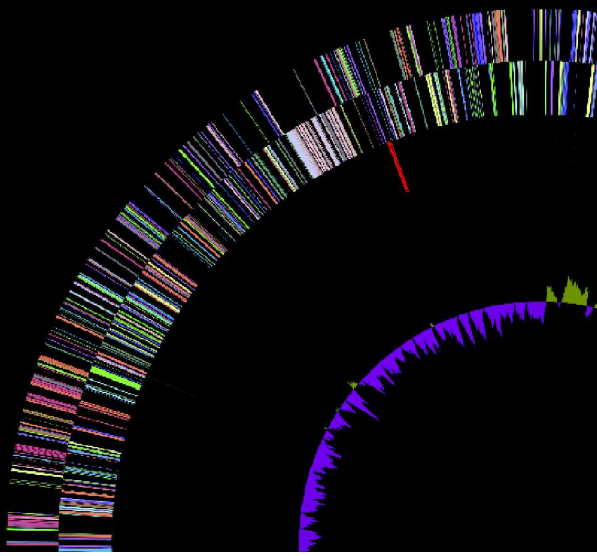
Estudio de los mecanismos de diversificación intraespecífica de *Salinibacter ruber*

Tesis doctoral
2016

Pedro Iñaki González Torres



Universitat d'Alacant
Universidad de Alicante

Universitat d'Alacant
Universidad de Alicante

Departamento de Fisiología, Genética y Microbiología
Universidad de Alicante

Centre de Regulació Genòmica (CRG)
Parc de Recerca Biomèdica de Barcelona (PRBB)

DOCTORADO EN BIOLOGÍA EXPERIMENTAL Y APLICADA

Estudio de los mecanismos de diversificación intraespecífica de *Salinibacter ruber*.

Memoria presentada para optar al grado de Doctor por la Universidad de Alicante

Pedro Iñaki González Torres

Alicante, Febrero 2016

Universitat d'Alacant
Universidad de Alicante

La Dra. **Josefa Antón Botella**, Profesora Titular del Departamento de Fisiología, Genética y Microbiología de la Universidad de Alicante y,

El Dr. **Juan Antonio Gabaldón Esteban**, Profesor ICREA y jefe de grupo en el Departamento de Bioinformatica y Genómica, del Centro de Regulación Genómica (CRG)

CERTIFICAN:

Que la memoria de Tesis doctoral titulada **"Estudio de los mecanismos de diversificación intraespecífica en *Salinibacter ruber*",** presentada por Pedro Iñaki González Torres, ha sido realizada bajo su dirección en el Departamento de Fisiología Genética y Microbiología de la Universidad de Alicante y en el Centre de Regulació Genòmica (CRG), Parc de Recerca Biomèdica de Barcelona (PRBB). Y para que conste a los efectos, firman en Alicante a 15 de Diciembre del año dos mil quince.

Fdo: Josefa Antón Botella. Fdo. Juan Antonio Gabaldón Esteban.

A mis padres y hermana

AGRADECIMIENTOS

Por fin llegó el momento en el que me enfrento a estas líneas, 5 años llenos de experiencias en los que siento que no sólo he crecido a nivel profesional sino también en el ámbito personal. Años llenos de experiencias en los que sois muchos los que me habéis brindado vuestro apoyo, momentos únicos que recorren mi mente a cámara rápida mientras escribo emocionado estas líneas…

En primer lugar, me gustaría dar las gracias a mis directores de tesis Pepa y Toni. Han pasado algunos años desde aquel seminario en la UA y aquel día en que aterricé en Barcelona con aquella caja de hielo seco. Me siento muy agradecido por la oportunidad que me habéis brindado formando parte de dos grupos de investigación increíbles en lo personal y en el aspecto investigador. Gracias por la confianza depositada en mi durante este periodo y por compartir conmigo vuestra experiencia. Gracias Toni por enseñarme que la edad es sólo un estado de ánimo.

No puedo olvidar la primera vez que pisé el que por entonces no sabía que sería el laboratorio donde crecería como investigador. Por aquel entonces recuerdo mis comienzos tímidos como joven Padawan de la mano de Manu y Cris. Fue con vosotros con los que empecé a dar los primeros pasos, aquellos inicios tambaleantes en los que di mis primeros pasos. Siempre recordaré aquellos momentos con mucho cariño. Gracias a vosotros chicos y muchas agracias también a Txo, Fer y Mary por todas las cosa que me habéis enseñado a lo largo de estos años. Tobal y Noe, como olvidar aquellas tardes de cumpleaños feliz antes de que anochezca jajaja, gracias por los buenos ratos y conversaciones, sois estupendos chicos. Judith, empezamos juntos este camino, mucho ánimo campeona ya queda menos. Gracias a Isabel, una de las mejores profes que he tenido, Jesús, Francis, Esther y Mónics. Gracias Loles por tu sentido del humor y Paco por todo lo que he aprendido de ti.

Las estancias en Barcelona me han reportado momentos buenísimos, no puedo olvidarme de ninguno de los miembros de mi otro equipo: Leszek gracias por todo lo que me enseñaste, Salva y esos "amistosos" jugando al futbol, Marina y sus buenos consejos, Gabriela, Damien, Alex, Jose, Fran, Esther y Eva. Gracias a todos vosotros por tan buenos momentos, conversaciones risas y por lo que me hicisteis crecer como persona. Gracias por recibirme desde el primer día como uno más y hacerme sentir parte del equipo!!

Gracias a mis amigos Luana y Eder por acogerme como a un hermano desde el primer día, sin duda, aquellas Navidades las guardo como uno de mis mejores recuerdos, gracias por hacer de Barcelona mi segundo hogar.

Gracias a las chicas del máster, Vir, Asun, Anna, Cindy por los buenos ratos que pasamos. Ánimo Asun y Anna que ya falta poco!! No puedo olvidarme de la gente del departamento de

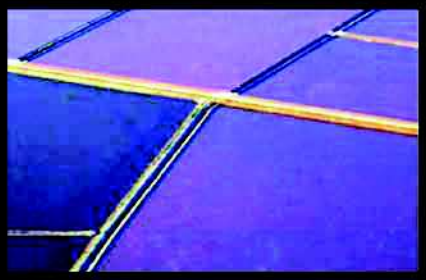
Bioquímica y Fitopatología, muchas gracias por todo lo que me habéis aportado durante estos años en la UA, Fede, Nuria, Edu, gracias por esos ratos de apoyo, a mis profesoras Mónica y Rosa, vosotras también tenéis la culpa de que me guste tanto esto. Gracias a nuestros vecinos de genética y fisio por tan buenos ratos y conversaciones.

Por último me gustaría dar las gracias a un pilar fundamental no sólo en esta tesis sino en mi vida, mis amigos y familia. Agradecer a mis padres y mi hermana el apoyo que me han dado siempre, desde el día en que decidí dedicarme a la Biología. Siempre habéis estado a mi lado y sin vosotros no hubiera podido llegar a donde estoy.

Sin duda durante este tiempo vosotros, si vosotros mis amigos de toda la vida habéis sido un apoyo imprescindible. Siempre os he tenido a mi lado en los momentos más difíciles y no sólo en esta etapa. Víctor, Muntó, Aaron, Dani, Charlie, esta tesis lleva mucho de todos vosotros, pues en ella hay mucha de vuestra alegría, apoyo, risas y buenos momentos que me han permitido continuar en los momentos más duros, porque es en esos momentos en los que más presentes habéis estado amigos míos.

Gracias a ti Silvia por todo lo que me has dado, gracias por recorrer estos años a mi lado.

GRACIAS A TODOS POR VUESTRO APOYO!!!



1.1- Ecología, distribución y microdiversidad de *Salinibacter ruber.*

1.1.1- Microbiota y microdiversidad en ambientes hipersalinos.

Los ambientes hipersalinos se caracterizan por presentar una concentración elevada de sales, superior a la del agua del mar, y constituyen un claro ejemplo de ambiente extremo. Dentro de este tipo de ambiente, las salinas solares constituyen un modelo de estudio muy interesante, ya que a lo largo de su circuito es posible hallar un gradiente de concentración de sales. Esta peculiaridad las convierte en el entorno perfecto para el estudio de estrategias adaptativas de los distintos microorganismos presentes a lo largo del mismo. El circuito de una salina solar comprende una serie de estanques contiguos. En su parte inicial se sitúan los denominados estanques preparadores que reciben el agua del mar. A continuación el agua circula hacia los concentradores, donde incrementa secuencialmente la salinidad. Al final de este circuito se localizan los cristalizadores, los de mayor salinidad y donde precipita el cloruro sódico. Las salinas solares contienen microorganismos halófilos de los tres dominios, *Archaea, Bacteria y Eukarya* (Oren, 2008). A medida que se incrementa la salinidad, tiene lugar un aumento en la densidad de microorganismos, acompañado de un descenso en la diversidad de especies y un incremento de la microdiversidad o diversidad intraespecífica, manifestándose de este modo el efecto selectivo de este ambiente extremo. Los estanques de baja salinidad, preparadores, presentan una microbiota similar a la del agua de mar (Oren, 2002b) pero, con el incremento de esta, se produce un aumento gradual en diversidad y abundancia de representantes del dominio *Archaea,* siendo el dominio predominante en los cristalizadores.

La caracterización de la microbiota de los cristalizadores y sus mecanismos de adaptación han suscitado especial interés ya que en estos estanques es posible encontrar microorganismos con óptimos de crecimiento en medios con elevada salinidad (2,5 a 5,2M), valores muy superiores a los del agua de mar (0,5M). Estos microorganismos, conocidos como halófilos extremos, han sido estudiados extensamente mediante técnicas de cultivo y moleculares. Los primeros estudios moleculares exploraron la diversidad a nivel de la secuencia del gen del rRNA 16S (Benlloch *et al*., 1995; Rodríguez Valera *et al*., 1999) y mediante hibridación *in situ* de

fluorescencia (FISH) con sondas específicas (Antón *et al*., 1999). Estos trabajos, realizados hace poco más de quince años, detectaron por primera vez la presencia relevante de organismos halófilos extremos del dominio *Bacteria* (Antón *et al*., 2000) en cristalizadores, entre ellos *Salinibacter ruber* como el representante mayoritario. Fruto de los resultados obtenidos cambió totalmente la percepción de estos estanques de elevada salinidad como un cultivo monoespecífico de representantes del dominio *Archaea.*

Actualmente se sabe que la abundancia de procariotas en los cristalizadores es del orden de 10^7-10^8 células/ml, de las cuales generalmente entre el 70-95% pertenecen el dominio *Archaea* y el 5-30% al dominio *Bacteria* (Antón *et al*., 2008). Los análisis moleculares llevados a cabo muestran como la mayoría de secuencias del gen del rRNA 16S de estos cristalizadores, y en otros ambientes hipersalinos, corresponden con la especie *Haloquadratum walsbyi* (Bolhuis *et al.,* 2004), representante más abundante del dominio *Archaea*, y *S. ruber* (Antón *et al.,* 2002), representante mayoritario del dominio *Bacteria*, encontrando representantes minoritarios de otras especies de ambos dominios. Dentro del dominio *Archaea* los organismos más halófilos pertenecen al orden *Halobacteriales* (filo *Euryarchaeota*) (Oren, 2002a). Recientemente, se ha descrito la presencia en cristalizadores de todo el mundo de miembros del nuevo filo *Nanohaloarchaea* (Ghai *et al*., 2012, Narasingarao *et al*., 2011, Gomariz *et al*., 2015). En el caso del dominio *Bacteria* abundan los representantes del filo *Bacteriodetes*, entre ellos en primer lugar *S. ruber* (Antón *et al*., 2000) y en algunas salinas *Salisaeta longa* (Vaisman y Oren, 2009). Aunque en general *Haloquadratum* y *Salinibacter* constituyen los géneros predominantes dentro de los dominios *Archaea* y *Bacteria* en la mayoría de cristalizadores, existen casos en los cuales no se sigue este patrón. Un ejemplo claro son las salinas de Maras, donde *Haloquadratum* no es la *Archaea* más abundante y *Salícola maranensis* es el representante más abundante del dominio *Bacteria* (Maturrano *et al*., 2006). En estas salinas no se ha detectado la presencia de *S. ruber* por técnicas moleculares aunque sí se pudo aislar por cultivo. Otro caso son algunas salinas de la región adriática, donde predomina *Halorubrum* spp. (Pasic *et al*., 2005).

