en los que consideramos una eficiencia del 100 % (tabla R3.4). Como se observa en la figura R3.10B, sólo la ribozima R2 presentó un porcentaje de escisión mayor al ser complementada por la proteína de los otros intrones que con la suya propia. Así, la IEP R1 generó una escisión de la ribozima R2 del 379 %, y la IEP S1 del 178 %. Con respecto a la escisión del ΔORF S1, éste presentó un 71 % de eficiencia cuando se estaba expresando la proteína R1 y un 15 % cuando lo hacía la IEP R2. Por su parte, la ribozima R1 mostró un porcentaje de escisión del 54 % con la proteína S1 y un 34 % con la R2. Como control negativo utilizamos un derivado de pKGEMA4, llamado pKGEMA4DV, que contiene una mutación hacia el final del dominio V de la ribozima que afecta a la triada catalítica. Esta mutación provoca la pérdida de la actividad catalítica y hace, por tanto, que la ribozima no sea funcional.

Posteriormente realizamos ensayos de movilidad (ver Material y Métodos, apartado M.10) en células de RMO17 con todas las quimeras sobre la

---

Figura R3.10: **Ensayos de escisión y movilidad de las quimeras proteína-ribozima.** (A) Representación de las seis construcciones donadoras de intrones quiméricos, donde la IEP de cada intrón se asocia a un ΔORF diferente. RmInt1 aparece en negro, RmInt2 en rojo, y SmedInt1 en celeste. (B) Análisis por extensión a partir de cebador realizada con RNA total extraído de células de RMO17 que contienen un plásmido con la IEP y el ΔORF del intrón indicado en cada muestra. Como control negativo se usó un mutante de RmInt1 en el dominio V. (C) El ensayo de movilidad se llevó a cabo en células de RMO17 que portan un plásmido donador (con la IEP y el ΔORF del intrón indicado) y otro receptor (con la diana específica para cada intrón, dependiendo de la ribozima presente en el plásmido donador, en orientación LAG). Una hibridación con la sonda Tc pone de manifiesto el plásmido receptor (R) y el producto de *homing* (H). (D) La eficiencia de escisión de cada quimera (calculada en relación a la molécula IEP-ΔORF del intrón de referencia), la eficiencia de *homing* sobre la diana en LAG (calculada como [H/(H+R)]*100) y sus correspondientes errores estándares se muestran en el gráfico de barras.

◀ (Página anterior)

secuencia de inserción correspondiente para cada ribozima. Así pues, para los plásmidos que contienen el ΔORF de RmInt1 el ensayo se realizó con la diana IS*Rm2011-2* (pJB0.6LAG y pJB0.6LEAD; tabla M.2), para RmInt2 con la IR de IS*Rm17* (pJBIS*Rm17*LAG y pJBIS*Rm17*LEAD; figura M.3C), y para SmedInt1 con la diana IS*Sme3* (pJBIS*Sme3*LAG y pJBIS*Sme3*LEAD; figura M.3B). Como se observa en la figura R3.10C, sólo cuatro construcciones presentaron eventos de movilidad tras realizar una hibridación DNA-DNA con la sonda Tc, que pone de manifiesto el plásmido receptor (R) y el producto de *homing* (H). En el apartado anterior analizamos la movilidad de los plásmidos que contienen la proteína y la ribozima del mismo intrón (forma derivada ΔORF), y observamos que RmInt1 era el intrón con mayor eficiencia de *homing* (tabla R3.4). Tres de las construcciones con intrones quiméricos no fueron capaces de invadir la diana en orientación LAG (S1-R2, R2-S1 y R2-R1). El resto mostraron un bajo porcentaje de invasión (R1-R2 un 5 % y R1-S1 un 6 %), salvo la quimera S1-R1 que presentó un 49 %. Con respecto a la invasión de la diana en LEAD, sólo la construcción S1-R1, además de la forma ΔORF de los tres intrones, presentó movilidad, con una eficiencia del 8 % (tabla R3.4).

R3.6 PAPEL DE LA IEP EN EL RECONOCIMIENTO DE LA DIANA: CONSTRUCCIÓN DE DIANAS QUIMÉRICAS

Los ensayos realizados con los intrones quiméricos proporcionaron un dato sorprendente: la quimera R1-R2, a pesar de presentar un alto nivel de escisión del ΔORF R2, no presentó invasión de su diana. No obstante, en apartados previos hemos puesto de manifiesto la movilidad del intrón RmInt2 tanto en su forma silvestre como derivada ΔORF. Por consiguiente, esta ausencia de movilidad sugiere que la proteína debe estar jugando un papel importante en el proceso de reconocimiento de la diana. Esto nos llevó a generar un plásmido receptor que incluye una diana donde la parte que contiene el *linker* y las IBSs, reconocidas por la ribozima R2, pertenece a su diana original (la IR de IS*Rm17*), y la zona distal de los exones 1 y 2 (en

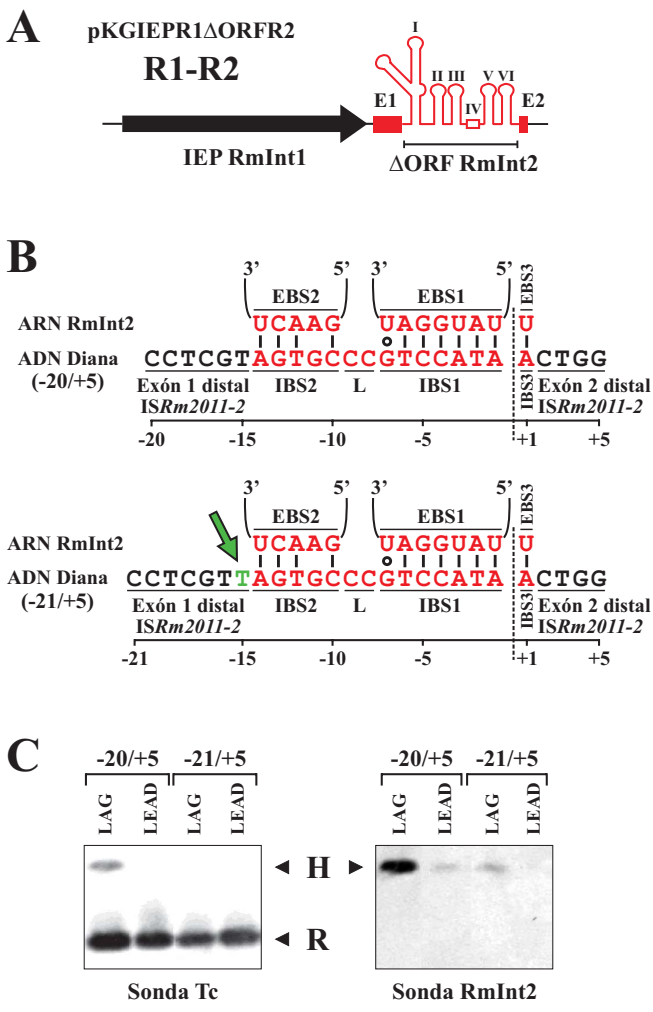


Figura R3.11: **Ensayo de movilidad sobre dianas quiméricas.** (A) Diagrama de la construcción donadora usada en estos ensayos. (B) Apareamiento de bases entre RmInt2 y dos dianas quiméricas compuestas por la zona distal de los exones 1 y 2 de IS*Rm2011-2* (en negro) y las IBSs y el *linker* de la IR de IS*Rm17* (en rojo). La diana quimérica -21/+5 contiene una T extra (en verde) entre la zona distal del exón 1 y la IBS2. La línea de puntos vertical indica el sitio de inserción del intrón. (C) Resultado del ensayo de movilidad realizado en células de RMO17 que portan un plásmido donador (el R1-R2) y otro receptor (con las dianas quiméricas en ambas orientaciones LAG y LEAD). Las hibridaciones DNA-DNA muestran dos tipos de bandas, la del plásmido receptor (R) y la del producto de *homing* (H).

teoría reconocidas por la proteína) se corresponden con la diana IS*Rm2011-2* (ver Material y Métodos, apartado M.2.2; figura M.3D y M.3E). Ésta se denominó diana quimérica -20/+5, para diferenciarla de una segunda diana quimérica que se propuso siguiendo las características descritas para RmInt1, donde se conoce que la T en posición -15 es crítica para el *homing* (Jiménez-Zurdo *et al*, 2003). Además, esa T -15 dentro de la diana IS*Rm2011-2* se encuentra separada de la IBS2 por un nucleótido. Por esta razón, a la quimera -20/+5 se le introdujo una T extra entre la zona distal del exón 1 y la IBS2, dando lugar a la diana quimérica -21/+5 (figura R3.11B).

Con el fin de realizar un ensayo de movilidad (ver Material y Métodos, apartado M.10), se transfirió la construcción donadora R1-R2 y cada uno de los plásmidos receptores con las dianas quiméricas (-20/+5 y -21/+5) en ambas orientaciones, LAG y LEAD, a células de RMO17. El DNA se digirió con *Sal*I y se hibridó con sondas específicas del plásmido receptor y del intrón. El resultado de esas hibridaciones DNA-DNA se muestra en la figura R3.11C. En ella observamos que con la sonda Tc sólo apareció una banda en el carril correspondiente al plásmido receptor que contiene la diana quimérica -20/+5 en LAG, cuya cuantificación reveló un porcentaje de movilidad del 27 %. No obstante, al hibridar con la sonda del intrón se observaron dos bandas adicionales muy tenues en las muestras pertenecientes a la diana quimérica -20/+5 en LEAD y a la -21/+5 en LAG. Las nuevas bandas, no cuantificables con la sonda Tc, ponen de manifiesto que ocurrieron eventos de invasión en esas dianas quiméricas pero en baja proporción.

R3.7 REGIÓN DE LA IEP IMPLICADA EN LA MOVILIDAD DEL INTRÓN: CONSTRUCCIÓN DE UNA IEP QUIMÉRICA

En el apartado anterior hemos demostrado la importancia que tiene la zona distal de los exones de la diana sobre la movilidad de un intrón constituido por la proteína de RmInt1 y la ribozima de RmInt2. En él sugerimos que la IEP juega un papel fundamental durante el proceso de movilidad. Molina-Sánchez *et al* (2010) pusieron de manifiesto que, en la región

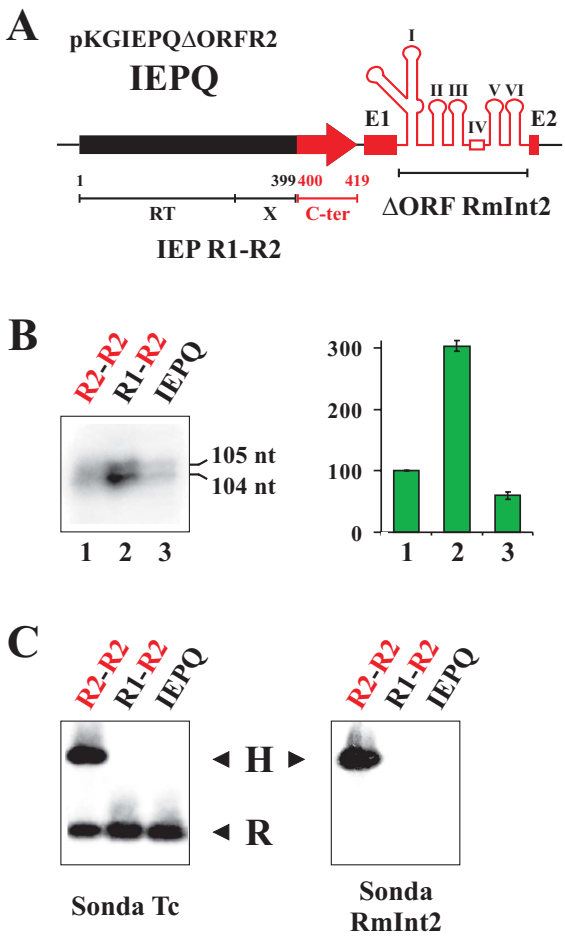


Figura R3.12: **Ensayos de escisión y movilidad de un plásmido que incluye un intrón con una IEP quimérica.** (A) Representación de la construcción donadora usada en estos ensayos, donde se ha intercambiado la parte C-terminal de la proteína de RmInt1 por la de RmInt2 (los últimos 20 aminoácidos de la proteína) en la quimera R1-R2. (B) Análisis por extensión a partir de cebador llevada a cabo con RNA total extraído de células de RMO17 que contienen el plásmido indicado en cada muestra. A la derecha se muestra la eficiencia de escisión de éstas con su correspondiente error estándar. (C) Resultados del ensayo de movilidad realizado en células de RMO17 que portan un plásmido donador y el plásmido receptor con la diana de RmInt2 en orientación LAG. Las hibridaciones DNA-DNA muestran dos tipos de bandas, la del plásmido receptor (R) y la del producto de homing (H).

C-terminal de la IEP, una α-hélice predicha y los últimos aminoácidos eran requeridos específicamente para la inserción de RmInt1 en su DNA diana. A su vez, la región C-terminal de la proteína es la más divergente entre la IEP de RmInt1 y RmInt2. Por todo ello, decidimos generar un plásmido donador de intrón que contiene la ribozima R2 y una proteína con la región RT y madurasa de RmInt1 y la región C-terminal de RmInt2 (desde el aminoácido 400 al 419, ambos inclusive; figura R3.12A). Esta proteína quimérica se obtuvo mediante una PCR de extensión por solapamiento (OE-PCR; ver Material y Métodos, apartado M.6.3; figura M.5) a partir de la forma ΔORF derivada de RmInt1 y RmInt2. Con esta nueva construcción se realizaron los mismos ensayos que con construcciones anteriores: extensión a partir de cebador y ensayo de movilidad.

Para conocer si esta IEP quimérica era capaz de complementar al ΔORF R2, y así generar productos de escisión, se hizo una extensión con el cebador 105mer a partir de RNA extraído de células de RMO17 que portan esa nueva construcción IEPQ. Como controles de la reacción de extensión se pusieron la quimera R2-R2 y R1-R2, para poder comparar el porcentaje de producto escindido que presenta esta nueva quimera IEPQ. Tomamos de referencia el nivel de escisión de la forma ΔORF derivada de RmInt2, al que le dimos el valor del 100 %. En la figura R3.12B se observa que la proteína quimérica no generó una complementación tan buena sobre la ribozima R2 como la IEP R1, incluso el nivel de producto escindido fue inferior al que presentó cuando se expresaba la IEP R2. El porcentaje de escisión de las dos construcciones quimeras en relación al plásmido R2-R2 fue del 303 % para R1-R2 y del 60 % para IEPQ. A pesar de esta baja cantidad de ribozima escindida mostrada por la IEPQ, realizamos un ensayo de movilidad usando, de nuevo, estas tres quimeras. Las hibridaciones DNA-DNA con sondas específicas del plásmido receptor y del intrón (figura R3.12C) mostraron eventos de movilidad tan solo en la forma derivada ΔORF de RmInt2.

R3.8 MOVILIDAD DE LOS INTRONES RmInt1 Y RmInt2 EN *E. coli*

Debido a la utilidad como herramienta biotecnológica que tienen este tipo de elementos móviles, decidimos comparar la movilidad de los intrones RmInt1 y RmInt2 en el fondo genético de la γ-proteobacteria *E. coli*. En esta bacteria se han llevado a cabo diversos estudios sobre la movilidad del intrón Ll.ltrB de *Lactococcus lactis* (Zhong *et al*, 2003) y de RmInt1 (García-Rodríguez *et al*, 2011).

Para determinar la movilidad de los intrones RmInt1 y RmInt2 en *E. coli* planteamos un ensayo de doble plásmido (ver Material y Métodos, apartado M.10), en el cual utilizamos los plásmidos donadores de intrón en su forma silvestre y derivada (pKG2.5 y pKGEMA4 para RmInt1, y pKGRmInt2 y pKGEMA4R2 para RmInt2, respectivamente), y los plásmidos receptores que portan la diana de cada intrón en ambas orientaciones, LAG y LEAD (figura R3.13). RmInt1 no mostró eventos de invasión sobre la diana en orientación LEAD. El porcentaje de movilidad que presentó dicho intrón en su forma derivada ΔORF sobre la diana en LAG fue cuatro veces superior a la eficiencia mostrada por su forma silvestre (27 % frente a 7 %). Sin embargo, la relación entre estas dos formas del intrón RmInt2 no fue la misma. Su forma derivada ΔORF mostró un porcentaje de movilidad sobre la diana en LAG del 34 %, mientras que el porcentaje observado en su forma silvestre fue del 41 %. RmInt2 también mostró movilidad sobre la diana en orientación LEAD, aunque sólo su forma silvestre (16 %).

R3.9 DISCUSIÓN

En este capítulo se ha llevado a cabo la descripción y caracterización funcional de dos nuevos intrones del grupo II relacionados con el intrón RmInt1 y de la secuencia diana a la que se encuentran asociados. Además, nuestro estudio ha corroborado en parte la coevolución ribozima-proteína que se ha sugerido que siguen estos retroelementos (Toor *et al*, 2001). La mayoría de intrones de este tipo identificados hasta la fecha interrumpen

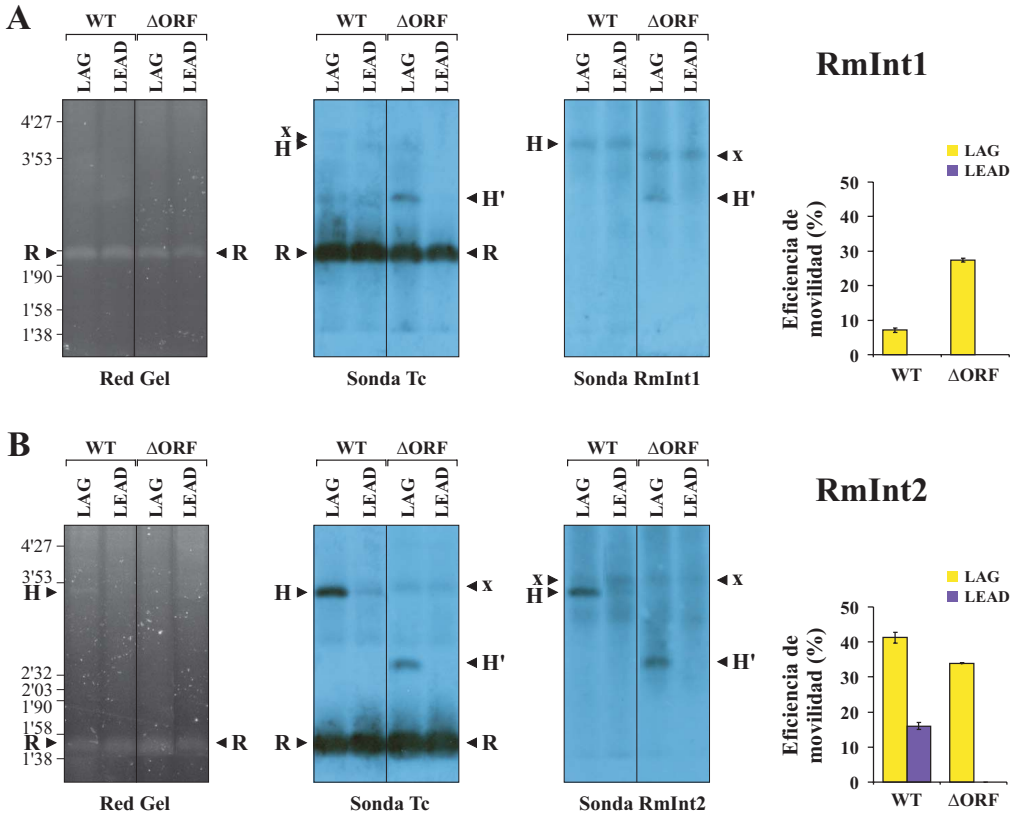


Figura R3.13: **Ensayo de movilidad de RmInt1 y RmInt2 en *E. coli*.** Este ensayo se llevó a cabo con las formas silvestres y derivadas ΔORF de RmInt1 (pKG2.5 y pKGEMA4) y RmInt2 (pKGRmInt2 y pKGEMA4R2), cuyo resultado se muestra en (A) y (B) respectivamente. Las imágenes muestran, de izquierda a derecha, el gel de agarosa y sus hibridaciones con una sonda específica del plásmido receptor y de los intrones. En cada carril se indica la orientación de la diana en el plásmido receptor usado: IS*Rm2011-2* para RmInt1 y la IR de IS*Rm17* para RmInt2. Las hibridaciones DNA-DNA muestran el plásmido receptor sin invadir (R) y el producto de homing en cada caso (H y H'). La x señala bandas inespecíficas de la hibridación. La eficiencia de movilidad de cada intrón (calculada como [H/(H+R)]*100), con su correspondiente error estándar, se muestra en el gráfico de barras.

elementos móviles como secuencias de inserción y transposones, generando una asociación que es posible que favorezca la propagación y dispersión de esos intrones dentro de un mismo genoma y entre distintas especies (Martínez-Abarca & Toro, 2000; Nisa-Martínez *et al*, 2007).

Una parte del capítulo ha estado destinada a la descripción y comparación de los tres intrones que hemos considerado como representativos dentro del grupo de intrones tipo RmInt1 (análisis realizado mediante árboles filogenéticos en el capítulo 2 de esta Tesis Doctoral; figura R2.1): RmInt1, RmInt2 y SmedInt1. Primeramente hemos estudiado en detalle la secuencia de inserción que invaden cada uno de ellos (figura R3.1), y, a su vez, hemos descrito una nueva IS perteneciente a la familia IS*66*. Las ISs de esta familia se caracterizan por tener un tamaño de unas 2'7 kb, IRs de 15-27 pb, DRs de 8 pb y tres ORFs orientados en la misma dirección. Entre los dos primeros (*tnpA* y *tnpB*) aparece una secuencia ATGA, mayoritariamente, o GTGA; el tercero es de mayor longitud y contiene el motivo catalítico DDE de la transposasa. La proteína codificada por *tnpA* muestra similitud con el ORFA de ISs pertenecientes a la familia IS*3* y contiene un dominio de unión al DNA HTH (Schneiker *et al*, 1999). La secuencia de inserción descrita en esta Tesis Doctoral (IS*Sme3*) también presenta esas características típicas de la familia IS*66* (dominio HTH, motivo catalítico DDE), con alguna excepción como el número de ORFs. Prácticamente todos los elementos de la familia IS*66* presentan un solapamiento en el marco de lectura al final de *tnpA* e inicio de *tnpB*. Entre estos ORFs, IS*Sme3* presenta la secuencia minoritaria GTGA, que sólo se encuentra en las dos ISs con las que hemos relacionado esta nueva IS*Sme3* estructuralmente (IS*Rsp1*) y respecto a secuencia (IS*693*; figura R3.2). La transposasa de IS*Rm2011-2* se produce gracias a un mecanismo de cambio en el marco de lectura durante su traducción (Selbitschka *et al*, 1995), sin embargo, los elementos de la familia IS*66* no presentan la estructura necesaria para ese cambio en el marco de lectura. Por este motivo, se ha sugerido que se producen las tres proteínas de manera independiente en proporciones adecuadas y forman un complejo que actúa como transposasa para promover la transposición (Han *et al*, 2001).

Como la mayoría de miembros de la familia IS*66* que contienen dos IRs, IS*Sme3* está flanqueada por secuencias repetidas directas de 8 pb, lo cual es indicativo de que ese elemento realiza el mecanismo de transposición y, como consecuencia, duplica una secuencia de 8 pb en el sitio diana. Se ha descrito que los tres ORFs que contiene IS*679* son esenciales para su transposición (Han *et al*, 2001), sin embargo, nosotros no hemos analizado experimentalmente esa característica en IS*Sme3*. La distribución de esta nueva IS es amplia dentro de los rizobios y, al igual que IS*679*, presenta diferentes DRs en cada sitio de inserción (tabla R3.1). Han *et al* (2001) proponen una secuencia terminal consenso de 7 pb (5'-GTAAGCG-3') derivada de las IRs de los elementos que componen la familia IS*66*. La secuencia de inserción IS*Sme3* presenta mayor identidad con IS*693*, y al igual que ella, sus IRs no comparten todos los nucleótidos con la secuencia consenso propuesta, sino que la posición 3 está ocupada por una G en vez de por una A.

Dentro de un genoma bacteriano las secuencias de inserción pueden localizarse en la cadena sentido o antisentido del DNA y, a su vez, dependiendo de su localización respecto al origen de replicación, en la hebra usada como molde para la síntesis de la cadena retardada (LAG) o líder (LEAD) en la horquilla de replicación. En *S. meliloti* 1021 sólo encontramos tres copias de RmInt1, todas ellas en orientación LAG: una en pSymA, insertada en ISa2, y dos en pSymB, una invadiendo ISb2 y otra dentro de la única copia de IS*Rm10-1* (Toro *et al*, 2003; Nisa-Martínez *et al*, 2007). Gracias a la secuenciación de la cepa GR4, hemos podido localizar y analizar las copias de RmInt1 presentes en ella (figura R3.3A). A diferencia de 1021, GR4 contiene 4 copias de intrón en el cromosoma, además de otras 5 copias en pRmeGR4c y 1 en pRmeGR4b; sólo 3 de estas 10 copias se encuentran en LEAD. Se ha demostrado que RmInt1 utiliza preferentemente la cadena retardada emergente durante la replicación como cebador para la reverso-transcripción del intrón (Martínez-Abarca *et al*, 2004). Tras analizar detalladamente las copias genómicas de RmInt2 y SmedInt1 (figura R3.3), se observa que éstos también se insertan mayoritariamente en la hebra que sirve de molde para la síntesis de la cadena retardada. Incluso RmInt2, que tiene la oportunidad

de invadir una IR en cada orientación, presenta mayor número de copias insertadas en una diana con orientación LAG.

RmInt1, como todo intrón del grupo II, presenta una estructura secundaria compuesta por seis dominios entre los que existen interacciones terciarias. El dominio IV presenta un ORF, que codifica una proteína que contiene los dominios RT y madurasa, y una región C-terminal carente de dominio endonucleasa (Martínez-Abarca *et al*, 1998). RmInt2 y SmedInt1 se consideran intrones tipo RmInt1, por lo que comparten la estructura secundaria y codifican una proteína que presenta los mismos dominios que la codificada por RmInt1. Ambos han acumulado mutaciones en su secuencia con respecto a la de RmInt1, que afectan al RNA del intrón y a la proteína. Sin embargo, mantienen los dominios más conservados. En el caso de la IEP (figura R3.6), se conserva el motivo catalítico (RYADD) para la actividad reverso transciptasa que reside en la región 5 del dominio RT (Matsuura *et al*, 1997) y los dos aminoácidos de tirosina (Y354 e Y355) en el dominio madurasa imprescindibles para la escisión del intrón (Molina-Sánchez *et al*, 2006). Además, los tres intrones comparten el motivo de la clase D [$LX_3AX_3PXLF(V/A)HW$] descrito por Molina-Sánchez *et al* (2010), que se encuentra entre los residuos 396-410. La región C-terminal contribuye a la función madurasa de la proteína, siendo el aminoácido más crítico para la escisión de RmInt1 el H409 (Molina-Sánchez *et al*, 2010), que se conserva en los nuevos intrones.