Además los ambientes hipersalinos presentan también los mayores números de partículas víricas (halovirus) en el conjunto de los ambientes acuáticos (Guixa-Boixareu *et al.,* 1996; Santos *et al*., 2010). En el caso de las salinas solares de Santa Pola se ha apreciado un aumento

creciente en el número de virus, desde $4x10^8$ hasta $2x10^9$ partículas víricas por mililitro a lo largo del gradiente de salinidad, superando en hasta dos órdenes de magnitud a la comunidad procariota, pudiendo tener un papel determinante en el control de las poblaciones procariotas halófilas. Los hospedadores más abundantes para estos virus serían los más afectados por los procesos de lisis vírica, estimulando de este modo los flujos de carbono y energía, y dirigiendo la evolución de los genomas de sus hospedadores (Rodríguez-Valera, *et al.,* 2009).

En los últimos años, la aplicación de nuevas técnicas y aproximaciones al estudio de los ambientes hipersalinos, tales como estudios metagenómicos y de *single cell genomics* (SCG), han permitido realizar una descripción mucho más precisa de su microbiota (Antón *et al*., 2012, Ghai *et al*., 2012, Narasingarao *et al*., 2011, Gomariz *et al*., 2015), de la diversidad intraespecífica de la misma y el descubrimiento de nuevos representantes como en el caso de las N*anohaloarchaea* (Ghai *et al*., 2012; Narasingarao *et al*., 2011). Recientemente, se están empleando estudios metagenómicos comparativos para obtener una caracterización más completa de la microbiota de diversos cristalizadores del mundo apreciando patrones de distribución de filotipos particulares de cada uno de ellos. El empleo de SCG ha permitido comparar la microbiota de los cristalizadores de las salinas de Santa Pola y South Bay Salt Works en Chula Vista (California,USA) (Zhaxybayeva *et al*., 2013). Del estudio se derivó que aunque ambos ambientes comparten representantes comunes en su microbiota, tales como las *Nanohaloarchaea*, sus comunidades microbianas son significativamente distintas. En Chula Vista los representantes mayoritarios del dominio *Bacteria* fueron diversas especies de *Proteobacterias* y *Bacteroidetes* mientras en Santa Pola la fracción bacteriana está constituida mayoritariamente por representantes de género *Salinibacter*, concluyendo que entre ambos ambientes existían comunidades con representantes bacterianos mayoritarios significativamente distintos.

1.1.2- *Salinibacter ruber*: Características, distribución y relevancia ecológica.

S. ruber constituye el primer miembro halófilo extremo del dominio *Bacteria* cuya relevancia ecológica ha sido demostrada (Antón *et al.,* 2008). Esta especie, que pertenece al Filo *Bacteroidetes,* Clase II *Sphingobacteria,* comparte rasgos fenotípicos y hábitat con las *Archaea* de la familia *Halobacteriaceae* **(tabla 1.I**). Ambas son halófilas extremas, aeróbicas y quimioorganotrofas (Antón *et al.,* 2002), empleando el potasio como soluto compatible a elevada concentración citoplasmática. Los aislados de esta especie presentan una coloración rojiza debido a la acumulación del carotenoide C40 salinixantina. Como las *Archaea* de la familia *Halobacteriaceae,* presenta en su membrana proteínas unidas a retinal: halorrodopsina (una bomba de cloruros activada por luz), xantorrodopsina (un análogo de la bacteriorrodopsina que funciona como bomba de protones fotosintética) y rodopsinas sensoriales. Asimismo el proteoma de *S. ruber* muestra un elevado contenido en aminoácidos ácidos, una baja proporción de aminoácidos básicos e hidrofóbicos y un punto isoeléctrico medio de 5.2 (Mogondin *et al.,* 2005). El contenido en G+C de la especie, determinado mediante HPLC (del inglés, *High-performance liquid chromatography*), se halla entre 66,3 y 67,7 mol% (Antón, 2002).

Su descubrimiento se llevó a cabo mediante el empleo de técnicas moleculares aplicadas a estudios de ecología microbiana de las salinas solares (Antón *et al*., 2000). En 1998 se detectó mediante FISH la presencia en una proporción relevante de microorganismos del Dominio *Bacteria* (Antón *et al.,*1999). Mediante DGGE (del inglés, *Denaturing gradient gel electrophoresis*) y genotecas del gen rRNA16S se reveló la presencia de dos filotipos bacterianos muy próximos (EBH-1 EBH-2) (del inglés, *Extremely Halophilic Bacterium*) (Antón *et al.,*2000), siendo el primero mucho más abundante que el segundo. Poco después se aislaron 5 cepas representantes del filotipo EBH-1 de las Salinas de Campos de Mallorca (M1, M8 y M31) y del Bras del Port en Santa Pola (Pola 13, Pola 18), utilizadas para la descripción de la especie *S. ruber* y seleccionando M31 (DSM 15338) como la cepa tipo (Antón *et al.,* 2002). Los estudios llevados a cabo durante los últimos 15 años se han centrado en la caracterización de la distribución, diversidad, abundancia, y microdiversidad genómica y metabolómica de esta especie (Peña *et al*., 2005, Rosselló-Mora *et al*., 2008, Peña *et al*., 2010, Antón *et al*., 2013).

Tabla I1. Comparación entre *Salinibacter ruber* y la familia *Halobacteriaceae (Oren, 2008).*

	S. ruber	*Halobacteriaceae*
Sal requerida	> 150 g/l	>150 g/l (en la mayoría)
Óptimo de sal	150-300 g/l	200-250 g/l
%G+C	66,2	50-71 (46,9 en *H.walsbyi*)
Soluto compatible	KCl	KCl
Enzimas	Dependientes de sal y tolerantes	Dependientes de sal
Lípidos	Bacterianos (éster+glicerol)	Arqueanos (éter+glicerol)
Carotenoides	C-40 (salinixantina)	C-50 (bacteriorruberina)
Proteínas de unión a retinal*	XR, HR, SR	BR, HR, SR

***XR: xantorrodopsina; HR: halorrodopsina; SR: rodopsina sensorial; BR: bacteriorrodopsina.**

El empleo de distintos métodos de detección moleculares (FISH, análisis de clones del gen del rRNA 16S, DGGE) además de técnicas de cultivo y más recientemente aproximaciones metagenómicas, ha permitido describir la distribución de este género en diferentes ambientes a lo largo del planeta (Antón *et al.,* 2008). Tal como muestra la **figura I1**, en Europa se han identificado representantes del género *Salinibacter* en los cristalizadores de las salinas de Santa Pola, Ibiza, Mallorca, Canarias y Delta del Ebro mediante técnicas moleculares y de cultivo. Su presencia en Asia ha sido detectada en El Lago Tuz (Turquía), encontrando representantes de filotipos claramente distintos a los descritos anteriormente EHB-I y EHB-II, y en Israel, en este último caso sólo mediante cultivo. En África se han detectado secuencias relacionadas con *S. ruber*, en tres lagos de Egipto y las salinas de Sfax (Túnez). En América se han detectado secuencias relacionadas con el género en el Gran Lago Salado de Utah (EEUU) y en las salinas de Guerrero Negro en Baja California (México), mientras que en las salinas de Maras (Perú) sólo se detectó su presencia mediante técnicas de cultivo (Maturrano *et al.,* 2006) siendo *Salicola marasensis* el representante más abundante del dominio *Bacteria*.

Aunque existen ambientes hipersalinos como las salinas de Maras en los cuales no se ha detectado una presencia relevante de *Salinibacter* o filotipos próximos, durante los últimos años se han identificado secuencias pertenecientes al filo *Bacteroidetes* alrededor de todo el mundo (Antón *et al.*, 2008). En algunos ambientes como las salinas de Campos (Mallorca) o del Bras

del Port (Santa Pola), el género *Salinibacter* llega a representar el 30% de la comunidad procariota, pudiendo ser detectado mediante técnicas moleculares y constituyendo un taxón *core* en el ecosistema (Pedrós-Alió, 2006). Según este autor, la biodiversidad de un ecosistema está constituida por dos elementos: el *core* y el *seed bank*. Los taxones más abundantes de una comunidad, con crecimiento activo, que juegan un papel funcional predominante y están sometidos a lisis viral continua constituirían el *core* de la misma. En otros casos, la presencia de *Salinibacter* sólo es detectable empleando técnicas de cultivo, y rara vez mediante técnicas moleculares, siendo uno de los taxones menos frecuentes y formando parte del *seed bank* (Pedrós- Alió, 2006), segundo componente descrito por este autor y que estaría constituido por los taxones menos frecuentes, no sometidos a depredación o lisis viral y que rara vez son detectados mediante técnicas moleculares.

Además de permitir la caracterización de la microbiota en este tipo de ambiente y su distribución a nivel global, el avance en técnicas moleculares y de secuenciación ha permitido observar su variación espacio temporal en ambientes hipersalinos como las salinas de San Diego mediante aproximaciones metagenómicas (Rodríguez-Brito *et al*., 2010) o las salinas de Bras del Port por medio de DGGE (Gomariz *et al.*, 2015). Del primer estudio se deriva que las especies más abundantes persisten a lo largo del tiempo, presentando fluctuaciones en abundancia relativa así como en su microdiversidad. El segundo muestra oscilaciones estacionales de filotipos relacionados con *Salinibacter* y otros *Bacteroidetes* en 5 estanques de las salinas de Santa Pola. En el mismo, tras el análisis multivariante de estos datos se observó que las variaciones estacionales apreciadas en la comunidad microbiota se correlacionaban con los cambios en la composición iónica del medio y las perturbaciones ambientales (Gomariz *et al*., 2015). Es lógico pensar que debido a las distintas condiciones ambientales presentes en salinas de todo el mundo, y en un mismo ambiente a lo largo del tiempo, se aprecien diferencias en la composición de su comunidad halófila, la abundancia relativa de las especies que la componen y, dentro de cada una de las especies, en las diferentes cepas englobadas y su microdiversidad.