En cuanto a la estructura primaria y secundaria del RNA de los tres intrones (figura R3.4), observamos que RmInt2 presenta mayor número de mutaciones que SmedInt1 con respecto a la secuencia de RmInt1, aunque en todos los casos se mantiene la triada catalítica (AGC) y la A protuberante a siete nucleótidos del final de intrón, esenciales para que ocurra la reacción de escisión (Peebles *et al*, 1995; Molina-Sánchez *et al*, 2011). Con respecto al dominio VI, Molina-Sánchez (2008) realizó un estudio sobre la estructura que presenta esta región de la ribozima en los intrones del grupo IIB. Analizó 107 intrones presentes en la base de datos creada por Zimmerly y colaboradores (Dai *et al*, 2003), y describió una estructura consenso a

todos ellos, que también es compartida por los intrones estudiados en este trabajo. Las mayores diferencias entre estos tres intrones son la longitud del bucle final, donde RmInt2 contiene 4 nt como la mayor parte de intrones IIB (65'4 %) y RmInt1 y SmedInt1 7 nt como el 20'5 % de éstos, y la presencia de regiones internas desapareadas en RmInt1 y SmedInt1 que, por lo general, no aparecen en este tipo de intrones (como se aprecia en RmInt2; figura R3.5).

Las interacciones IBS-EBS también se han visto modificadas debido a las mutaciones presentes en el RNA de los nuevos intrones y al cambio de la diana reconocida por éstos. En este trabajo hemos puesto de manifiesto que RmInt2 y SmedInt1 mantienen la capacidad de escisión y movilidad, por tanto, las interacciones IBS-EBS tienen que estar ocurriendo. La unión IBS1-EBS1 es esencial y suficiente para que se produzca la escisión de RmInt1 (Barrientos-Durán *et al*, 2011), mientras que los apareamientos IBS-EBS 2 y 3 son prescindibles para la escisión pero necesarios durante el proceso de movilidad de estos intrones (Jiménez-Zurdo *et al*, 2003; Barrientos-Durán *et al*, 2011). Hemos definido la interacción IBS2-EBS2 para los nuevos intrones en base a lo que se conoce para RmInt1, pero no la hemos comprobado experimentalmente. Por esta razón, no podemos descartar que la secuencia correspondiente a la EBS2 se haya movido con respecto a las posiciones descritas en RmInt1 y sea diferente a la propuesta. No obstante, si la EBS2 de SmedInt1 incluyera la secuencia propuesta en este trabajo, entonces estaría formada por 4 nucleótidos y el *linker* por 2 (figura R3.7).

Jiménez-Zurdo *et al* (2003)han descrito los requerimientos mínimos de la diana para que tenga lugar el proceso de *homing* de RmInt1. Las posiciones críticas de las regiones distales, que no pueden ser compensadas con un cambio en el RNA del intrón, son la T -15, en mayor medida, y la G +4. Esto es indicativo de que la zona distal de los exones puede representar un sitio de contacto con la proteína. Además, cuando el resto de posiciones de la zona distal son mutadas individualmente, el intrón sigue presentando movilidad; sin embargo, si el exón 1 se acorta hasta la posición -15 la movilidad se reduce drásticamente (diana -15/+10 analizada en Jiménez-

Zurdo *et al*, 2003). Cambios de esta magnitud los encontramos de manera natural en la diana de RmInt2, donde la zona distal de los exones difiere dependiendo de su localización genómica (tabla R3.3). Este intrón ha sido capaz de invadir secuencias que sólo comparten las posiciones T -15 y C -16 en el exón 1, y las posiciones A +2 y C +3 en el exón 2. Por tanto, podría considerarse que una diana -16/+3 es el requisito mínimo para que se produzca la movilidad de RmInt2, aunque para una elevada eficiencia de *homing* sea necesario un cierto grado de conservación de las posiciones -20 a -17, como se ha sugerido para RmInt1 (Jiménez-Zurdo *et al*, 2003).

R3.9.1 Caracterización funcional de los nuevos intrones

Otra parte del capítulo ha estado centrada en los procesos de escisión y movilidad que llevan a cabo este tipo de intrones. Durante más de una década se ha estado caracterizando el intrón del grupo II RmInt1, tanto estructuralmente como en los procesos de escisión y movilidad que realiza (Martínez-Abarca *et al*, 2000; Nisa-Martínez *et al*, 2007; Molina-Sánchez *et al*, 2010; Chillón *et al*, 2011). El descubrimiento de dos nuevos intrones muy relacionados filogenéticamente con RmInt1, pero diferentes en cuanto a su secuencia diana, nos llevó a aplicar sobre éstos los conocimientos acerca de la escisión y movilidad aprendidos con RmInt1.

Los intrones del grupo II se escinden a través de un intermediario en forma de lazo (Lambowitz & Zimmerly, 2011), aunque se ha descrito que RmInt1 también se escinde en forma de círculos (Molina-Sánchez *et al*, 2006). Para estudiar la escisión de RmInt2 y SmedInt1 llevamos a cabo un ensayo de extensión a partir de cebador. Mediante esta técnica no se pueden diferenciar los productos escindidos a partir de una molécula de intrón lineal y de una en forma de lazo, ya que el tamaño de banda de ambos productos es el mismo; sin embargo, sí se puede detectar la escisión de una molécula de intrón circular, que daría un producto de un nucleótido superior. Esta diferencia de tamaños se debe a la reverso transcriptasa usada en el ensayo, la AMV RT, que no es capaz de leer más allá del enlace producido en las

moléculas en forma de lazo, y se para antes de éste, pero sí atraviesa la unión de las moléculas circulares y la pausa tiene lugar tras dicha unión. Así pues, las bandas de 97 nt y 105 nt presentes en el gel de la figura R3.8B corresponden a productos escindidos a partir de moléculas en forma de lazo o lineal de los intrones SmedInt1 y RmInt2 respectivamente. No obstante, mediante un experimento de RACE 5′ se ha demostrado que la banda de 97 nt detectada en los análisis de extensión a partir de cebador realizados con RmInt1 no se corresponde con formas lineales del intrón (Barrientos-Durán, 2008). Otra técnica, la RT-PCR, ha permitido comprobar que la banda de 98 nt, de mucha menor intensidad que la de 97 nt, encontrada en los ensayos de extensión de cebador llevados a cabo con RmInt1 corresponde al producto escindido de moléculas circulares de dicho intrón (Molina-Sánchez *et al*, 2006). Sin embargo, en ninguno de nuestros ensayos de extensión a partir de cebador encontramos una banda de 98 nt, en el caso de SmedInt1, o de 106 nt, en el caso de RmInt2. Estos resultados sugieren que, a pesar de considerarse intrones tipo RmInt1, en la escisión de estos nuevos intrones no se detectan moléculas circulares. Aunque no podemos descartar que se estén formando en una proporción más baja que en el caso de RmInt1.

A pesar de que el proceso de movilidad de este tipo de intrones requiere el apareamiento de bases IBS2-EBS2 (Jiménez-Zurdo *et al*, 2003), y que en RmInt2 y SmedInt1 dicho apareamiento no es total, se han detectado eventos de *homing* en ambos intrones (figura R3.8D). Conociendo los niveles de *homing* que presenta RmInt1 en su forma silvestre (Martínez-Abarca *et al*, 2004), observamos que esa forma del intrón RmInt2 muestra un porcentaje de movilidad sobre la diana en orientación LAG (68 %) superior a la mostrada por SmedInt1 y RmInt1, con un 51 % y 52 % respectivamente; y que los nuevos intrones se mueven más eficientemente sobre la diana en LEAD (28 % para RmInt2 y 35 % para SmedInt1) que RmInt1 (12 %). Esa mejor movilidad de los nuevos intrones puede explicarse por la distinta longitud de los exones que flanquean al intrón en el plásmido usado como donador. Los plásmidos donadores de RmInt2 y SmedInt1 generados en esta Tesis Doctoral contienen una diana -20/+5 (exones cortos), mientras que el plás-

mido donador de RmInt1, pKG2.5 (Martínez-Abarca *et al*, 2000), incluye una diana con exones largos (-175/+466). Estudios posteriores muestran un ligero aumento en el porcentaje de *homing* de RmInt1 (57 % sobre una diana en LAG) al usar un plásmido donador que contiene una diana -50/+146 (Nisa-Martínez *et al*, 2007). No se ha analizado la movilidad de este intrón desde una construcción con exones cortos (-20/+5), pero se ha demostrado que la escisión *in vitro* se ve afectada por la longitud de los exones, siendo más eficiente en construcciones -15/+1 (Costa *et al*, 2006).

En las muestras de RmInt2 aparecen dos bandas adicionales a las encontradas en SmedInt1. La hipótesis más plausible para explicar su aparición se basa en el reestablecimiento de la secuencia diana dentro del plásmido donador (D) y del plásmido correspondiente al producto de *homing* (H). Estos plásmidos contienen una copia de RmInt2, que permite que la G +1 del intrón actúe como IBS3 y se una al U presente en la EBS3 de dicho intrón. Por tanto, en ellos puede darse un evento de invasión de la nueva diana constituida. Este reestablecimiento de la diana no ocurre cuando el plásmido usado en el ensayo contiene una copia de RmInt1 o SmedInt1, ya que en su EBS3 presentan una G (que no interacciona con la G +1 del intrón). No obstante, se ha demostrado la existencia de dímeros de intrón al analizar mediante RT-PCR y qPCR la escisión de RmInt1 desde un plásmido que contiene una EBS3 mutada a una C, donde puede producirse la unión IBS3-EBS3 (Molina-Sánchez *et al*, 2011).

La hipótesis propuesta por Toor *et al* (2001), apoyada por Simon *et al* (2009), de una coevolución entre la ribozima y la proteína de los intrones del grupo II, nos llevó a generar construcciones quimeras que contienen el ΔORF y la IEP de intrones distintos. Con esto pretendíamos verificar la existencia de una correlación funcional en ese modelo evolutivo. A priori, era de esperar que las construcciones compuestas por la proteína y la ribozima del mismo intrón tuvieran un porcentaje de escisión determinado (que nosotros consideramos como el 100 %), y que, al cambiar de proteína, la ribozima presentara niveles menores de escisión. Este supuesto se cumple para RmInt1 y SmedInt1 (carriles 7-9 y 4-6 respectivamente en la figura

R3.10). Sin embargo, para RmInt2 ocurre todo lo contrario (carriles 1-3 en la figura R3.10). Resulta sorprendente que la IEP R1 complemente a la ribozima R2 mejor que su proteína natural, produciendo una escisión de ésta casi cuatro veces superior a la escisión generada por la IEP R2. Este efecto se observa también en la combinación S1-R2, aunque no de manera tan acusada, ya que el nivel de escisión de la ribozima R2 originado en dicha quimera no llega al doble del producido en la construcción con ambos componentes procedentes de RmInt2 (R2-R2). No obstante, la ribozima S1 no presenta mayor porcentaje de escisión cuando es complementada por la proteína R1 en vez de por la suya propia como le ocurre al ΔORF R2. Además, podemos decir que la proteína R2 es la que peor complementa a todas las ribozimas, incluida la suya, ya que contribuye a una escisión ínfima de éstas. Esto puede deberse a los cambios producidos en el dominio madurasa de la IEP (figura R3.6), ya que se ha demostrado que RmInt1 necesita la actividad de dicho dominio para llevar a cabo su escisión (Molina-Sánchez *et al*, 2006).

Sabiendo que existe una correspondencia entre los valores de escisión y movilidad de RmInt1 (Barrientos-Durán *et al*, 2011), y que todas las construcciones ΔORF, en mayor o menor medida, presentaron escisión de la ribozima que portan, era de esperar que todas mostraran eventos de *homing*. Sin embargo, sólo cuatro de ellas mostraron un porcentaje de movilidad sobre la diana en orientación LAG superior al 40 %. Tres de estas cuatro muestras son los intrones que contienen ambos componentes, ribozima y proteína, del mismo intrón. Hay que destacar que la quimera S1-R1 (carril 6 en la figura R3.10) no muestra correspondencia entre sus niveles de escisión y movilidad aunque presenta una mayor eficiencia de escisión que su quimera inversa R1-S1. Pero lo más llamativo de la gráfica mostrada en la figura R3.10 es que, a pesar del gran porcentaje de producto de escisión presentado por la quimera R1-R2, esa ribozima R2 escindida no es capaz de invadir su diana (carril 3 en la figura R3.10). Con el fin de corroborar la participación de la proteína en el reconocimiento de la diana, planteamos un ensayo de movilidad de la quimera R1-R2 sobre dos dianas quiméricas distintas (figura R3.11B). El resultado de esos ensayos de movilidad (figura

R3.11C) indica que realmente la proteína forma parte del reconocimiento de la diana en la zona distal del exón 1, del exón 2, o de ambos a la vez. Incluso la adición de un nucleótido en la zona distal del exón 1 reduce casi totalmente la movilidad del intrón al variar el nucleótido presente en el resto de posiciones. Este dato sugiere que el exón 1 es el que debe estar influyendo en mayor medida en la interacción con la proteína. La eficiencia con la que se mueve la ribozima R2 sobre la diana quimérica -20/+5 en LAG en presencia de la IEP R1 (27 %) no alcanza los valores de *homing* mostrados por la forma ΔORF derivada de RmInt2 (69 %), lo cual indica que la interacción de la proteína con los exones no es el único requisito para que se produzca la movilidad del intrón. De hecho, la proteína interviene en la estabilización del RNA escindido mediante la formación de complejos ribonucleoproteicos (RNP; Huang *et al*, 2011), por lo que una asociación incorrecta entre estas dos moléculas puede disminuir los niveles de invasión de la diana.

En este capítulo hemos analizado la secuencia primaria y secundaria de los tres intrones objeto de estudio. Al compararlos hemos descubierto mutaciones en sus ribozimas así como en sus proteínas, siendo la zona más diferente entre éstas la región C-terminal. Siguiendo esta premisa, y basándonos en la hipótesis de que la proteína puede estar jugando un papel importante en el reconocimiento de la diana durante la movilidad del intrón, creamos una construcción que contiene la ribozima R2 y una IEP quimérica entre la de RmInt1 y RmInt2 (quimera IEPQ). Los resultados obtenidos de los ensayos llevados a cabo con esta nueva construcción apuntan a que la proteína quimérica generada presenta capacidad funcional en la escisión pero no en la movilidad de la ribozima. Estudios posteriores con quimeras que abarquen regiones más precisas de los distintos dominios de la IEP ayudarán a resolver los aminoácidos de la proteína implicados en el reconocimiento de la diana, como ha ocurrido con la IEP del intrón Ll.ltrB (LtrA), en la que se han descrito seis zonas (A-F) dentro de los dominios RT y madurasa dependiendo de la región del RNA con la que interaccionan (Gu *et al*, 2010).

La ausencia de movilidad en la quimera IEPQ podría tener su explicación en la interacción que existe entre uno de los subdominios de la ribozima (dIVa) y la IEP. Se ha demostrado que el subdominio dIVa del intrón Ll.LtrB presenta un sitio de alta afinidad de unión a la IEP (Wank *et al*, 1999; Watanabe & Lambowitz, 2004); y análisis realizados con LtrA sugieren que la región N-terminal de su dominio RT interacciona con ese sitio de alta afinidad del dIVa (Cui *et al*, 2004). Más tarde, Nisa-Martínez *et al* (2007) descri-

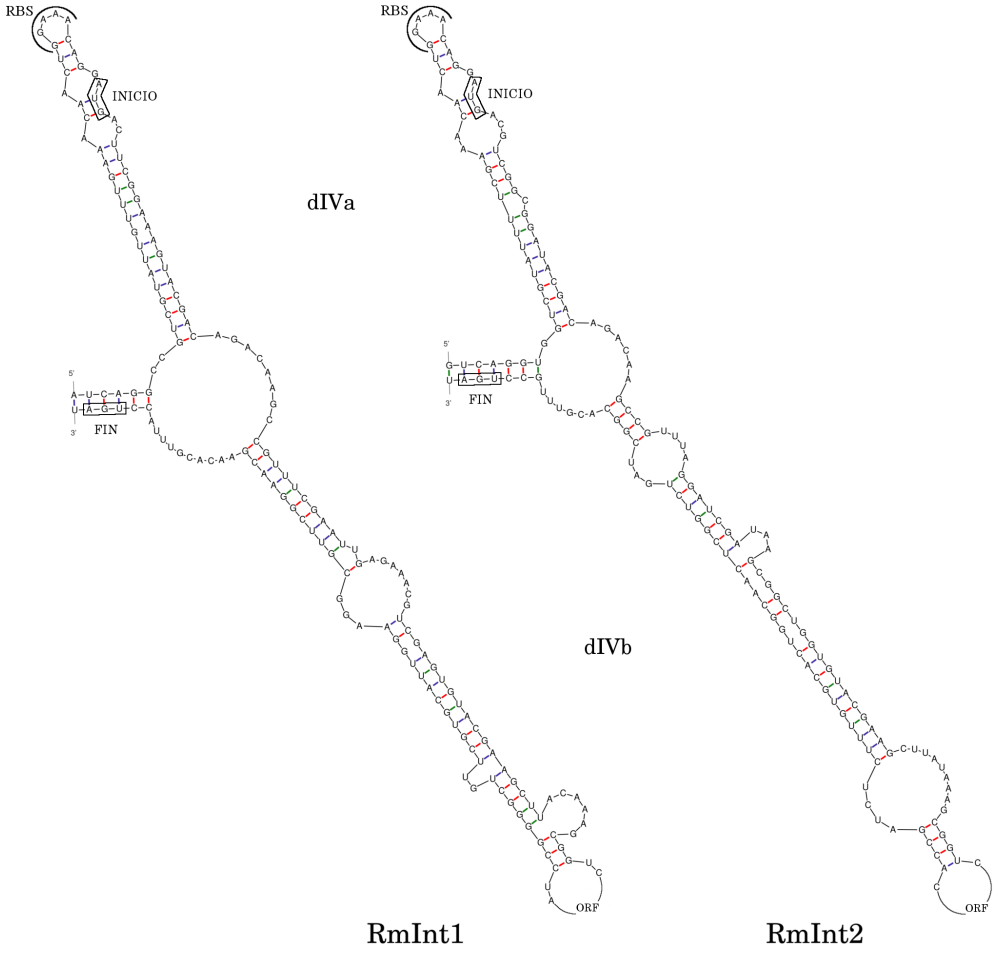


Figura R3.14: **Comparación de la estructura secundaria predicha de los subdominios IVa y IVb de RmInt1 y RmInt2.** En un cuadro se enmarcan los nucleótidos de inicio y fin de la proteína codificada por el DIV.

bieron los límites del subdominio dIVa dentro del intrón RmInt1 y mostraron su importancia en el proceso de *homing* de dicho intrón. En este trabajo hemos analizado *in silico* la estructura secundaria de los subdominios IVa y IVb de RmInt2 y la hemos comparado con la estructura que presentan dichos subdominios en RmInt1 (figura R3.14). Esta comparación pone de manifiesto que el dIVa de ambos intrones mantiene la misma estructura secundaria a pesar de su diferente estructura primaria. Esto sugiere que, aunque la región de la IEP R1 que se una al dIVa sea similar en ambos intrones, dicha unión puede verse afectada por la variación nucleotídica que presenta el subdominio dIVa en la ribozima R2 con respecto a RmInt1. Por otro lado, analizando el subdominio dIVb, observamos que en RmInt2 éste presenta un bucle en una zona más adelantada que donde lo presenta RmInt1. Ese tallo inicial en el dIVb del que RmInt2 carece es de gran importancia en la movilidad de RmInt1 (Nisa-Martínez *et al*, 2007). Por tanto, dicha zona podría explicar la falta de movilidad de la ribozima R2 observada.

En *E. coli* se han identificado varios intrones del grupo II que se encuentran dispersos por un gran número de poblaciones de esta bacteria (Ferat *et al*, 1994; Dai & Zimmerly, 2002). Además de la movilidad de estos intrones (Zhuang *et al*, 2009), se ha descrito que el intrón Ll.ltrB presenta movilidad en *E. coli* (Cousineau *et al*, 1998). En la cepa *E. coli* HMS174 (DE3), RmInt1 en su forma derivada ΔORF muestra una eficiencia de movilidad del 10'4 % sobre una diana en orientación LAG (García-Rodríguez *et al*, 2011), un valor inferior al mostrado en la cepa *E. coli* DH5α (27 %). RmInt2 presenta un porcentaje de invasión mayor que RmInt1, tanto en su forma de intrón silvestre como derivada (41 % y 34 % respectivamente). Este resultado pone de manifiesto que RmInt2 presenta una mayor eficiencia de movilidad que RmInt1 en el fondo genético de *E. coli*, por lo que supondría un candidato alternativo para ser utilizado como herramienta biotecnológica (Perutka *et al*, 2004; Yao *et al*, 2005). Además, el estudio de diseño de intrones programados para invadir secuencias de interés llevado a cabo

con RmInt1 y el desarrollo de un algoritmo para encontrar posibles dianas puede extrapolarse a este nuevo intrón (García-Rodríguez *et al*, 2011).

Los estudios presentados en este trabajo muestran el proceso evolutivo de los intrones del grupo II sobre un mismo hospedador. La secuenciación del genoma completo de *S. meliloti* GR4 reveló que RmInt1 y RmInt2 conviven en esa cepa, así como RmInt1 y SmedInt1 conviven en *S. medicae* RMO09 (capítulos 1 y 2). Estos tres intrones, estrechamente relacionados, son un ejemplo de cómo un cambio en la diana afecta a la expansión y distribución de este tipo de elementos. Una evidencia de ello es la similitud que muestran las dianas donde se han encontrado los intrones RmInt1 y SmedInt1 en distintas localizaciones (tabla R3.5). Ocasionalmente, RmInt1 se encuentra insertado en la diana IS*Rm10-1* (tabla R2.11), que comparte la mayoría de nucleótidos del exón 1 y 2 con IS*Rm2011-2* salvo la posición -2 (donde contiene una A que sigue apareando con la T del intrón) y parte de la zona distal de los exones. Los nucleótidos que sí se mantienen entre ambas dianas, y que son importantes para la movilidad de RmInt1 (Jiménez-Zurdo *et al*, 2003), son la T -15 y la G +4. Las mismas características de IS*Rm10-1* frente a IS*Rm2011-2* las presenta una de las dianas donde se ha encontrado SmedInt1, la del plásmido pSMED01 en la que interrumpe al gen Smed_4905 (tabla R3.2). El intrón presente en la cepa *S. fredii* NGR234 (con un 98 % de identidad con RmInt1; tabla R2.2) podría ser una forma interme-

Tabla R3.5: Distintas secuencias diana donde se han encontrado intrones tipo RmInt1. Las IBS 1 y 2, en el exón 1, y la IBS3, en el exón 2, aparecen subrayadas. También se muestran las EBSs del intrón en cada caso.

Diana	**Exón 1**	**Exón 2**	**EBS2**	**EBS1**	**EBS3**
IS*Rm2011-2*	CCTCGTTTTCATCGATGAGA	CCTGG	ATGAA	TTTCGTC	G
IS*Rm10-1*	GCTGATCTTCATCGATGAAA	CCGGC	ATGAA	TTTCGTC	G
Smed_4905	GGAAGTGTTCATGGATGAAA	CGCGT	AAGAA	TTTCGTC	G
IS*Sme3*	AAGGGTGTTCTGTGACGAGA	CGCCG	AAGAA	TTTCGTC	G
NGRIS-7*	ACTGGTCTTTATCGACGAGA	CCTGG	ATGAA	TTTCGTC	G

* IS donde se ha encontrado insertado el intrón RmInt1 presente en la cepa *S. fredii* NGR234

dia entre RmInt1 y SmedInt1, puesto que la diana que invade (la secuencia de inserción NGRIS-7, perteneciente a la misma familia que IS*Rm2011-2*), presenta unas regiones similares a IS*Rm2011-2* y otras a IS*Sme3* (tabla R3.5). Además, sería interesante determinar la diana que ha invadido el intrón encontrado en *E. adhaerens* 5D19 (un 93 % idéntico a RmInt1; tabla R2.2), ya que podría contener una secuencia intermedia entre la diana NGRIS-7 e IS*Sme3*.

Debido a que las interacciones IBS-EBS en RmInt1 y SmedInt1 son equivalentes, se puede decir que el proceso para generar nuevos intrones primero conlleva un cambio en la secuencia del sitio de unión a la proteína (exones distales), y posteriormente ocurren mutaciones en el resto de la diana (acompañadas con cambios en la secuencia del intrón). Esto sugiere, al igual que mostraban las relaciones filogenéticas (figura R2.1), que la diferenciación del intrón RmInt2 ha sido anterior a la diferenciación de SmedInt1, ya que RmInt2 muestra mutaciones tanto en la zona distal de los exones como en las interacciones IBS-EBS (figura R3.7).

CONCLUSIONS

1. Genome sequencing based on a 77x coverage direct shotgun and a 17x coverage 3 kb paired end libraries is a successful strategy to obtain the complete sequence of a multireplicon bacterial genome.

2. Fold Coverage Index (FCI) has been a key tool for assigning contigs to replicons as well as for identifying repeat elements and determining their copy number in the genome.

3. The 7.14 Mb *Sinorhizobium meliloti* GR4 genome comprises five replicons: one chromosome of 3.6 Mb length, two accessory plasmids of 176 kb (pRmeGR4a) and 226 kb (pRmeGR4b), and two symbiotic plasmids of 1.4 Mb (pRmeGR4c) and 1.7 Mb (pRmeGR4d).

4. The group II intron RmInt1 is widely spread within the sequenced *S. meliloti* species, however, occurrence of group II intron RmInt2 is restricted to the genome it was discovered (*S. meliloti* GR4) despite of the widespread distribution of both target IS elements (IS*Rm2011-2* and IS*Rm17*). These evidences suggest a recent acquisition of RmInt2 by *S. meliloti* species.

5. Deletion of *groEL3* gene in *S. meliloti* AF14 mutant does not explain the observed RmInt1 excision decrease.

6. The two RmInt1 phylogenetically related introns, RmInt2 and SmedInt1, are functional. Like RmInt1, their mobility presents a bias for targets in the template for lagging strand DNA synthesis.