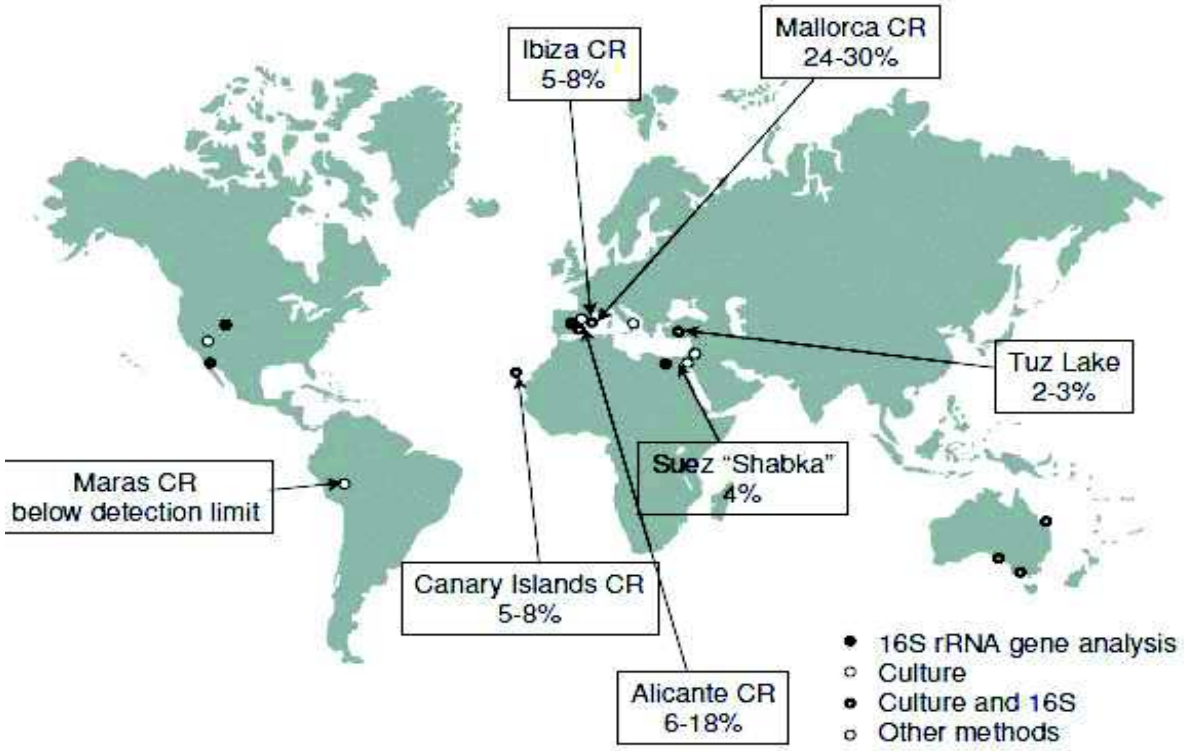


Figura I1. Distribución de los de los representantes del género *Salinibacter*, mostrando el método de detección y el porcentaje de abundancia en los casos en que se llevaron a cabo recuentos por FISH (adaptada de Antón *et al*., 2008).

1.1.3- Microdiversidad fenotípica en *S. ruber*. Antecedentes en estudios metabolómicos

Dado que los análisis multilocus de secuencias (MLSA) (véase apartado 1.1.4) y de patrones de restricción, visualizados por electroforesis en gel de campo pulsado (PFGE), para las cepas de *S. ruber* aisladas en las localizaciones mediterráneas, atlánticas o peruanas no mostraron patrones de segregación geográficos, se recurrió al análisis metabolómico de las mismas por medio de la espectrometría de masas de alta resolución (Rosselló-Mora *et al*., 2008). Se emplearon 28 cepas aisladas, entre ellas las 10 usadas en el de MLSA (véase apartado 1.1.4; **tabla I2**). La idea fue explorar patrones geográficos de cepas aisladas utilizando datos más directos de interacción con el medio (fenotipo) más a que a nivel genotípico. La aplicación de una metodología pionera de caracterización metabolómica permitió observar que, a pesar de ser

organismos muy semejantes, las distintas cepas de *S. ruber* presentan características que se correlacionaban con su origen geográfico (Rosselló-Mora et al., 2008). Las diferencias que marcaban una segregación geográfica se encontraban a nivel de la expresión diferencial de algunos metabolitos, en general componentes de la envuelta celular, ácidos grasos y terpenoides, por lo que la incipiente especiación apreciada podría manifestarse en diferencias a nivel transcripcional o post-transcripcional. Los aislados atlánticos se situaron en un espacio intermedio entre los mediterráneos y los peruanos. Estos resultados concordaron con los derivados de los análisis de correspondencias (CA) (**figura I6**) realizados con genes de transferencia horizontal inter-dominio (HGT) y que se explicarán más adelante (véase apartado 1.1.4.1). A una escala más pequeña, se encontraron diferencias entre las cepas mediterráneas en metabolitos implicados en metabolismo primario (biosíntesis y metabolismo de carbohidratos, aminoácidos y ácidos grasos).

Como se acaba de mencionar, los datos metabolómicos permitieron explorar diferencias y patrones fenotípicos biogeográficos generales pero también diferencias a pequeña escala para una misma localización. De hecho, la comparación metabolómica de las cepas coaisladas M8 y M31 puso de manifiesto diferencias metabolómicas significativas, relacionadas sobre todo con la producción de metabolitos sulfatados y /o glicosilados de la fracción extracelular (Peña *et al*., 2010). Estas diferencias en los componentes de sus paredes celulares y los metabolitos que secretan al medio se pueden atribuir a las diferencias genómicas encontradas en las zonas hipervariables en genes que codifican para sulfotransferasas y glicosiltransferasas implicadas en la biosíntesis de componentes de pared celular. Las diferencias fenotípicas entre ambas cepas podrían relacionarse con la distinta susceptibilidad y respuesta a infección por fagos de M8 y M31. Experimentos de susceptibilidad a virus (Peña *et al.,* 2010) sugieren que las diferencias genómicas y metabolómicas apreciadas pueden afectar a proteínas superficiales haciendo que los mecanismos de evasión a fagos sean distintos. Por lo tanto, estas diferencias genómicas entre M8 y M31 no pueden considerarse neutrales desde una perspectiva ecológica, pudiendo afectar a la microdiversificación y evolución de la especie ya no sólo por su presión selectiva, sino por los fenómenos de transferencia horizontal de genes en los que participan partículas víricas. Del mismo modo, las distintas tasas de crecimiento en experimentos de respuesta a condiciones de

Tabla I3. Cepas *de S. ruber* analizadas desde 1999 a 2013 (Peña *et al*., 2014).

Área geográfica	Cepas	Lugar y fecha de aislamiento	Tipo de análisis	Referencias
Mediterránea	M1, M8[a,b], M31[a,b]	Salinas de Campos, Mallorca, 1999	PFGE, LGT, metabolómica , genómica, MLSA, filogenia	Antón *et al*., (2002), Peña *et al*, (2005), Rosselló-Mora *et al*., (2008), Soria-Carrasco *et al*., (2008). Peña *et al*., (2010). Antón *et al*., (2013)
	RM30, RM84, RM101, RM103, RM117, RM129, RM131, RM141, RM150, RM158, RM159, RM172, RM174, RM179, RM186, RM216, RM224, RM225, RM240, RM272, RM20	Salinas de Campos, Mallorca, 2006	PFGE, metabolómica, LGT	Antón *et al*., (2013)
	Pola 13[a,b], Pola 18[a,b]	Salinas Bras del Port, Alicante, 1999	PFGE, MLSA, LGT, metabolómica	Antón *et al.,* (2002), Peña *et al*., (2005), Rosselló-Mora *et al*., 2008 Antón *et al*., (2013).
	SP3, SP7, SP8, SP10, SP11, SP15-24, SP26, SP28-30, SP32, SP35, SP36, SP38-40, SP51, SP57, SP73, SP79, SP84, SP86, SP99, SP100	Salinas Bras del Port, Alicante, 2007	PFGE, MLSA, LGT, metabolómica	Antón *et al*., (2013)
	E1[a], E3[a,b], E7[a,b], E11, E12[a]	San Carles de la Ràpita, Tarragona, 2001	PFGE, MLSA, metabolómica , LGT	Peña et al., 2005 Rosselló-Mora *et al*., 2008 Antón et al., 2013
	IL3[a]	Ses Salines, Ibiza, 2001	PFGE, MLSA, LGT, metabolómica	Peña et al., (2005) Rosselló-Mora et al., (2008) Antón *et al*.,(2013)
	A1	Salinas deS'Avall, Mallorca, 2001	PFGE, MLSA, LGT	Peña et al., 2005
	ES4[a]	Salnas Eilat, Israel , 2001	PFGE, MLSA, LGT	Peña et al., 2005 Rosselló-Mora *et al*., 2008
Atlántica	C3[a], C4[a], C5, C6[a], C7, C9[a,b], C12[a], C14[ab], C15[a], C16, C17[a], C18, C22[a], C24, C25A[a], C26[a], C27[a], C28, C29[a]	Gran Canaria, Islas Canarias, 2001	PFGE, MLSA, LGT, metabolómica	Peña *et al*., 2005 Rosselló-Mora *et al*., 2008 Antón *et al*., (2013)
Peruana	PR1[a,b], PR2, PR3[a,b], PR4, PR6[a], PR8[a]	Salinas Maras, Perú, 2001/2003	PFGE, MLSA, LGT, metabolómica	Peña *et al*. (2005) Rosselló-Mora *et al*., (2008) Antón *et al*., (2013)

[a] Cepas analizadas en Peña *et al*, (2005). [b] Cepas analizadas por MLSA en Rosselló-Mora *et al*., (2008).
PFGE, Electroforesis en gel de campo pulsado; LGT, transferencia horizontal de genes; MLSA, análisis mediante multilocus.

estrés (desecación o radiación) y de competencia entre las cepas M8 y M31 de *S. ruber* apoyarían que dos cepas genéticamente similares pueden presentar comportamientos ecológicos distintos frente a determinadas condiciones (Peña *et al*; 2010).

Con el objetivo de explorar si esta diversidad metabólica es un atributo extensible a todas las cepas de *S. ruber* se realizó un estudio extensivo metabolómico (Antón *et al*., 2013), incluyendo 57 nuevos aislados (22 coaislados de Mallorca en 2006 y 35 de Santa Pola en 2008) y un subgrupo de aislados de la colección antigua (M8, M31, Pola13,Pola18 y IL3) (**tabla I2**). La colección antigua incluye 35 cepas aisladas de áreas mediterráneas, peruanas y atlánticas, entre ellas algunos aislados empleados en la descripción de la especie (Antón et al., 2002) y los usados en los primeros estudios de microdiversidad (Peña *et al*., 2005) y metabolómicos (Rosselló-Mora *et al*., 2008) con *Salinibacte*r. Como primera conclusión, se observó una enorme diversidad de metabolitos en las tres fracciones analizadas (extracelular, citoplasmática y celular insoluble). En segundo lugar, las cepas de la colección antigua formaron un agrupamiento claro independientemente del lugar de aislamiento. Por último aquellas cepas que fueron aisladas al tiempo en un mismo punto geográfico presentaron perfiles metabolómicos diferentes. La mayoría de metabolitos anotados pertenecieron a las rutas metabólicas de lípidos, aminoácidos, biosíntesis de metabolitos secundarios, metabolismo de otros aminoácidos y metabolismo de terpenoides y policétidos. Dentro del metabolismo de lípidos, las rutas de biosíntesis de ácidos grasos y metabolismo de esfingolípidos acumularon la mayoría de cambios en metabolitos que permitieron establecer las agrupaciones biogeográficas mencionadas. Así pues, el metabolismo de lípidos parece ser la vía metabólica más versátil en *S. ruber*.

1.1.4- Microdiversidad genómica.