7. Excision and mobility performance of chimeric introns combining different IEPs and ribozymes from the studied introns (RmInt1, RmInt2 and SmedInt1) provide evidences of the coevolution between these two group II intron elements (IEP and RNA) as suggested by their phylogenetic relationships.

8. Genetic assays with a chimeric intron combining the RmInt1 IEP and the RmInt2 ribozyme indicate that RmInt1 phylogenetically related group II introns recognize the inner region of their target sites by forming IBS-EBS interactions by the ribozyme while the IEP recognizes the distal zone of exons 1 and 2.

9. RmInt2 intron shows mobility in the heterologous host *Escherichia coli*, becoming an additional group II intron to be used as a gene-targeting vector for genetic enginering.

APÉNDICE

genomeA
Journals.ASM.org

Complete Genome Sequence of the Alfalfa Symbiont *Sinorhizobium/Ensifer meliloti* Strain GR4

Francisco Martínez-Abarca, Laura Martínez-Rodríguez, José Antonio López-Contreras, José Ignacio Jiménez-Zurdo, Nicolás Toro

Grupo de Ecología Genética de la Rizosfera, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas (CSIC), Granada, Spain

We present the complete nucleotide sequence of the multipartite genome of *Sinorhizobium/Ensifer meliloti* GR4, a predominant rhizobial strain in an agricultural field site. The genome (total size, 7.14 Mb) consists of five replicons: one chromosome, two expected symbiotic megaplasmids (pRmeGR4c and pRmeGR4d), and two accessory plasmids (pRmeGR4a and pRmeGR4b).

Received 5 December 2012 **Accepted** 18 December 2012 **Published** 14 February 2013

Citation Martínez-Abarca F, Martínez-Rodríguez L, López-Contreras JA, Jiménez-Zurdo JI, Toro N. 2013. Complete genome sequence of the alfalfa symbiont *Sinorhizobium/Ensifer meliloti* strain GR4. Genome Announc. 1(1):e00174-12. doi:10.1128/genomeA.00174-12.

Address correspondence to Nicolás Toro, nicolas.toro@eez.csic.es.

The primary source of biologically fixed nitrogen in crops is found in the symbiotic interaction between legume plants and certain soil microorganisms, collectively referred to as rhizobia (1). *Sinorhizobium* bacteria are the microsymbionts of *Medicago* (e.g., *Medicago sativa* and *Medicago truncatula*), *Melilotus*, and *Trigonella* legume species.

The genome sequences of four *Sinorhizobium meliloti* strains are publicly available (2–4). *S. meliloti* GR4 was isolated as the predominant rhizobial strain (i.e., nearly 50% of the isolates) from an agricultural field with a well-documented crop history in Granada, Spain (5, 6). Besides the chromosome and the expected symbiotic megaplasmids (pRmeGR4c and pRmeGR4d), this strain harbors two accessory plasmids, designated pRmeGR4a and pRmeGR4b. The latter contains a region that was identified as the genetic determinant of the particularly high competitiveness for nodulation on alfalfa roots along with a plethora of mobile genetic elements (7–10).

A highly pure genomic DNA sample of *S. meliloti* GR4 was sequenced on a GS FLX Titanium platform (Roche Diagnostics) at Macrogen, Inc. (South Korea), on the basis of both shotgun and 3-kb paired-end libraries, resulting in 70-fold genome coverage. Raw sequence data fit the quality standards of the Genomes On-Line Database (GOLD) (11). Sequencing reads were *de novo* assembled (Newbler 3.0), resulting in a total of 12 scaffolds (>40 kb each) and 51 contigs (<3 kb each). Most of the gaps (78%) were closed using customized informatics scripts (L. Martínez-Rodríguez, J. A. López-Contreras, F. Martínez-Abarca, and N. Toro, unpublished data). The remaining gaps (except two, corresponding to repeated sequences) were manually closed by combining Southern blot hybridization data and a detailed observation of relevant sequencing reads with the Tablet tool (http://bioinf.scri.ac.uk/tablet/). The genome was annotated using the Integrated Microbial Genomes (IMG) Expert Review (ER) service (12). Replicon sizes and the G+C content of the chromosome and plasmids pRmeGR4a, pRmeGR4b, pRmeGR4c (related to pSymA), and pRmeGR4d (related to pSymB) are 3,618,794 bp (62.8%), 175,986 bp (60.0%), 225,725 bp (58.6%), 1,417,856 bp (60.4%), and 1,701,197 bp (62.4%), respectively. The complete genome consists of 6,700 protein-coding sequences: 3,334 on the chromosome, 1,541 on pRmeGR4d, 1,393 on pRmeGR4c, 247 on pRmeGR4b, and 185 on pRmeGR4a. Similarly to other *S. meliloti* genomes, 3 *rrn* chromosomal operons and 55 tRNA loci (52 on the chromosome and 3 on plasmids) were identified. In addition, 1,066 noncoding RNA genes (sRNAs) were predicted in this genome based on those identified in the *S. meliloti* 1021 and 2011 reference strains (12, 13). Genome comparisons using the MUMmer package (14) revealed a high degree of synteny of the chromosome and the largest plasmid (pRmeGR4d) to the corresponding replicons of the other four sequenced *S. meliloti* strains. This synteny is less pronounced in the symbiotic megaplasmid pRmeGR4c. The two smaller plasmids, pRmeGR4a and pRmeGR4b, did not evidence signs of synteny with any rhizobial genomic region, which, together with their low G+C content, suggests that horizontal transfer has been the major contribution to the mosaic arrangement of these accessory replicons.

Nucleotide sequence accession numbers. The accession no. for GR4 chromosome, pRmeGR4a, pRmeGR4b, pRmeGR4c, and pRmeGR4d are CP003933, CP003934, CP003935, CP003936, and CP003937, respectively.

ACKNOWLEDGMENTS

This work was supported by research grants BIO2011-24401 from the Spanish Ministerio de Ciencia e Innovación and CSD 2009-0006 of Programme Consolider-Ingenio, both including ERDF (European Regional Development Funds).

L.M.-R. and J.A.L.-C. were supported by predoctoral fellowships (Programs JAE-Predoc and I3P from Consejo Superior de Investigaciones Científicas, respectively). We are particularly grateful to M. G. Claros and R. Bautista from Plataforma Andaluza de Bioinformática (Universidad de Málaga) Spain, to A. J. Fernández-González for bioinformatics support, and to the Estación Experimental del Zaidín—CSIC for DNA sequencing services.

REFERENCES

1. **Sahgal M, Johri BN.** 2003. Taxonomy of rhizobia: current status. Curr. Sci. **90**:486–487.
2. **Galibert F, Finan TM, Long SR, Puhler A, Abola P, Ampe F, Barloy-**

Hubler F, Barnett MJ, Becker A, Boistard P, Bothe G, Boutry M, Bowser L, Buhrmester J, Cadieu E, Capela D, Chain P, Cowie A, Davis RW, Dreano S, Federspiel NA, Fisher RF, Gloux S, Godrie T, Goffeau A, Golding B, Gouzy J, Gurjal M, Hernandez-Lucas I, Hong A, Huizar L, Hyman RW, Jones T, Kahn D, Kahn ML, Kalman S, Keating DH, Kiss E, Komp C, Lelaure V, Masuy D, Palm C, Peck MC, Pohl TM, Portetelle D, Purnelle B, Ramsperger U, Surzycki R, Thebault P, Vandenbol M, Vorholter FJ, Weidner S, Wells DH, Wong K, Yeh KC, Batut J. 2001. The composite genome of the legume symbiont *Sinorhizobium meliloti*. Science **293**:668–672.

3. **Schneiker-Bekel S, Wibberg D, Bekel T, Blom J, Linke B, Neuweger H, Stiens M, Vorhölter FJ, Weidner S, Goesmann A, Pühler A, Schlüter A.** 2011. The complete genome sequence of the dominant *Sinorhizobium meliloti* field isolate SM11 extends the *S. meliloti* pan-genome. J. Biotechnol. **155**:20–33.
4. **Galardini M, Mengoni A, Brilli M, Pini F, Fioravanti A, Lucas S, Lapidus A, Cheng JF, Goodwin L, Pitluck S, Land M, Hauser L, Woyke T, Mikhailova N, Ivanova N, Daligault H, Bruce D, Detter C, Tapia R, Han C, Teshima H, Mocali S, Bazzicalupo M, Biondi EG.** 2011. Exploring the symbiotic pangenome of the nitrogen-fixing bacterium *Sinorhizobium meliloti*. BMC Genomics **12**:235.
5. **Casadesús J, Olivares J.** 1979. Rough and fine linkage mapping of the *Rhizobium meliloti* chromosome. Mol. Gen. Genet. **174**:203–209.
6. **Muñoz E, Villadas PJ, Toro N.** 2001. Ectopic transposition of a group II intron in natural bacterial populations. Mol. Microbiol. **41**:645–652.
7. **Soto MJ, Zorzano A, Olivares J, Toro N.** 1992. Nucleotide sequence of *Rhizobium meliloti* GR4 insertion sequence ISRm3 linked to the nodulation competitiveness locus *nfe*. Plant Mol. Biol. **20**:307–309.
8. **Soto MJ, Zorzano A, Olivares J, Toro N.** 1992. Sequence of ISRm4 from *Rhizobium meliloti* strain GR4. Gene **120**:125–126.
9. **Soto MJ, Zorzano A, Mercado-Blanco J, Lepek V, Olivares J, Toro N.** 1993. Nucleotide sequence and characterization of *Rhizobium meliloti* nodulation competitiveness genes *nfe*. J. Mol. Biol. **229**:570–576.
10. **Zekri S, Soto MJ, Toro N.** 1998. ISRm4-1 and ISRm9, two novel insertion sequences from *Sinorhizobium meliloti*. Gene **207**:93–96.
11. **Liolios K, Chen IM, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM, Kyrpides NC.** 2010. The genomes on line Database (gold) in 2009: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res. **38**:D346–D354.
12. **Markowitz VM, Mavromatis K, Ivanova NN, Chen IM, Chu K, Kyrpides NC.** 2009. IMG ER: a system for microbial genome annotation expert review and curation. Bioinformatics **25**:2271–2278.
13. **del Val C, Rivas E, Torres-Quesada O, Toro N, Jiménez-Zurdo JI.** 2007. Identification of differentially expressed small non-coding RNAs in the legume endosymbiont *Sinorhizobium meliloti* by comparative genomics. Mol. Microbiol. **66**:1080–1091.
14. **Schlüter JP, Reinkensmeier J, Daschkey S, Evguenieva-Hackenberg E, Janssen S, Jänicke S, Becker JD, Giegerich R, Becker A.** 2010. A genome-wide survey of sRNAs in the symbiotic nitrogen-fixing alphaproteobacterium *Sinorhizobium meliloti*. BMC Genomics **11**:245.

BIBLIOGRAFÍA

BIBLIOGRAFÍA

Abascal F, Zardoya R, Posada D (2005) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, **21**, 2104–2105.

Adessi C, Matton G, Ayala G, *et al* (2000) Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Research*, **28**, e87–e87.

Ansorge WJ (2009) Next-generation DNA sequencing techniques. *New Biotechnology*, **25**, 195–203.

Assefa S, Keane TM, Otto TD, Newbold C, Berriman M (2009) ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics*, **25**, 1968–1969.

Bakke P, Carney N, DeLoache W, *et al* (2009) Evaluation of Three Automated Genome Annotations for Halorhabdus utahensis. *PLoS ONE*, **4**, e6291.

Barloy-Hubler F, Jebbar M (2009) *Sinorhizobium meliloti* Megaplasmids and Symbiosis in *S. meliloti*. In *Microbial Megaplasmids* (edited by E Schwartz), no. 11 in Microbiology Monographs, pp. 91–118. Springer Berlin Heidelberg.

Barrientos-Durán A, Chillón I, Martínez-Abarca F, Toro N (2011) Exon sequence requirements for excision in vivo of the bacterial group II intron RmInt1. *BMC Molecular Biology*, **12**, 24.

Barrientos-Durán AM (2008) *Bases moleculares para la aplicación biotecnológica del intrón del grupo II RmInt1 de* Sinorhizobium meliloti. Ph.D. thesis, Granada.

Beauregard A, Curcio MJ, Belfort M (2008) The Take and Give Between Retrotransposable Elements and their Hosts. *Annual Review of Genetics*, **42**, 587–617.

Beckloff N, Starkenburg S, Freitas T, Chain P (2012) Bacterial Genome Annotation. In *Microbial Systems Biology* (edited by A Navid), no. 881 in Methods in Molecular Biology, pp. 471–503. Humana Press.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2011) GenBank. *Nucleic acids research*, **39**, D32–37.

Bentley DR, Balasubramanian S, Swerdlow HP, *et al* (2008) Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, **456**, 53–59.

Beringer JE (1974) R Factor Transfer in *Rhizobium leguminosarum*. *Journal of General Microbiology*, **84**, 188–198.

Bigot S, Saleh O, Lesterlin C, *et al* (2005) KOPS: DNA motifs that control *E. col*i chromosome segregation by orienting the FtsK translocase. *The EMBO journal*, **24**, 3770–3780.

Bigot S, Sivanathan V, Possoz C, Barre FX, Cornet F (2007) FtsK, a literate chromosome segregation machine. *Molecular Microbiology*, **64**, 1434–1441.

Biondi EG, Toro N, Bazzicalupo M, Martínez-Abarca F (2011) Spread of the group II intron RmInt1 and its insertion sequence target sites in the plant endosymbiont *Sinorhizobium meliloti*. *Mobile genetic elements*, **1**, 2–7.

Bittner AN, Foltz A, Oke V (2007) Only One of Five groEL Genes Is Required for Viability and Successful Symbiosis in *Sinorhizobium meliloti*. *Journal of Bacteriology*, **189**, 1884–1889.

Bittner AN, Oke V (2006) Multiple groESL Operons Are Not Key Targets of RpoH1 and RpoH2 in *Sinorhizobium meliloti*. *Journal of Bacteriology*, **188**, 3507–3515.

Boetzer M, Pirovano W (2012) Toward almost closed genomes with GapFiller. *Genome Biology*, **13**, R56.

Bonen L, Vogel J (2001) The ins and outs of group II introns. *Trends in Genetics*, **17**, 322–331.

Boorstein WR, Craig EA (1989) Primer extension analysis of RNA. In *Methods in Enzymology* (edited by JNA James E Dahlberg), vol. Volume 180, pp. 347–369. Academic Press.

Boudvillain M, de Lencastre A, Pyle AM (2000) A tertiary interaction that links active-site domains to the 5′ splice site of a group II intron. *Nature*, **406**, 315–318.

Boudvillain M, Marie Pyle A (1998) Defining functional groups, core structural features and inter-domain tertiary contacts essential for group II intron self-splicing: a NAIM analysis. *The EMBO Journal*, **17**, 7091–7104.

Cabanes D, Boistard P, Batut J (2000) Identification of *Sinorhizobium meliloti* Genes Regulated during Symbiosis. *Journal of Bacteriology*, **182**, 3632–3637.

Campbell V, Legendre P, Lapointe FJ (2011) The performance of the Congruence Among Distance Matrices (CADM) test in phylogenetic analysis. *BMC Evolutionary Biology*, **11**, 64.

Candales MA, Duong A, Hood KS, *et al* (2011) Database for bacterial group II introns. *Nucleic Acids Research*, **40**, D187–D190.

Capela D, Barloy-Hubler F, Gouzy J, *et al* (2001) Analysis of the chromosome sequence of the legume symbiont *Sinorhizobium meliloti* strain 1021. *Proceedings of the National Academy of Sciences*, **98**, 9877–9882.

Casadesús J, Olivares J (1979) Rough and fine linkage mapping of the *Rhizobium meliloti* chromosome. *Molecular & general genetics : MGG*, **174**, 203–209.

Cerveau N, Leclercq S, Bouchon D, Cordaux R (2011) Evolutionary Dynamics and Genomic Impact of Prokaryote Transposable Elements. In *Evolutionary Biology – Concepts, Biodiversity, Macroevolution and Genome Evolution* (edited by P Pontarotti), pp. 291–312. Springer Berlin Heidelberg.

Cevallos M, Cervantes-Rivera R, Gutiérrez-Ríos R (2008) The repABC plasmid family. *Plasmid*, **60**, 19–37.

Chillón I, Martínez-Abarca F, Toro N (2011) Splicing of the *Sinorhizobium meliloti* RmInt1 group II intron provides evidence of retroelement behavior. *Nucleic Acids Research*, **39**, 1095–1104.

Coros CJ, Landthaler M, Piazza CL, *et al* (2005) Retrotransposition strategies of the *Lactococcus lactis* Ll.LtrB group II intron are dictated by host identity and cellular environment. *Molecular Microbiology*, **56**, 509–524.

Coros CJ, Piazza CL, Chalamcharla VR, Smith D, Belfort M (2009) Global Regulators Orchestrate Group II Intron Retromobility. *Molecular Cell*, **34**, 250–256.

Costa M, Michel F, Toro N (2006) Potential for alternative intron-exon pairings in group II intron RmInt1 from *Sinorhizobium meliloti* and its relatives. *RNA (New York, N.Y.)*, **12**, 338–341.

Costa M, Michel F, Westhof E (2000) A three-dimensional perspective on exon binding by a group II self-splicing intron. *The EMBO journal*, **19**, 5007–5018.

Cousineau B, Lawrence S, Smith D, Belfort M (2000) Retrotransposition of a bacterial group II intron. *Nature*, **404**, 1018–1021.

Cousineau B, Smith D, Lawrence-Cavanagh S, *et al* (1998) Retrohoming of a Bacterial Group II Intron. *Cell*, **94**, 451–462.

Cui X, Davis G (2007) Mobile group II intron targeting: applications in prokaryotes and perspectives in eukaryotes. *Frontiers in bioscience: a journal and virtual library*, **12**, 4972–4985.

Cui X, Matsuura M, Wang Q, Ma H, Lambowitz AM (2004) A Group II Intron-encoded Maturase Functions Preferentially In Cis and Requires Both the Reverse Transcriptase and X Domains to Promote RNA Splicing. *Journal of Molecular Biology*, **340**, 211–231.

Dai L, Toor N, Olson R, Keeping A, Zimmerly S (2003) Database for mobile group II introns. *Nucleic Acids Research*, **31**, 424–426.

Dai L, Zimmerly S (2002) Compilation and analysis of group II intron insertions in bacterial genomes: evidence for retroelement behavior. *Nucleic Acids Research*, **30**, 1091–1102.

Darling AE, Mau B, Perna NT (2010) progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS ONE*, **5**, e11147.

Ditta G, Stanfield S, Corbin D, Helinski DR (1980) Broad host range DNA cloning system for gram-negative bacteria: construction of a gene bank of *Rhizobium meliloti*. *Proceedings of the National Academy of Sciences*, **77**, 7347–7351.

Dourado AC, Alves PIL, Tenreiro T, *et al* (2009) Identification of *Sinorhizobium* (*Ensifer*) medicae based on a specific genomic sequence unveiled by M13-PCR fingerprinting. *International microbiology: official journal of the Spanish Society for Microbiology*, **12**, 215–226.

Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences*, **100**, 8817–8822.

Drummond AJ, Ashton B, Cheung M, Heled J, Kearse M (2009) Geneious v4. 8. 2009. *Biomatters Ltd., Auckland, New Zealand*.

Duan J, Heikkila JJ, Glick BR (2010) Sequencing a bacterial genome: An overview. In *Current Research, Technology and Education Topics in Applied Microbiology and Microbial Biotechnology,*, vol. 2. Mendez-Vilas, A.

Eid J, Fehr A, Gray J, *et al* (2009) Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, **323**, 133–138.

Ettema TJG, Andersson SGE (2009) The α-proteobacteria: the Darwin finches of the bacterial world. *Biology Letters*, **5**, 429–432.

Fedorova O, Mitros T, Pyle AM (2003) Domains 2 and 3 Interact to Form Critical Elements of the Group II Intron Active Site. *Journal of Molecular Biology*, **330**, 197–209.

Fedorova O, Pyle AM (2005) Linking the group II intron catalytic domains: tertiary contacts and structural features of domain 3. *The EMBO Journal*, **24**, 3906–3916.

Fedorova O, Solem A, Pyle AM (2010) Protein-facilitated folding of group II intron ribozymes. *Journal of molecular biology*, **397**, 799–813.

Fedorova O, Zingler N (2007) Group II introns: structure, folding and splicing mechanism. *Biological Chemistry*, **388**.

Ferat JL, Le Gouar M, Michel F (1994) Multiple group II self-splicing introns in mobile DNA from *Escherichia coli*. *Comptes rendus de l'Académie des sciences. Série III, Sciences de la vie*, **317**, 141–148.

Ferat JL, Michel F (1993) Group II self-splicing introns in bacteria. *Nature*, **364**, 358–361.

Fernández-López M, Muñoz-Adelantado E, Gillis M, Willems A, Toro N (2005) Dispersal and Evolution of the *Sinorhizobium meliloti* Group II RmInt1 Intron in Bacteria that Interact with Plants. *Molecular Biology and Evolution*, **22**, 1518–1528.

Figurski DH, Helinski DR (1979) Replication of an origin-containing derivative of plasmid RK2 dependent on a plasmid function provided in trans. *Proceedings of the National Academy of Sciences*, **76**, 1648–1652.

Finan TM, Weidner S, Wong K, *et al* (2001) The complete sequence of the 1,683-kb pSymB megaplasmid from the N2-fixing endosymbiont *Sinorhi-*

zobium meliloti. *Proceedings of the National Academy of Sciences*, **98**, 9889–9894.

Finn RD, Mistry J, Schuster-Böckler B, *et al* (2006) Pfam: clans, web tools and services. *Nucleic Acids Research*, **34**, D247–D251.

Forde BM, O'Toole PW (2013) Next-generation sequencing technologies and their impact on microbial genomics. *Briefings in Functional Genomics*.

Galardini M, Biondi E, Bazzicalupo M, Mengoni A (2011a) CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. *Source code for biology and medicine*, **6**, 11.

Galardini M, Mengoni A, Brilli M, *et al* (2011b) Exploring the symbiotic pangenome of the nitrogen-fixing bacterium *Sinorhizobium meliloti*. *BMC genomics*, **12**, 235.

Galibert F, Finan TM, Long SR, *et al* (2001) The Composite Genome of the Legume Symbiont *Sinorhizobium meliloti*. *Science*, **293**, 668–672.

García-Rodríguez FM, Barrientos-Durán A, Díaz-Prado V, Fernández-López M, Toro N (2011) Use of RmInt1, a Group IIB Intron Lacking the Intron-Encoded Protein Endonuclease Domain, in Gene Targeting. *Applied and Environmental Microbiology*, **77**, 854–861.

García-Rodríguez FM, Toro N (2000) *Sinorhizobium meliloti* nfe (nodulation formation efficiency) genes exhibit temporal and spatial expression patterns similar to those of genes involved in symbiotic nitrogen fixation. *Molecular plant-microbe interactions: MPMI*, **13**, 583–591.

García-Rodríguez FM, Zekri S, Toro N (2000) Characterization of the *Sinorhizobium meliloti* genes encoding a functional dihydrodipicolinate synthase (dapA) and dihydrodipicolinate reductase (dapB). *Archives of microbiology*, **173**, 438–444.

Gil R, Latorre A (2012) Factors Behind Junk DNA in Bacteria. *Genes*, **3**, 634–650.

Gilbert W, Maxam A (1973) The Nucleotide Sequence of the lac Operator. *Proceedings of the National Academy of Sciences of the United States of America*, **70**, 3581–3584.

González V, Acosta JL, Santamaría RI, *et al* (2010) Conserved Symbiotic Plasmid DNA Sequences in the Multireplicon Pangenomic Structure of *Rhizobium etli*. *Applied and Environmental Microbiology*, **76**, 1604–1614.

González V, Santamaría RI, Bustos P, *et al* (2006) The partitioned *Rhizobium etli* genome: Genetic and metabolic redundancy in seven interacting replicons. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 3834–3839.

Gourbeyre E, Siguier P, Chandler M (2010) Route 66: investigations into the organisation and distribution of the IS66 family of prokaryotic insertion sequences. *Research in Microbiology*, **161**, 136–143.

Grigoriev A (1998) Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Research*, **26**, 2286–2290.

Gu SQ, Cui X, Mou S, Mohr S, Yao J, Lambowitz AM (2010) Genetic identification of potential RNA-binding regions in a group II intron-encoded reverse transcriptase. *RNA*, **16**, 732–747.

Guindon S, Gascuel O (2003) A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology*, **52**, 696–704.

Guo H, Karberg M, Long M, Jones JP, Sullenger B, Lambowitz AM (2000) Group II Introns Designed to Insert into Therapeutically Relevant DNA Target Sites in Human Cells. *Science*, **289**, 452–457.

Guo X, Flores M, Morales L, *et al* (2007) DNA Diversification in Two *Sinorhizobium* Species. *Journal of Bacteriology*, **189**, 6474–6476.

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series*.

Hamill S, Pyle AM (2006) The Receptor for Branch-Site Docking within a Group II Intron Active Site. *Molecular Cell*, **23**, 831–840.

Han CG, Shiga Y, Tobe T, Sasakawa C, Ohtsubo E (2001) Structural and Functional Characterization of IS679 and IS66-Family Elements. *Journal of Bacteriology*, **183**, 4296–4304.

Harris TD, Buzby PR, Babcock H, *et al* (2008) Single-Molecule DNA Sequencing of a Viral Genome. *Science*, **320**, 106–109.

Hernandez-Lucas I, Ramirez-Trujillo JA, Gaitan MA, *et al* (2006) Isolation and characterization of functional insertion sequences of rhizobia. *FEMS Microbiology Letters*, **261**, 25–31.

Higuchi R, Fockler C, Dollinger G, Watson R (1993) Kinetic PCR Analysis: Real-time Monitoring of DNA Amplification Reactions. *Nature Biotechnology*, **11**, 1026–1030.

van Hijum SAFT, Zomer AL, Kuipers OP, Kok J (2005) Projector 2: contig mapping for efficient gap-closure of prokaryotic genome sequence assemblies. *Nucleic acids research*, **33**, W560–566.

Huang T, Shaikh TR, Gupta K, *et al* (2011) The group II intron ribonucleoprotein precursor is a large, loosely packed structure. *Nucleic Acids Research*, **39**, 2845–2854.

Jacquier A, Michel F (1987) Multiple exon-binding sites in class II self-splicing introns. *Cell*, **50**, 17–29.

Jarvie T, Harkins T (2008) 3K Long-Tag Paired End sequencing with the Genome Sequencer FLX System. *Nature Methods*, **5**.

Jiang J, Li J, Kwan HS, *et al* (2012) A cost-effective and universal strategy for complete prokaryotic genomic sequencing proposed by computer simulation. *BMC Research Notes*, **5**, 80.