Con el objetivo de caracterizar la diversidad intraespecífica existente en *S. ruber* se han llevado a cabo estudios filogenéticos con cepas aisladas pertenecientes al filotipo EHB-I basados en el análisis del operón ribosómico (**Tabla I2)**, comparando las secuencias tanto de los genes del operón ribosómico como las sus espaciadores intergénicos con una amplia colección de aislados (Antón *et al*., 2002; Peña *et al*., 2005). Una de las conclusiones más relevantes de estos

estudios fue la elevada homogenidad a nivel del operon ribosómico entre las cepas comparadas. Sin embargo, el uso posterior de técnicas de análisis genómico como PFGE y la Amplificación al Azar de DNA Polimórfico (RAPD, del inglés Random Amplified Polymorphic DNA), comparando patrones genómicos de restricción y amplificación respectivamente, revelaron diferencias en cuanto al contenido y la distribución de material genético (Peña *et al*; 2005). Se utilizaron los patrones de PFGE para explorar una posible distribución biogeográfica de los genotipos estudiados y, aunque no se observó una distribución clara (Peña *et al.,* 2005) con estos datos, sí se pudo discernir un patrón biogeográfico tras el análisis metabolómico de las mismas cepas (Roselló-Mora *et al*., 2008) como se comentó anteriormente. El empleo de FPGE ha mostrado patrones distintos para cada nueva cepa *de S.ruber* aislada (**figura I2**), lo que refuerza la idea de que *S. ruber* representa un ejemplo de microdiversificación a corta escala, aunque ha de considerarse que tales diferencias podrían explicarse mediante reordenamientos génicos sin la implicación de diferencias notables en el contenido génico. Esta diversidad intraespecífica se ha observado también en el patrón de abundancia de plásmidos en diversas cepas (Peña *et al*., datos no publicados).

Con el objetivo de profundizar en la caracterización de la distribución biogeográfica existente en la colección de cepas de *S. ruber* y sus patrones de diversificación, se llevó a cabo un MLSA (**Tabla I2**), empleando 10 cepas y 8 genes *housekeeping* considerados filogenéticamente informativos (Roselló-Mora *et al*., 2008). A diferencia de lo observado en otros organismos extremófilos (Whitaker *et al.*, 2003), las reconstrucciones filogenéticas derivadas de este análisis no mostraron una clara segregación geográfica. Los arboles filogenéticos de cada uno de los 8 genes resultaron incongruentes (**Figura I3**), lo cual es esperable cuando la recombinación actúa durante la divergencia de cepas (Papke *et al*., 2004). Mediante un análisis de descomposición de divisiones (*split decomposition*) (Bandelt y Dress, 1992) se analizó en profundidad la posibilidad de que la recombinación estuviera influyendo en la diversificación de la especie. Este análisis se realizó para los 8 genes de las 10 cepas dando como resultado una red más que un árbol claramente resuelto (Rosselló-Mora *et al*., 2008), mostrando la presencia de divisiones paralelas debido a señales conflictivas (que pueden derivarse de eventos de recombinación). Esto, junto con la incongruencia de las filogenias de los

genes individuales, apunta al papel de la recombinación como un mecanismo predominante en la evolución de la especie. Si la recombinación actúa como elemento de diversificación u homogenización en esta especie en concreto es una cuestión que debe explorarse con mayor detenimiento, dado que ambos escenarios podrían derivarse de la misma (Papke *et al*., 2007).

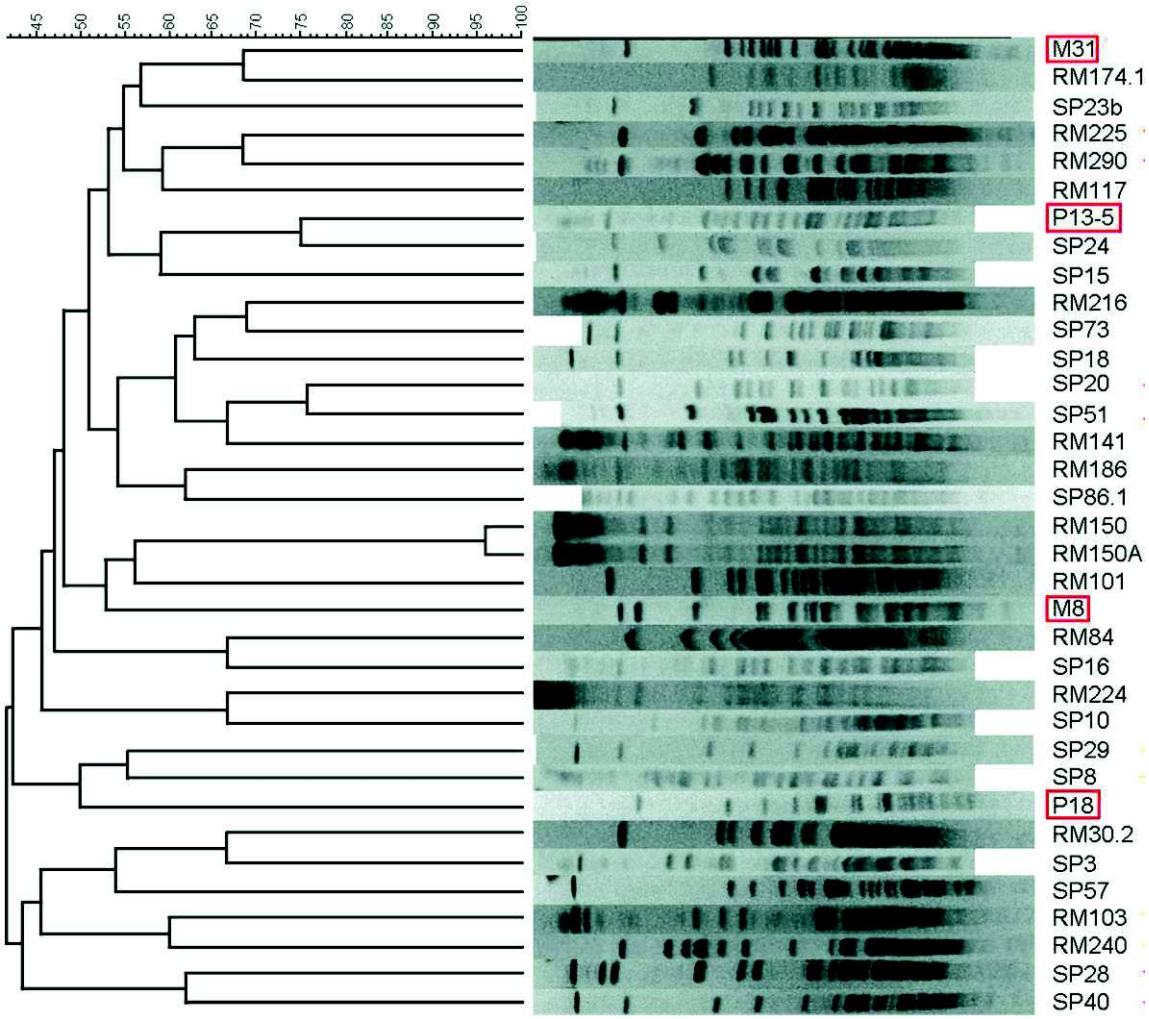


Figura I2. Dendrograma de similitud (izquierda) obtenido mediante el análisis de presencia/ausencia de bandas de productos de digestión genómicos con XbaI separados por PFGE para diferentes cepas de *S. ruber* (derecha). Encuadradas en rojo las cepas empleadas en el año 2000 para la descripción de la especie (adaptada de Antón *et al*., 2013).

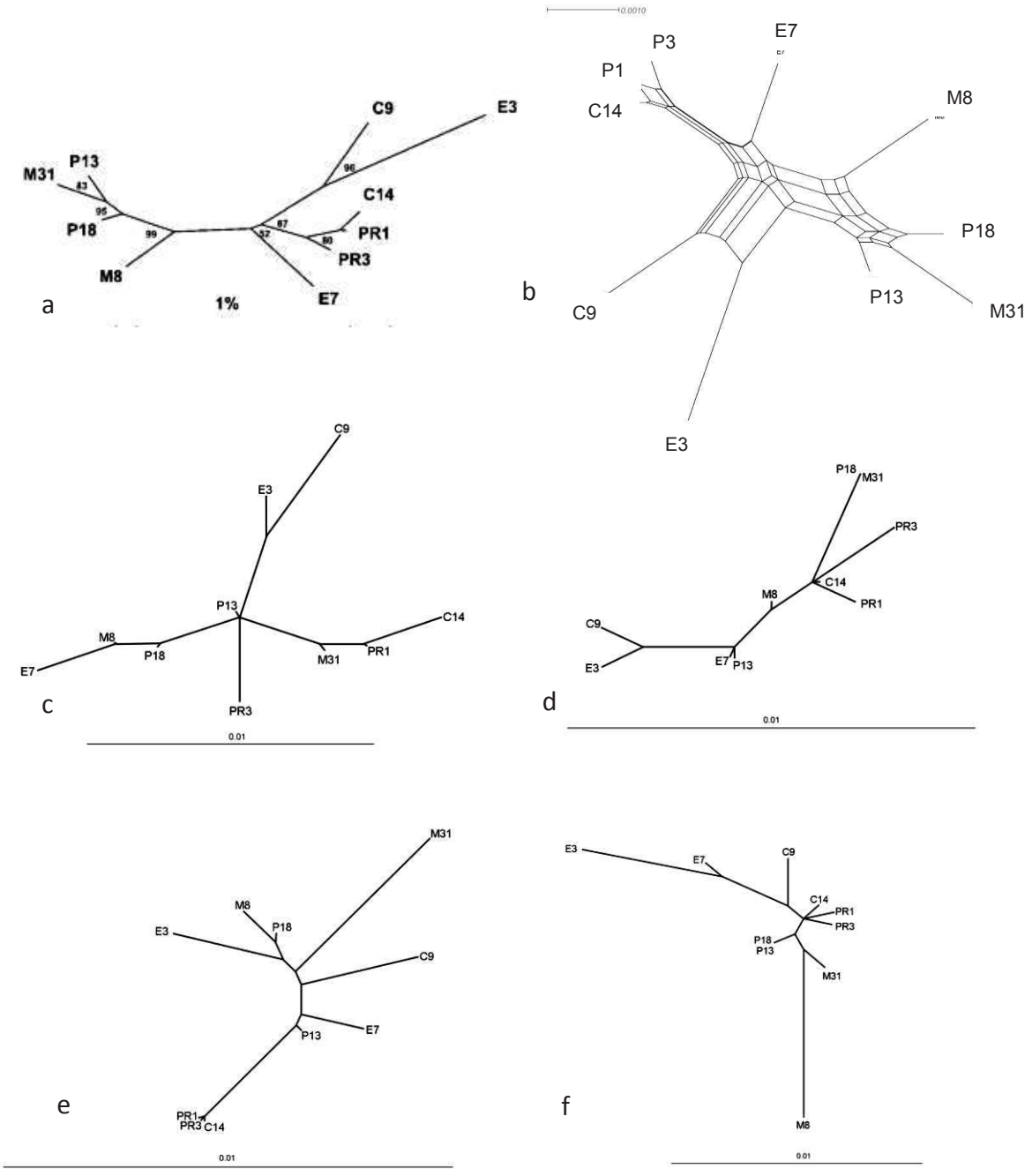


Figura I3. Figura A): Reconstrucción filogenética basada en el algoritmo PHYML empleando el alineamiento de 7995 nucleótidos correspondientes a 8 genes *housekeeping*, incluyendo SSU rRNA (Rosselló-Mora *et al*., 2008). Figura B): Análisis de descomposición de divisiones de los datos mostrados en la figura A. Figuras C) D) E) y F): Árboles filogenéticos de 4 de los 8 genes empleado en MLSA de 10 cepas de *S.ruber* aisladas en diferentes localizaciones: C=*eno*F (enolasa) D= *gap* (gliceraldehido-3-fosfato deshidrogenasa) E= *pyr*G (CTP sintasa) F= S5 (proteína ribososómica).(Adaptado de Peña *et al*., 2014).