Jiménez-Zurdo JI, García-Rodríguez FM, Barrientos-Durán A, Toro N (2003) DNA Target Site Requirements for Homing in Vivo of a Bacterial Group

II Intron Encoding a Protein Lacking the DNA Endonuclease Domain. *Journal of Molecular Biology*, **326**, 413–423.

Jones KM, Kobayashi H, Davies BW, Taga ME, Walker GC (2007) How rhizobial symbionts invade plants: the *Sinorhizobium–Medicago* model. *Nature Reviews Microbiology*, **5**, 619–633.

Kircher M, Kelso J (2010) High-throughput DNA sequencing – concepts and limitations. *BioEssays*, **32**, 524–536.

Kislyuk AO, Haegeman B, Bergman NH, Weitz JS (2011) Genomic fluidity: an integrative view of gene diversity within microbial populations. *BMC Genomics*, **12**, 32.

Koonin EV (2006) The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? *Biology Direct*, **1**, 1–23.

Koonin EV (2009) Evolution of genome architecture. *The International Journal of Biochemistry & Cell Biology*, **41**, 298–306.

Kuhn S, Stiens M, Pühler A, Schlüter A (2008) Prevalence of pSmeSM11a-like plasmids in indigenous *Sinorhizobium meliloti* strains isolated in the course of a field release experiment with genetically modified *S. meliloti* strains. *FEMS Microbiology Ecology*, **63**, 118–131.

Lambowitz AM, Mohr G, Zimmerly S (2005) Group II Intron Homing Endonucleases: Ribonucleoprotein Complexes with Programmable Target Specificity. In *Homing Endonucleases and Inteins* (edited by DM Belfort, DDW Wood, DBL Stoddard, DV Derbyshire), no. 16 in Nucleic Acids and Molecular Biology, pp. 121–145. Springer Berlin Heidelberg.

Lambowitz AM, Zimmerly S (2004) Mobile Group II Introns. *Annual Review of Genetics*, **38**, 1–35.

Lambowitz AM, Zimmerly S (2011) Group II Introns: Mobile Ribozymes that Invade DNA. *Cold Spring Harbor Perspectives in Biology*, **3**.

Leclercq S, Cordaux R (2012) Selection-Driven Extinction Dynamics for Group II Introns in Enterobacteriales. *PLoS ONE*, **7**, e52268.

Leclercq S, Giraud I, Cordaux R (2011) Remarkable Abundance and Evolution of Mobile Group II Introns in Wolbachia Bacterial Endosymbionts. *Molecular Biology and Evolution*, **28**, 685–697.

Lee H, Tang H (2012) Next-Generation Sequencing Technologies and Fragment Assembly Algorithms. In *Evolutionary Genomics* (edited by M Anisimova), no. 855 in Methods in Molecular Biology, pp. 155–174. Humana Press.

Lee KB, Backer PD, Aono T, *et al* (2008) The genome of the versatile nitrogen fixer *Azorhizobium caulinodans* ORS571. *BMC Genomics*, **9**, 271.

Legendre P, Lapointe FJ (2004) Assessing Congruenceamong Distance Matrices: Single-Malt Scotch Whiskies Revisited. *Australian & New Zealand Journal of Statistics*, **46**, 615–629.

Lehmann K, Schmidt U (2003) Group II introns: structure and catalytic versatility of large natural ribozymes. *Critical reviews in biochemistry and molecular biology*, **38**, 249–303.

Lewin B (2007) *Genes IX*. Jones & Bartlett Publ.

Li J, Jiang J, Leung FC (2012a) 6–10 × pyrosequencing is a practical approach for whole prokaryote genome studies. *Gene*, **494**, 57–64.

Li Z, Ma Z, Hao X, Wei G (2012b) Draft Genome Sequence of *Sinorhizobium meliloti* CCNWSX0020, a Nitrogen-Fixing Symbiont with Copper Tolerance Capability Isolated from Lead-Zinc Mine Tailings. *Journal of Bacteriology*, **194**, 1267–1268.

Lobry JR (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Molecular Biology and Evolution*, **13**, 660–665.

López-Guerrero MG, Ormeño-Orrillo E, Acosta JL, *et al* (2012) Rhizobial extrachromosomal replicon variability, stability and expression in natural niches. *Plasmid*, **68**, 149–158.

Lozano L, Hernández-González I, Bustos P, *et al* (2010) Evolutionary Dynamics of Insertion Sequences in Relation to the Evolutionary Histories of the Chromosome and Symbiotic Plasmid Genes of *Rhizobium etli* Populations. *Applied and Environmental Microbiology*, **76**, 6504–6513.

Lund PA (2009) Multiple chaperonins in bacteria – why so many? *FEMS Microbiology Reviews*, **33**, 785–800.

Magrane M, Consortium U (2011) UniProt Knowledgebase: a hub of integrated protein data. *Database*, **2011**, bar009–bar009.

Mahillon J, Chandler M (1998) Insertion sequences. *Microbiology and molecular biology reviews : MMBR*, **62**, 725–774.

Malik HS, Burke WD, Eickbush TH (1999) The age and evolution of non-LTR retrotransposable elements. *Molecular Biology and Evolution*, **16**, 793–805.

Mardis E, McPherson J, Martienssen R, Wilson RK, McCombie WR (2002) What is Finished, and Why Does it Matter. *Genome Research*, **12**, 669–671.

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, **24**, 133–141.

Mardis ER (2013) Next-Generation Sequencing Platforms. *Annual Review of Analytical Chemistry*, **6**.

Margulies M, Egholm M, Altman WE, *et al* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Markowitz VM, Mavromatis K, Ivanova NN, Chen IMA, Chu K, Kyrpides NC (2009) IMG ER: a system for microbial genome annotation expert review and curation. *Bioinformatics*, **25**, 2271–2278.

Marqués S, Ramos JL, Timmis KN (1993) Analysis of the mRNA structure of the Pseudomonas putida TOL meta fission pathway operon around the transcription initiation point, the xylTE and the xylFJ regions. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, **1216**, 227–236.

Martínez-Abarca F, Barrientos-Durán A, Fernández-López M, 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 Research*, **32**, 2880–2888.

Martínez-Abarca F, García-Rodríguez F, Toro N (2000) Homing of a bacterial group II intron with an intron-encoded protein lacking a recognizable endonuclease domain. *Molecular microbiology*, **35**, 1405–1412.

Martínez-Abarca F, Martínez-Rodríguez L, López-Contreras JA, Jiménez-Zurdo JI, Toro N (2013) Complete Genome Sequence of the Alfalfa Symbiont *Sinorhizobium/Ensifer meliloti* Strain GR4. *Genome announcements*, **1**.

Martínez-Abarca F, Toro N (2000) RecA-independent ectopic transposition in vivo of a bacterial group II intron. *Nucleic acids research*, **28**, 4397–4402.

Martínez-Abarca F, Zekri S, Toro N (1998) Characterization and splicing in vivo of a *Sinorhizobium meliloti* group II intron associated with particular insertion sequences of the IS630-Tc1/IS3 retroposon superfamily. *Molecular Microbiology*, **28**, 1295–1306.

Matsuura M, Saldanha R, Ma H, *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.

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. *Proceedings of the National Academy of Sciences*, **74**, 560–564.

Meade HM, Long SR, Ruvkun GB, Brown SE, Ausubel FM (1982) Physical and genetic characterization of symbiotic and auxotrophic mutants of *Rhizobium meliloti* induced by transposon Tn5 mutagenesis. *Journal of Bacteriology*, **149**, 114–122.

Medhekar B, Miller JF (2007) Diversity-generating retroelements. *Current Opinion in Microbiology*, **10**, 388–395.

Médigue C, Moszer I (2007) Annotation, comparison and databases for hundreds of bacterial genomes. *Research in microbiology*, **158**, 724–736.

Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R (2005) The microbial pan-genome. *Current Opinion in Genetics & Development*, **15**, 589–594.

Mercado-Blanco J, Olivares J (1993) Stability and transmissibility of the cryptic plasmids of *Rhizobium meliloti* GR4. *Archives of Microbiology*, **160**, 477–485.

Mercado-Blanco J, Olivares J (1994) The Large Nonsymbiotic Plasmid pRmeGR4a of *Rhizobium meliloti* GR4 Encodes a Protein Involved in Replication That Has Homology with the RepC Protein of Agrobacterium Plasmids. *Plasmid*, **32**, 75–79.

Metzker ML (2010) Sequencing technologies — the next generation. *Nature Reviews Genetics*, **11**, 31–46.

Michel F, Feral J (1995) Structure and Activities of Group II Introns. *Annual Review of Biochemistry*, **64**, 435–461.

Michel F, Kazuhiko U, Haruo O (1989) Comparative and functional anatomy of group II catalytic introns — a review. *Gene*, **82**, 5–30.

Miller JR, Koren S, Sutton G (2010) Assembly Algorithms for Next-Generation Sequencing Data. *Genomics*, **95**, 315–327.

Mills DA, McKay LL, Dunny GM (1996) Splicing of a group II intron involved in the conjugative transfer of pRS01 in lactococci. *Journal of Bacteriology*, **178**, 3531–3538.

Milne I, Bayer M, Cardle L, *et al* (2010) Tablet—next generation sequence assembly visualization. *Bioinformatics*, **26**, 401–402.

Mitsui H, Sato T, Sato Y, Ito N, Minamisawa K (2004) *Sinorhizobium meliloti* RpoH1 is required for effective nitrogen-fixing symbiosis with alfalfa. *Molecular Genetics and Genomics*, **271**, 416–425.

Mohr G, Ghanem E, Lambowitz AM (2010) Mechanisms Used for Genomic Proliferation by Thermophilic Group II Introns. *PLoS Biol*, **8**, e1000391.

Molina-Sánchez MD (2008) *Intrones del grupo II en* Sinorhizobium meliloti*: Contribución de la proteína y la ribozima codificadas por RmInt1 a su escisión y movilidad*. Ph.D. thesis, Granada. Tesis Univ. Granada. Departamento de Genética. Leída el 28 de abril de 2008.

Molina-Sánchez MD, Barrientos-Durán A, Toro N (2011) Relevance of the Branch Point Adenosine, Coordination Loop, and 3′ Exon Binding Site for in Vivo Excision of the *Sinorhizobium meliloti* Group II Intron RmInt1. *Journal of Biological Chemistry*, **286**, 21154–21163.

Molina-Sánchez MD, Martínez-Abarca F, Toro N (2006) Excision of the *Sinorhizobium meliloti* Group II Intron RmInt1 as Circles in Vivo. *Journal of Biological Chemistry*, **281**, 28737–28744.

Molina-Sánchez MD, Martínez-Abarca F, Toro N (2010) Structural features in the C-terminal region of the *Sinorhizobium meliloti* RmInt1 group II intron-encoded protein contribute to its maturase and intron DNA-insertion function. *FEBS Journal*, **277**, 244–254.

Muñoz E, Villadas PJ, Toro N (2001) Ectopic transposition of a group II intron in natural bacterial populations. *Molecular microbiology*, **41**, 645–652.

Muñoz-Adelantado E, San Filippo J, Martínez-Abarca F, García-Rodríguez FM, Lambowitz AM, Toro N (2003) Mobility of the *Sinorhizobium meliloti* Group II Intron RmInt1 Occurs by Reverse Splicing into DNA, But Requires an Unknown Reverse Transcriptase Priming Mechanism. *Journal of Molecular Biology*, **327**, 931–943.

Murray HL, Mikheeva S, Coljee VW, *et al* (2001) Excision of Group II Introns as Circles. *Molecular Cell*, **8**, 201–211.

Ng P, Tan JJS, Ooi HS, *et al* (2006) Multiplex sequencing of paired-end ditags (MS-PET): a strategy for the ultra-high-throughput analysis of transcriptomes and genomes. *Nucleic Acids Research*, **34**, e84–e84.

Nisa Martínez R (2011) *Desarrollo de intrones del grupo II como vectores de reconocimiento génico y su aplicación en genómica funcional en microorganismos y plantas*. Ph.D. thesis, Granada.

Nisa-Martínez R, Jiménez-Zurdo JI, Martínez-Abarca F, Muñoz-Adelantado E, Toro N (2007) Dispersion of the RmInt1 group II intron in the *Sinorhizobium meliloti* genome upon acquisition by conjugative transfer. *Nucleic Acids Research*, **35**, 214–222.

Nylander JAA (2004) MrModeltest v2.

Paradis E, Claude J, Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20**, 289–290.

Peebles CL, Zhang M, Perlman PS, Franzen JS (1995) Catalytically critical nucleotide in domain 5 of a group II intron. *Proceedings of the National Academy of Sciences*, **92**, 4422–4426.

Perutka J, Wang W, Goerlitz D, Lambowitz AM (2004) Use of Computer-designed Group II Introns to Disrupt *Escherichia coli* DExH/D-box Protein and DNA Helicase Genes. *Journal of Molecular Biology*, **336**, 421–439.

Pini F, Galardini M, Bazzicalupo M, Mengoni A (2011) Plant-Bacteria Association and Symbiosis: Are There Common Genomic Traits in Alphaproteobacteria? *Genes*, **2**, 1017–1032.

Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.

Pyle AM (2010) The tertiary structure of group II introns: implications for biological function and evolution. *Critical Reviews in Biochemistry and Molecular Biology*, **45**, 215–232.

Pyle AM, Lambowitz AM (2006) Group II Introns: Ribozymes That Splice RNA and Invade DNA. *The RNA World: The Nature of Modern RNA Suggests a Prebiotic RNA World*, **43**, 469.