1.1.4.1- El genoma de *S.ruber*.

Hasta la fecha se ha llevado a cabo la secuenciación de dos cepas de *S. ruber*: la cepa tipo M31 y M8, ambas aisladas al mismo tiempo en las salinas de Campos de Mallorca junto a una tercera cepa, M1. De todas las cepas de *S. ruber* aisladas, M8 y M31 son las más cercanas filogenéticamente tal y como muestran los patrones obtenidos con el uso de las técnicas PFGE y RAPD (Peña *et al*., 2005). Ambas presentan idéntica secuencia a nivel del operón del rRNA, incluyendo los espaciadores, lo que impide la detección selectiva de las mismas mediante FISH con sondas marcadas para dicha región.

Si bien con anterioridad se disponía de secuencias parciales (Peña *et al*., 2005), en el año 2005 se publicó la secuencia completa del genoma de la cepa tipo *S. ruber* M31 (DSM 15388) (NC_007677.1) (Mogondin *et al.,* 2005). El genoma de M31 se compone de un plásmido de 35.505 pb con un contenido en G+C del 57.9% y un cromosoma de 3.551.823 pb con un contenido en G+C de 66.29%. Se identificaron mediante búsqueda automática un total de 2.934 ORFs en el cromosoma y 33 en el plásmido. El análisis del contenido en G+C a lo largo del cromosoma de M31 permitió describir tres zonas que presentaban un contenido inferior al valor promedio del resto del cromosoma y a las que se nombró posteriormente islas genómicas (Pasic *et al.,* 2009).

Posteriormente se secuenció el genoma de la cepa M8 (NC_014032.1), de 3,619,417 pb con un porcentaje de G+C del 66,12% (Peña *et al.,* 2010) (**figura I4**). El genoma de esta cepa se compone de un cromosoma y 4 plásmidos. Se detectaron y anotaron un total de 3.086 genes en el cromosoma. Su proteoma teórico presentó la distribución bimodal típica de organismos halófilos con un punto isoeléctrico de 5,05. La **tabla I3** muestra la comparación de las características genómicas más relevantes de las cepas M8 y M31.

La secuenciación de los genomas de las cepas M8 y M31 permitió realizar un estudio comparativo (Peña *et al*., 2010) con el fin de evaluar la diversidad funcional y genómica existente entre dos cepas tan cercanas, su significado ecológico así como el origen de la misma, constituyendo el primer análisis de microevolución llevado a cabo con microorganismos aislados de este ambiente. Pese a ser muy semejantes, cada una de las cepas contiene un conjunto

Tabla I3. Comparación de las principales características genómicas de las cepas M8 y M31 (DSM 13588) *de S. ruber* (Peña *et al*., 2011).

	S. ruber M8				*S. ruber* M31 DSM 13588
Cromosoma					
Logitud (Kb)	3.619.447				3.551.823
%GC	66.12%				66.29%
Número ORF	3086				2934
rRNAs	3				3
tRNAs	43				44
Número total de sitios alineados	3.348.702				3.283.845
% ANI total	98.45				nd
% Similitud (aa)	94.3				
% Similiud (nt)	93.5				
Ortólogos totales	2604				
Plásmidos	4				1
Nombre	pSR11	Psr56	pSR61	pSR84	pSR35
Longitud (Kb)	11,23	56,53	61,37	84,34	35,5
% GC	63,29	60,03	59,58	63,19	59,70
Número ORF	13	38	50	70	33
Ortólogos totales	1	11	16	28	Nd

de genes específicos que no comparte con la otra y que constituyen un 10% del genoma. Los genes compartidos, constituyen el *core* genoma de la especie, entendiendo como tal el conjunto de genes que están presentes en todos los miembros de una especie, mientras que los específicos de cepa constituyen el genoma accesorio de la misma. El *pangenoma* de la especie estaría constituido por la suma de ambos (Mira *et al*., 2010). De entre los genes compartidos, el 22% son específicos de especie al no presentar homólogos con genes de otros microorganismos depositados en las bases de datos. Su comparación permitió detectar una región singular de más de 300 kb, a la que se denominó zona conservada (CR), en la cual no se apreciaban sustituciones no sinónimas, siendo su dN/dS cero. El alineamiento de los genomas de ambas cepas permitió apreciar inserciones y delecciones de genes, que junto a los genes divergentes, aquellos que

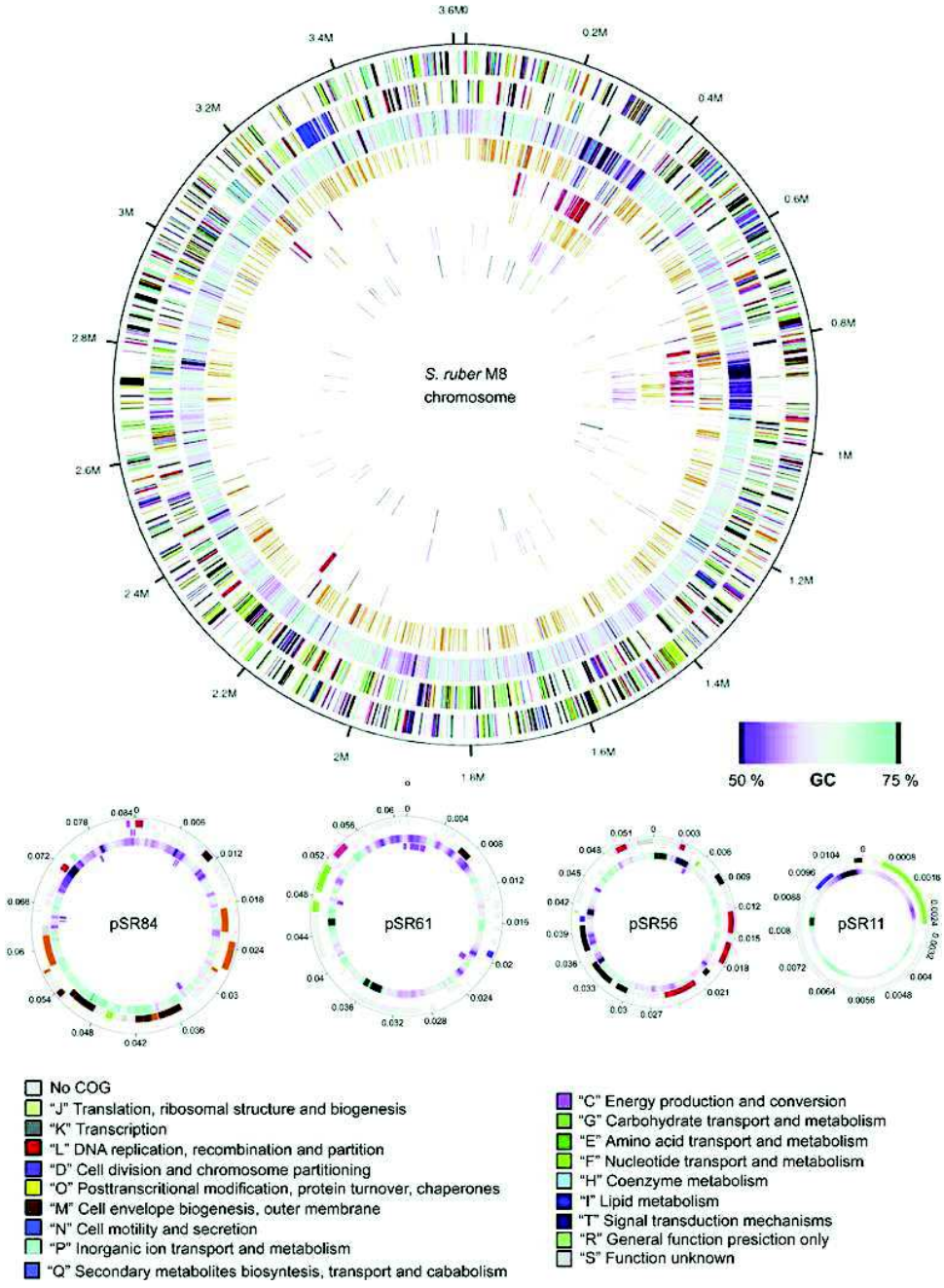


Figura I4. Representación circular de los replicones de M8. Para el cromosoma, de fuera hacia dentro**: Círculos 1-2**: anotación COG**. Círculo 3**: porcentaje GC**. Círculo 4**: genes divergentes en *S. ruber* M8/M31 marcados en azul y naranja los genes con un % identidad por debajo del 50% o 90%respectivamente. **Círculo 5**: genes presentes en M8 y otros organismos de la base de datos pero no en M31. **Círculo 6**: Genes específicos *de S . ruber* M8. **Círculo 7**: transposasas. **Círculo 8**: LGT desde *Archaea.* Para los plásmidos: Los dos círculos externos representan las ORFs en las cadenas + y -, coloreada según su categoría COG (adaptada de Peña *et al*., 2010).

presentan una identidad menor al 90%, dispersos en el cromosoma, constituyen la mayor parte de las diferencias existentes entre ambas. Este alineamiento además permitió apreciar una gran sintenia entre ambos genomas, detectando tres zonas de discontinuidad entre ellos (**figura I5**). Dos de ellas correspondieron con zonas de contenido génico distinto, mientras que la tercera correspondió con una inserción en el genoma de M3 que alberga el sistema de modificación-restricción de esta cepa, situado en un plásmido en el caso de M8. Las dos primeras zonas de discontinuidad se denominaron regiones hipervariables (HRVs) (correspondientes a 2 de las anteriormente mencionadas islas genómicas) y presentaron % G+C menor que el del resto del genoma así como un índice de uso de codones (CAI) bajo. Las diferencias entre los genomas de las cepas M8 y M31 de *S. ruber* se concentran en estas zonas concretas del genoma. En ellas se hallan sobre-representados los genes específicos y los muy divergentes así como los involucrados funcionalmente en la síntesis de las envolturas celulares (pared y membrana) entre los que se incluyen sulfotransferasas y glicosiltransferasas. Además, contienen genes de transposasas, que favorecerían eventos de recombinación diferentes en cada cepa.

1.1.4.2- Transfencia horizontal en *S.ruber*.

Como se mencionó anteriormente, *S. ruber* comparte muchas características fenotípicas con las *Haloarchaea* de su ambiente, por lo que pese a ser un representante del filo *Bacteroidetes*, fenotípicamente se comporta como una *Archaea* halofílica. De este modo *S. ruber* parece un claro candidato como ejemplo de organismo en el cual genes adquiridos mediante transferencia horizontal de genes (HGT) inter-dominio otorgan una clara ventaja adaptativa. Estudios filogenéticos confirmaron la presencia de 40 genes HGT inter-dominio en M8, 34 de los cuales estaban presentes en la cepa M31 formando parte del *core* genoma (Peña *et al*., 2010). Esto, junto al hecho de presentar la mayoría de ellos un GC similar al del del genoma y una identidad de secuencia nucleotidica (ANI) con su ortólogo en la otra cepa generalmente mayor del 90%, sugiere que se traten de eventos de transferencia antiguos, acontecidos en el origen de la especie. Una parte de estos genes codifican para transportadores e iones, sistemas de transducción de señales y proteínas relacionadas con las rodopsinas.

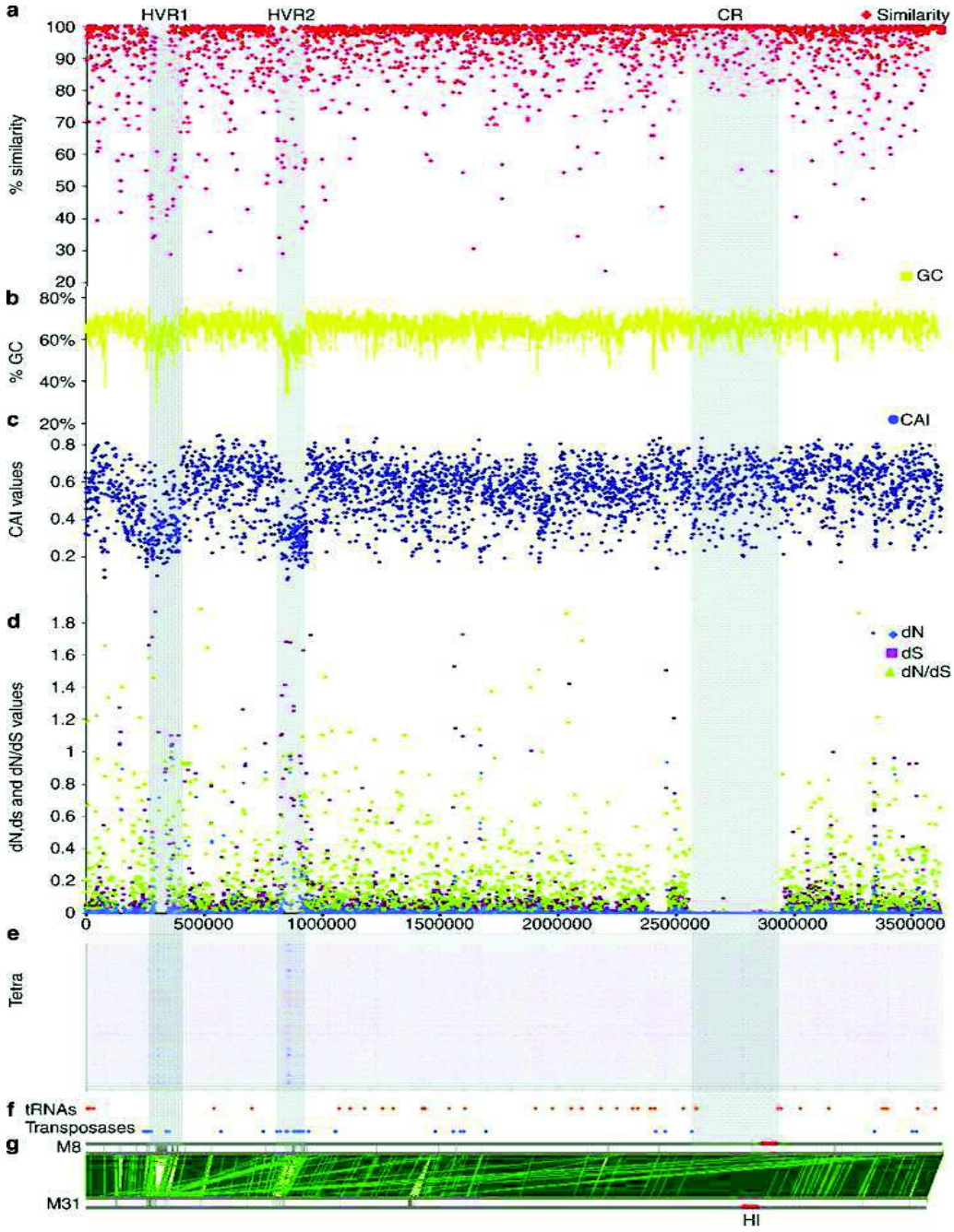


Figura I5. Resumen y datos comparativos del cromosoma de S.ruber.M8. (a) Identidad respecto a los ortólogos de M31. (b) %GC. (c) CAI. (d) dN, dS y dN/dS para los homólogos. TETRA (patrón de frecuencia de tetrnucleótidos en secuencias de ADN). (f) RNAs transferentes (tRNAs) y transposasas a lo largo del cromosoma. (g) Alineamiento completo de los genomas de M8 y M31 desde el origen de replicación (Peña *et al*., 2010).

Con el objetivo de discernir si existía alguna tendencia biogeográfica en los patrones de distribución de los HGT, se analizaron los patrones de presencia/ausencia de los 40 genes en 92 cepas de *S.ruber* de áreas biogeográficas diversas: mediterránea (Santa Pola, Mallorca, Ebro), atlántica (Canarias) y peruana (**tabla 2I**). Dado que los ambientes hipersalinos presentan una distribución parcheada en la Tierra, la presencia de poblaciones microbianas endémicas en este tipo de ambiente apoyaría un posible escenario de especiación alopátrica (Zhaxybyeva *et al*., 2013). Aunque se consideró un conjunto de cepas diferente, a diferencia de los análisis

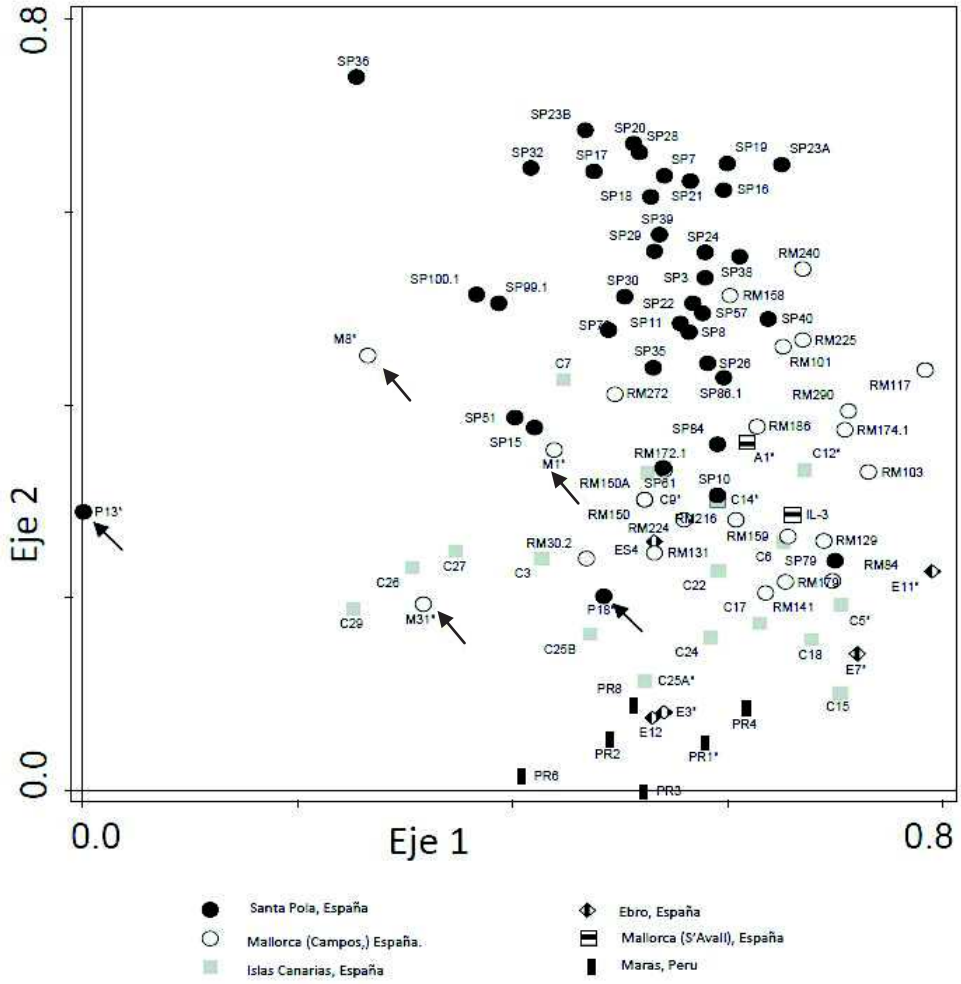


Figura I6. Análisis de redundancias de los datos de presencia/ausencia para los 40 HGT distribuidos en las 92 cepas de *S.ruber*. Las flechas señalan las cepas aisladas en 1999 de las salinas de Santa Pola (círculos negros) y Mallorca (círculos blancos) (Adaptada de Peña *et al*., 2014).

metabolómicos mencionados anterioemente, los análisis MLSA basados en marcadores genéticos no mostraron ningún agrupamiento geográfico. Las 92 cepas analizadas incluyeron cepas empleadas en la descripción de la especie (Antón *et al*., 2002). los primeros estudios de microdiversidad (Peña *et al*., 2005), en los primeros estudios metabolómicos (Rosselló-Mora *et al*.,2008), y las 22 cepas coaisladas en Mallorca (2006) las 35 de Santa Pola (2008), incluidas en el último estudio metabolómico (Antón *et al*., 2013) (**tabla I2**). Los patrones de presencia/ausencia se analizaron mediante un análisis de correspondencias (CA), observando una relación clara entre la distribución de cepas y el punto de aislamiento de las mismas (**figura I6**), lo que indicaría que cepas relacionadas geográficamente han incorporado genes HGT similares. Además se apreciaron diferencias en los agrupamientos de cepas aisladas en el mismo punto geográfico en diferentes tiempos (M1, M8 y M31 aislados en 1999; cepas de RM, P13 y P18 en 2006). Estos resultados muestran nos solo el enorme grado de microdiversidad genómica sino la tasa de cambio elevada a la que se dan estos (menos de 10 años).

1.2- Mecanismos de diversificación intraespecífica.

Los procesos de generación de diversidad en microorganismos han sido ampliamente estudiados. Los procariotas contienen genomas altamente dinámicos que reordenan e intercambian, adquieren o pierden, información genética relevante (Dutta y Pan., 2002, Abby y Dauvin., 2007). Los mecanismos que contribuyen a la diversificación génica en procariotas se clasifican en dos grupos: i) Modificaciones de la información génica interna de secuencias preexistentes por mecanismos de mutación o recombinación homóloga intergénica, ii) Pérdida o adquisición de genes mediante mecanismos de transferencia horizontal. En procariotas se pueden producir duplicaciones o pérdidas de genes pre-existentes, por escisión o formación de pseudogenes, o adquirir genes mediante con procedencias distintas dentro de un genoma mediante procesos de transferencia horizontal (HGT).

1.2.1- Transferencia horizontal de genes y mecanismos de captación de DNA.

La HGT es un fenómeno ampliamente distribuido en los tres dominios del árbol de la vida además de en partículas víricas. Consiste en el intercambio de DNA entre dos organismos y su establecimiento en el organismo receptor. Es el principal mecanismo de generación de diversidad microbiana junto con los procesos de mutación, contribuyendo notablemente a la adaptación y evolución del organismo receptor (Ochman *et al*., 2000, Thomas y Nielsen, 2005). El impacto y la frecuencia con que tiene lugar este tipo de fenómeno depende de diversos factores: la microbiota existente en determinado ambiente, la presión selectiva del mismo y los mecanismos barrera o de captación del organismo receptor del organismo receptor (Thomas y Nielsen, 2005). Trabajos llevados a cabo con genomas procariotas completos estiman que el porcentaje de genes de HGT oscila entre 1,56 y 14,47% del total del genoma (García-Vallvé *et al*., 2000).