Reeve W, Chain P, O'Hara G, *et al* (2010) Complete genome sequence of the *Medicago* microsymbiont *Ensifer* (*Sinorhizobium*) *medicae* strain WSM419. *Standards in Genomic Sciences*, **2**, 77–86.

Richardson EJ, Watson M (2012) The automatic annotation of bacterial genomes. *Briefings in Bioinformatics*.

Richter DC, Schuster SC, Huson DH (2007) OSLay: optimal syntenic layout of unfinished assemblies. *Bioinformatics (Oxford, England)*, **23**, 1573–1579.

Robertsen BK, Aman P, Darvill AG, McNeil M, Albersheim P (1981) Host-Symbiont Interactions : V. The structure of acidic extracellular polysaccharides secreted by *Rhizobium leguminosarum* and *Rhizobium trifolii*. *Plant physiology*, **67**, 389–400.

Rocha E (2008) The organization of the bacterial genome. *Annual review of genetics*, **42**, 211–233.

Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P (1996) Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. *Analytical Biochemistry*, **242**, 84–89.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Rothberg JM, Hinz W, Rearick TM, *et al* (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348–352.

Rothberg JM, Leamon JH (2008) The development and impact of 454 sequencing. *Nature Biotechnology*, **26**, 1117–1124.

Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular cloning*, vol. 2. Cold spring harbor laboratory press New York.

Sanger F, Coulson A (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, **94**, 441–448.

Sanger F, Coulson A, Friedmann T, *et al* (1978) The nucleotide sequence of bacteriophage φX174. *Journal of Molecular Biology*, **125**, 225–246.

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.

Sanjuan J, Olivares J (1989) Implication of nifA in regulation of genes located on a *Rhizobium meliloti* cryptic plasmid that affect nodulation efficiency. *Journal of Bacteriology*, **171**, 4154–4161.

Sayers EW, Barrett T, Benson DA, *et al* (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **39**, D38–D51.

Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. *Human Molecular Genetics*, **19**, R227–R240.

Schneiker S, Kosier B, Pühler A, Selbitschka W (1999) The *Sinorhizobium meliloti* Insertion Sequence (IS) Element ISRm14 Is Related to a Previously Unrecognized IS Element Located Adjacent to the *Escherichia coli* Locus of Enterocyte Effacement (LEE) Pathogenicity Island. *Current Microbiology*, **39**, 274–281.

Schneiker-Bekel S, Wibberg D, Bekel T, *et al* (2011) The complete genome sequence of the dominant *Sinorhizobium meliloti* field isolate SM11 extends the *S. meliloti* pan-genome. *Journal of biotechnology*, **155**, 20–33.

Selbitschka W, Arnold W, Jording D, Kosier B, Toro N, Pühler A (1995) The insertion sequence element ISRm2011-2 belongs to the IS630-Tc1 family of transposable elements and is abundant in *Rhizobium meliloti*. *Gene*, **163**, 59–64.

Selbitschka W, Zekri S, Pühler A, Schröder G, Toro N (1999) The *Sinorhizobium meliloti* insertion sequence (IS) elements ISRm102F34-1/ISRm7 and ISRm220-13-5 belong to a new family of insertion sequence elements. *FEMS Microbiology Letters*, **172**, 1–7.

Sernova NV, Gelfand MS (2008) Identification of replication origins in prokaryotic genomes. *Briefings in Bioinformatics*, **9**, 376–391.

Shearman C, Godon JJ, Gasson M (1996) Splicing of a group II intron in a functional transfer gene of *Lactococcus lactis*. *Molecular Microbiology*, **21**, 45–53.

Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nature Biotechnology*, **26**, 1135–1145.

Shendure J, Porreca GJ, Reppas NB, *et al* (2005) Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science*, **309**, 1728–1732.

Sibley CD, MacLellan SR, Finan T (2006) The *Sinorhizobium meliloti* chromosomal origin of replication. *Microbiology*, **152**, 443–455.

Siefert J (2009) Defining the Mobilome. In *Horizontal Gene Transfer* (edited by M Gogarten, J Gogarten, L Olendzenski), vol. 532 of *Methods in Molecular Biology*, pp. 13–27. Humana Press.

Siguier P, Filée J, Chandler M (2006a) Insertion sequences in prokaryotic genomes. *Current Opinion in Microbiology*, **9**, 526–531.

Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M (2006b) ISfinder: the reference centre for bacterial insertion sequences. *Nucleic acids research*, **34**, D32–6.

Siguier P, Varani A, Perochon J, Chandler M (2012) Exploring Bacterial Insertion Sequences with ISfinder: Objectives, Uses, and Future Developments. In *Mobile Genetic Elements* (edited by Y Bigot), no. 859 in Methods in Molecular Biology, pp. 91–103. Humana Press.

Simon DM, Clarke NAC, McNeil BA, *et al* (2008) Group II introns in Eubacteria and Archaea: ORF-less introns and new varieties. *RNA*, **14**, 1704–1713.

Simon DM, Kelchner SA, Zimmerly S (2009) A Broadscale Phylogenetic Analysis of Group II Intron RNAs and Intron-Encoded Reverse Transcriptases. *Molecular Biology and Evolution*, **26**, 2795–2808.

Singh NN, Lambowitz AM (2001) Interaction of a group II intron ribonucleoprotein endonuclease with its DNA target site investigated by DNA footprinting and modification interference. *Journal of Molecular Biology*, **309**, 361–386.

Smith D, Zhong J, Matsuura M, Lambowitz AM, Belfort M (2005) Recruitment of host functions suggests a repair pathway for late steps in group II intron retrohoming. *Genes & Development*, **19**, 2477–2487.

Soto MJ, Zorzano A, Mercado-Blanco J, Lepek V, Olivares J, Toro N (1993) Nucleotide sequence and characterization of *Rhizobium meliloti* nodulation competitiveness genes nfe. *Journal of molecular biology*, **229**, 570–576.

Soto MJ, Zorzano A, Olivares J, Toro N (1992a) Nucleotide sequence of *Rhizobium meliloti* GR4 insertion sequence ISRm3 linked to the nodulation competitiveness locus nfe. *Plant molecular biology*, **20**, 307–309.

Soto MJ, Zorzano A, Olivares J, Toro N (1992b) Sequence of ISRm4 from *Rhizobium meliloti* strain GR4. *Gene*, **120**, 125–126.

Stiens M, Schneiker S, Keller M, Kuhn S, Pühler A, Schlüter A (2006) Sequence analysis of the 144-kilobase accessory plasmid pSmeSM11a, isolated from a dominant *Sinorhizobium meliloti* strain identified during a long-term field release experiment. *Applied and environmental microbiology*, **72**, 3662–3672.

Stiens M, Schneiker S, Pühler A, Schlüter A (2007) Sequence analysis of the 181-kb accessory plasmid pSmeSM11b, isolated from a dominant *Sinorhizobium meliloti* strain identified during a long-term field release experiment. *FEMS microbiology letters*, **271**, 297–309.

Studier F (1991) Use of bacteriophage T7 lysozyme to improve an inducible T7 expression system. *Journal of Molecular Biology*, **219**, 37–44.

Sugiura N (1978) Further analysts of the data by akaike' s information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, **7**, 13–26.

Swofford DL (2002) *PAUP 4.0 b10: Phylogenetic analysis using parsimony*. Sinauer Associates, Sunderland, MA, USA.

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Molecular Biology and Evolution*, **24**, 1596–1599.

Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research*, **28**, 33–36.

Toor N, Hausner G, Zimmerly S (2001) Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA*, **7**, 1142–1152.

Toor N, Robart AR, Christianson J, Zimmerly S (2006) Self-splicing of a group IIC intron: 5′ exon recognition and alternative 5′ splicing events implicate the stem–loop motif of a transcriptional terminator. *Nucleic Acids Research*, **34**, 6461–6471.

Toro N (1985) *Estudio genético de las características simbióticas de* Rhizobium meliloti. Ph.D. thesis, Granada.

Toro N (2003) Bacteria and Archaea Group II introns: additional mobile genetic elements in the environment. *Environmental microbiology*, **5**, 143–151.

Toro N, Jiménez-Zurdo JI, García-Rodríguez FM (2007) Bacterial group II introns: not just splicing. *FEMS microbiology reviews*, **31**, 342–358.

Toro N, Martínez-Abarca F (2013) Comprehensive phylogenetic analysis of bacterial group II intron-encoded ORFs lacking the DNA endonuclease domain reveals new varieties. *PloS one*, **8**, e55102.

Toro N, Martínez-Abarca F, Fernández-López M, Muñoz-Adelantado E (2003) Diversity of group II introns in the genome of *Sinorhizobium meliloti* strain 1021: splicing and mobility of RmInt1. *Molecular genetics and genomics : MGG*, **268**, 628–636.

Toro N, Molina-Sánchez M, Fernández-López M (2002) Identification and characterization of bacterial class E group II introns. *Gene*, **299**, 245–250.

Toro N, Olivares J (1986) Characterization of a large plasmid of *Rhizobium meliloti* involved in enhancing nodulation. *Molecular and General Genetics MGG*, **202**, 331–335.

Touchon M, Rocha EPC (2007) Causes of Insertion Sequences Abundance in Prokaryotic Genomes. *Molecular Biology and Evolution*, **24**, 969–981.

Toussaint A, Chandler M (2012) Prokaryote Genome Fluidity: Toward a System Approach of the Mobilome. In *Bacterial Molecular Networks* (edited by J Helden, A Toussaint, D Thieffry), vol. 804, pp. 57–80. Springer New York, New York, NY.

Vance CP (1998) Legume Symbiotic Nitrogen Fixation: Agronomic Aspects. In *The Rhizobiaceae* (edited by HP Spaink, A Kondorosi, PJJ Hooykaas), pp. 509–530. Springer Netherlands.

Venkova-Canova T, Chattoraj DK (2011) Transition from a plasmid to a chromosomal mode of replication entails additional regulators. *Proceedings of the National Academy of Sciences*, **108**, 6199–6204.

Villadas PJ, Velazquez E, Martínez-Molina E, Toro N (1995) Identification of nodule-dominant *Rhizobium meliloti* strains carrying pRmeGR4b-type plasmid within indigenous soil populations by PCR using primers derived from specific DNA sequences. *FEMS Microbiology Ecology*, **17**, 161–168.

Voelkerding KV, Dames SA, Durtschi JD (2009) Next-Generation Sequencing: From Basic Research to Diagnostics. *Clinical Chemistry*, **55**, 641–658.

Wank H, SanFilippo J, Singh RN, Matsuura M, Lambowitz AM (1999) A Reverse Transcriptase/Maturase Promotes Splicing by Binding at Its Own Coding Segment in a Group II Intron RNA. *Molecular Cell*, **4**, 239–250.

Wartell RM, Reznikoff WS (1980) Cloning DNA restriction endonuclease fragments with protruding single-stranded ends. In *Gene*, vol. 9, pp. 307–319.

Watanabe K, Lambowitz AM (2004) High-affinity binding site for a group II intron-encoded reverse transcriptase/maturase within a stem–loop structure in the intron RNA. *RNA*, **10**, 1433–1443.

Weidner S, Baumgarth B, Göttfert M, *et al* (2013) Genome Sequence of *Sinorhizobium meliloti* Rm41. *Genome Announcements*, **1**.

Wiley G, Macmil S, Qu C, *et al* (2009) Methods for Generating Shotgun and Mixed Shotgun/Paired-End Libraries for the 454 DNA Sequencer. In *Current Protocols in Human Genetics*. John Wiley & Sons, Inc.

Williams KP, Sobral BW, Dickerman AW (2007) A Robust Species Tree for the Alphaproteobacteria. *Journal of Bacteriology*, **189**, 4578–4586.

Yahaya MO (2012) Review of techniques for gene sequencing, annotation and comparative genomics. *International Journal of Applied*, **1**, 25–28.

Yao J, Zhong J, Lambowitz AM (2005) Gene targeting using randomly inserted group II introns (targetrons) recovered from an *Escherichia coli* gene disruption library. *Nucleic Acids Research*, **33**, 3351–3362.

Young J, Crossman L, Johnston A, *et al* (2006) The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. *Genome biology*, **7**, R34.

Zekri S, Soto MJ, Toro N (1998) ISRm4-1 and ISRm9, two novel insertion sequences from *Sinorhizobium meliloti*. *Gene*, **207**, 93–96.

Zekrí S, Toro N (1996) Identification and nucleotide sequence of *Rhizobium meliloti* insertion sequence ISRm6, a small transposable element that belongs to the IS3 family. *Gene*, **175**, 43–48.

Zekri S, Toro N (1998) A new insertion sequence from *Sinorhizobium meliloti* with homology to IS1357 from *Methylobacterium sp.* and IS1452 from *Acetobacter pasteurianus*. *FEMS microbiology letters*, **158**, 83–87.

Zhong J, Karberg M, Lambowitz AM (2003) Targeted and random bacterial gene disruption using a group II intron (targetron) vector containing a retrotransposition-activated selectable marker. *Nucleic Acids Research*, **31**, 1656–1664.

Zhong J, Lambowitz AM (2003) Group II intron mobility using nascent strands at DNA replication forks to prime reverse transcription. *The EMBO Journal*, **22**, 4555–4565.

Zhuang F, Karberg M, Perutka J, Lambowitz AM (2009) EcI5, a group IIB intron with high retrohoming frequency: DNA target site recognition and use in gene targeting. *RNA*, **15**, 432–449.