La HGT es un proceso que implica varias etapas desde la captación del DNA: i) su entrada a la célula y selección, ii) su integración y estabilización mediante diferentes mecanismos de recombinación en la célula receptora y iii) su expresión y replicación adecuada en el entorno génico (Thomas y Nielsen, 2005, Barkay y Smets, 2005) **(figura I7**). Tradicionalmente se han considerado en organismos procariotas 3 mecanismos principales de transferencia de material genético: transformación, conjugación y transducción (**figura I8**).

La transformación consiste en la captación de DNA libre en el ambiente para su posterior integración en el genoma (Chen *et al*., 2005). La transformación natural es un proceso vinculado al estado de competencia celular, aquel que permite que una célula pueda incorporar DNA libre del medio en condiciones naturales de crecimiento. Se han detectado transformantes naturales en los principales grupos taxonómicos de *Bacteria* y *Archaea* (Kleter *et al*., 2005). A diferencia de los genes responsables de los mecanismos de conjugación o transfección, usualmente codificados en elementos genéticos móviles (MGE), los responsables del estado de competencia suelen albergarse en el cromosoma repartidos por diferentes regiones. Generalmente, la competencia está regulada por un conjunto de 20-50 genes, el regulón *com*, cuyas proteínas interaccionen estructuralmente entre sí de manera coordinada. Muchos de los genes implicados

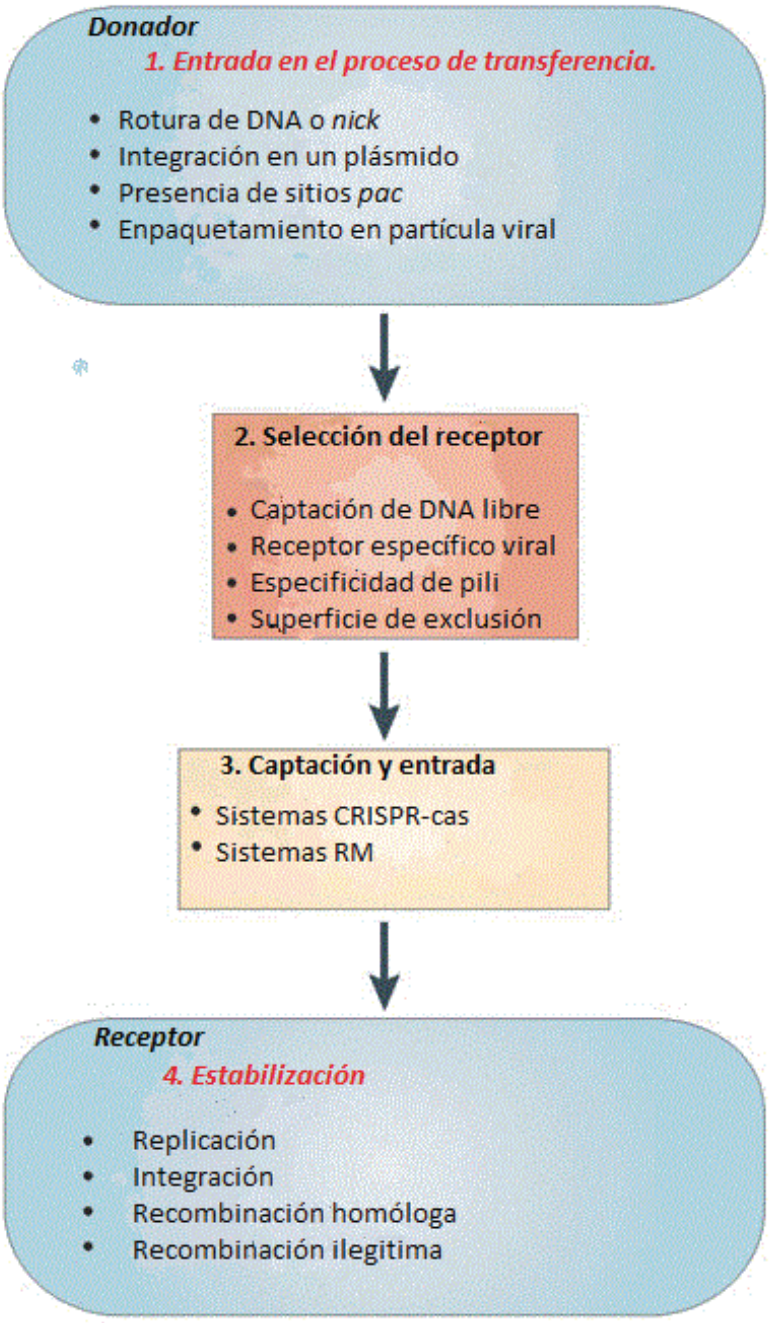


Figura I7. Principales etapas del proceso de HGT en procariotas. Cada recuadro representa una etapa en el proceso de transferencia desde el organismo donador al receptor, detallándose factores y eventos destacables en cada una de ellas (adaptada de Thomas y Nielsen.,2005).

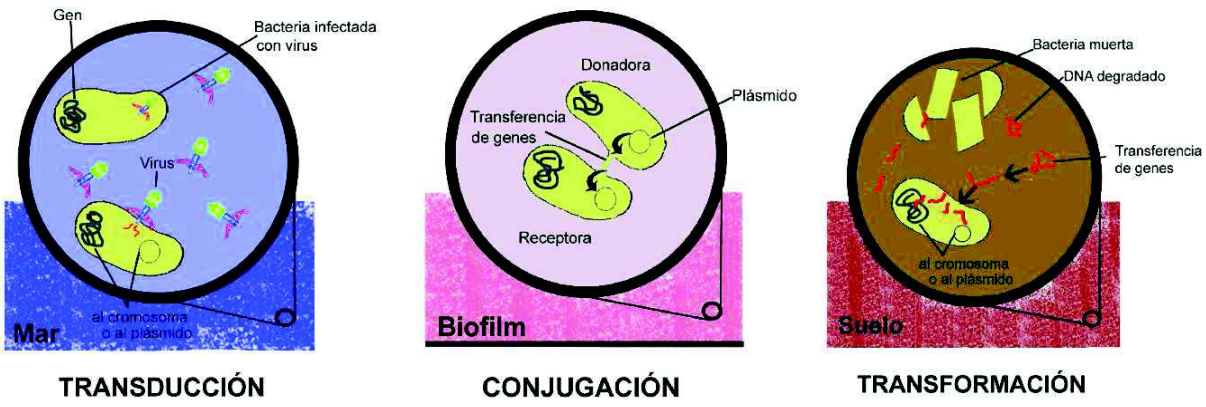


Figura I8. Principales mecanismos de HGT en procariotas: Transducción, conjugación y transformación (adaptada de Barkay y Smets., 2005).

están ampliamente conservados entre bacterias Gram positivas y Gram negativas, aunque existen especializaciones para acomodar las diferencias estructurales de sus envolturas. Las maquinarias de transformación se han estudiado ampliamente en diferentes especies bacterianas como *Bacillus subtilus, Streptococcus pneumoniae, Haemophilus influenza, Neisseria gonorrhoeae y Helicobacter pylori* (González-Candelas y Francino., 2012). En muchos de estos casos, los genes implicados guardan relación estructural con los sistemas de secreción pili tipo IV (Imam *et al*., 2011; Claverys *et al*., 2006). En el caso de organismos Gram positivos como *S. pneumoniae y B. subtilus,* se ha caracterizado la cascada de mecanismos de inducción en situaciones de estrés para los genes contenidos en el operón *com*, actuando en la misma mecanismos de activación/inhibición de tipo *quorum sensing* (Claverys *et al.,* 2006).

La conjugación es un mecanismo de trasferencia de DNA célula-célula específicamente relacionado con la tranferencia de elementos plasmídicos debido normalmente a que el plásmido transferido, plásmido conjugativo, contiene elementos génicos que complementan los presentes en la célula receptora y restituyen el sistema de conjugación. Se trata de un proceso unidireccional en el cual las células que contienen el plásmido actúan como donadoras y las que carecen del mismo como receptoras. Como la transformación, es un mecanismo de transferencia de DNA ampliamente distribuido en *Bacteria* e incluso en *Archaea* (Stedman *et al*., 2000). La maquinaria de conjugación normalmente está codificada en genes *tra* localizados en plásmidos conjugativos y codifican para un *pili* especializado, en Gram negativos, que conecta la célula

donadora y la receptora y a través del cual se transfiere una copia del plásmido generada por círculo rodante. En gram positivas, uno de los mecanismos de conjugación mejor descritos es el sistema es el de secreción tipo IV (Imam *et al*., 2011). En este grupo bacteriano los plásmidos conjugativos codifican proteínas de adhesión de membrana que contribuyen a la agregación celular (Burrus *et al*., 2002).

El tercer mecanismo de transferencia de DNA es la transducción, que consiste en transferencia de material genético de una célula a otra por medio de un virus que contiene parte del material genético de la célula donadora encapsidado (Jiang y Paul., 1998). El tamaño de la molécula transferida está limitado por la capacidad de la cápside del virión. En el caso de virus lisogénicos o temperados, tras una infección, su DNA es capaz de integrarse en el genoma del hospedador. Cuando se dan las señales ambientales oportunas, los virus se escinden del genoma y entran en fase lítica. Muchas veces la escisión no resulta precisa, arrastrando parte del cromosoma del hospedador cercano al punto de integración y empaquetándolos en el virión. Este virión transportará el material genético de la bacteria donadora a la receptora, que puede estabilizarlo mediante recombinación homóloga reemplazando una región análoga o mediante su integración en islas genómicas (apartado 1.2.3). En muchas comunidades microbianas, como es el caso de las comunidades de halófilos extremos, los fagos juegan un papel importantísimo, presentando una abundancia de hasta dos órdenes de magnitud mayor que la de los procariotas de esta comunidad (Guixa-Boixerau et al., 1996, Santos *et al*., 2010). Es en estos casos en los cuales los fagos pueden jugar un papel predominante en los procesos de transducción y en la evolución de las poblaciones microbianas (Rodríguez-Valera *et al*., 2009), más aun cuando hay indicios de que los virus de un determinado ambiente pueden infectar a hospedadores de otros ecosistemas (Breitbart y Rohwer., 2005, Short y Suttle 2005), pudiendo actuar como vehículos transmisores de genes entre ecosistemas.

1.2.2- Recombinación homóloga: funciones biológicas e implicaciones ecológicas.

Los procesos de recombinación son los que permiten la fijación e integración del DNA captado por alguno de los tres mecanismos descritos anteriormente y una vez evadidos los sistemas barrera de los que se hablará posteriormente (apartado 1.2.4). A excepción de la entrada de un plásmido o replicón independiente, la estabilización del DNA tendrá lugar mediante recombinación, conduciendo a la modificación de secuencias internas o a adquisición de nuevos genes. Existen tres tipos de recombinación: homóloga (HR) y no homóloga, diferenciando entre ilegítima (IR) y específica de sitio en este segundo caso.

La recombinación homóloga implica el intercambio de fragmentos genéticos o genómicos adquiridos reemplazando otros homólogos preexistentes. Involucra la presencia de regiones con identidad elevada y está mediada por la proteína *RecA*. La recombinación no homóloga específica de sitio requiere de pequeñas regiones de homología entre ambas moléculas de DNA, la foránea a integrar y la receptora, y suele involucrar, aunque no necesariamente, una integrasa de fago, mientras que la ilegítima esta mediada por transposasas y no requiere de ninguna región homóloga (Thomas y Nielsen., 2005).

Además de por recombinación homóloga y no homóloga, la integración de DNA foráneo puede darse mediante mecanismos de recombinación ilegítima facilitada por homología (de Vries y Wackernagel., 2002), en la que uno de los flancos de la región integrada presenta homología y por tanto participa el sistema de recombinación homóloga. La eficiencia de los mecanismos de recombinación no homóloga es de varios órdenes de magnitud menor que la de la homóloga (Brigulla y Wackernagel., 2010). Además de la recombinación homóloga y la ilegítima, otros mecanismos de integración pueden favorecer la estabilización del DNA foráneo. Existen multitud de enzimas codificadas en elementos móviles que favorecen la integración de los mismos sin necesidad de la presencia de regiones homólogas. Un ejemplo son las recombinasas específicas de sitio responsables de la inserción de fagos temperados y elementos conjugativos integrativos como los transposones (apartado 1.2.3).

Atendiendo a los mecanismos de recombinación involucrados, es posible clasificar los eventos de recombinación en dos categorías: homólogos y no homólogos. Los primeros

implicarían la transferencia de contenido génico similar al de la célula receptora reemplazándose el preexistente mediante mecanismo de recombinación homóloga. Este proceso se conoce como conversión génica e involucra generalmente organismos con una identidad de secuencia elevada (generalmente superior al 99%). Es por tanto un proceso que tiene lugar entre cepas cercanas filogenéticamente con una elevada identidad de secuencia y contribuiría al intercambio de secuencias homólogas ortólogas, aunque de tratarse de grandes regiones genómicas pudieran aportar cierto contenido no homólogo, siempre que estén flanqueadas por secuencias de elevada homología. La recombinación no homóloga contribuiría a la adquisición de secuencias nuevas o a ortólogos de otras especies. Tal como se ilustra en la **figura I9**, atendiendo a su relación entre las secuencias preexistentes en la célula receptora y las incorporadas mediante transferencia horizontal podrían clasificarse en:

a) Adquisición de un nuevo gen no presente en la célula receptora por ser parte del genoma accesorio o procedente de un clado filogenéticamente distante.

b) Adquisición de un parálogo de determinado gen procedente de un ancestro evolutivamente distante.

c) Adquisición de un ortólogo filogenéticamente distante seguido de la eliminación del gen ancestral por el xenólogo (ortólogo incongruente filogenéticamente).

Actualmente encontramos 3 hipótesis que ponen de manifiesto los beneficios evolutivos de la recombinación homóloga en organismos procaritotas. El primero de ellos es la de reparación del DNA dañado, en la cual el DNA foráneo actuaría como molde para reparar roturas de doble hebra (Bernstein *et al*., 1981). Esta hipótesis se ve sustentada por el hecho de que se han observado elevadas tasas de recombinación homóloga en organismos que habitan ambientes con elevado estrés (García-Gonzalez *et al*., 2013). La función de reparación implica en ocasiones recombinación no efectiva, proceso por el cual un fragmento de DNA es reemplazado por otro idéntico. El proceso de reparación involucraría cepas muy cercanas, y aunque en muchas ocasiones no es detectable, se considera muy frecuente y constituiría la función biología más importante de la recombinación homóloga (Michod *et al*., 2008). Existe también una segunda hipótesis alternativa según la cual originariamente los mecanismos de incorporación de DNA se establecieron como mecanismo de captación del mismo como fuente

de nutrientes (Redfield *et al*., 2001).

Por último, en tercer lugar, varias hipótesis postulan que la recombinación homóloga actúa como mecanismo que modula la diversidad al eliminar mutaciones deletéreas e incorporar nuevas (Narra y Ochman., 2006). La estabilización de nuevas secuencias, salvo en la adquisición de plásmidos, requiere de la integración de las mismas mediante procesos de recombinación. Junto con las mutaciones puntuales, la recombinación homóloga contribuye a la variación de los genomas afectando a la clonalidad de la comunidad. Cuando la trasferencia implica la adquisición de nuevos genes flanqueados por regiones homólogas, la adquisición de los mismos bajo condiciones ambientales de selección neutras o positivas conlleva una mejora en los

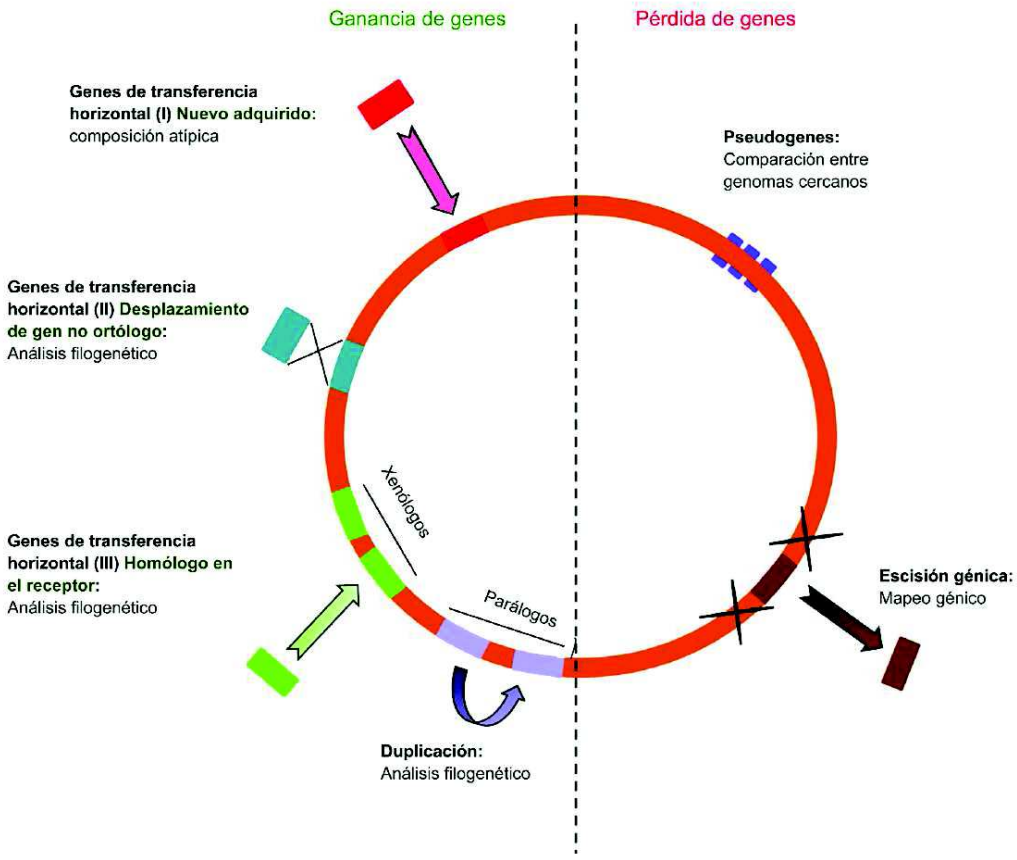


Figura I9. Dinámica y técnicas de detección de mecanismos por los que se produce la variación el repertorio génico en procariotas (adaptación de Abby y Daubin., 2007).

procesos de adaptación al ambiente. Un claro ejemplo es la adquisición de resistencias a antibióticos o de factores de patogenicidad por parte de bacterias patógenas obligadas u oportunistas, en donde el fenómeno de recombinación homóloga se observó por primera vez (Sprat *et al*., 1992; Hacker y Carniel 2001). Sin embargo, los mecanismos de recombinación homóloga, debido al efecto de algunos sistemas barrera (apartado 1.2.4) y a la inefectividad de este tipo de mecanismos a la hora de recombinar secuencias de baja identidad, afectan mayoritariamente a secuencias homólogas entre especies cercanas, apreciándose una clara correlación negativa de las tasas de recombinación y la divergencia de secuencias entre dos organismos (Fraser *et al*., 2007 ; Williams *et al*., 2012). En estos casos actuarían preferentemente mecanismos de recombinación ilegítima o específica de sitio.

Como se comentará más adelante (apartado 1.2.5), hasta hace poco los estudios de recombinación homóloga se basaban en MLSA, empleando muy pocos genes, generalmente menos de 8, y tamaños muestrales pequeños (Vos y Didelot, 2009). Este reducido número de genes no permitía explorar en profundidad su impacto sobre la dinámica génica ni sobre la arquitectura genómica y las regiones afectadas por la recombinación (Didelot *et al*., 2010). En conjunto, existían limitaciones a la hora de tener una idea más aproximada sobre su impacto en la microevolución de las diferentes especies procariotas. El incremento de secuencias de genomas completos está permitiendo abordar algunas de estas cuestiones abiertas hasta la fecha, entre ellas si los procesos de recombinación homóloga tienen un papel homogenizador o no en determinada especie. Análisis genéticos y poblacionales han demostrado que el flujo de genes entre grupos poblacionales filogenéticamente definidos es comparable al derivado de la reproducción sexual en eucariotas.

El impacto de la recombinación homóloga sobre una determinada especie o población y el efecto sobre la microdiversidad y evolución de la misma, homogenizador o diversificador de la especie, dependerá de varios factores. Por un lado influirían factores intrínsecos, de los que se hablará en los siguientes apartados, como los mecanismos barrera de esta especie y la presencia de elementos génicos móviles (MGE) que promuevan el intercambio de DNA en la célula. Entre los factores extrínsecos encontramos la microdiversidad ambiental y el *pool* génico disponible o supergenoma (**figura I11**), entendiendo este último como el conjunto de secuencias accesibles

del ambiente, y del que se hablará más adelante (Hackel y Carniel 2001).

Algunos autores sostienen que los mecanismos de HGT podrían actuar como fuerzas que homogenicen poblaciones de organismos filogenéticamente relacionados (Whitaker *et al*., 2005, Papke 2007, Adam *et al*., 2010), mientras que otros argumentan que la frecuencia en que tienen lugar estos fenómenos en la naturaleza no es la suficiente para observar tal efecto (Cohan 2006). Se han llevado a cabo estudios de simulación para explorar mediante modelos neutrales el impacto de la variación de las tasas de recombinación. El objetivo de estas simulaciones fue el de explorar los niveles requeridos para explicar el agrupamiento o *clustering* filogenético observado en la naturaleza (Fraser *et al*., 2007), representado por poblaciones bacterianas con elevado grado de identidad genómica global. Cuando las tasas de recombinación se sitúan bastante por debajo de las de mutación, la situación correspondería con un escenario poblacional clonal, en el marco de la teoría de ecotipos (Cohan, 2006). Dicha población mantendría un elevado grado de *clustering*, y dichos *clusters* estarían compuestos por líneas clonales sometidas a selección natural. En el caso de que la tasa de recombinación sea mucho mayor que la de mutación, tendríamos un escenario en el cual se equipararía la diversidad de alelos aunque el número de genotipos únicos sería mayor que en el caso anterior, presentándose un *clustering* más pronunciado aunque transitorio. Con valores de relación entre las tasas de mutación y recombinación (r/m) entre 0.25 y 4 se tendrían poblaciones diferenciables y estables. En este último escenario la recombinación actuaría como fuerza cohesiva sobre la población, eliminando los vínculos entre alelos, incluso entre genes colindantes. No se requeriría del efecto de la selección natural a la hora de generar la apariencia de *clusters* únicos, ya que los procesos que afecten a la proliferación y muerte de la población serían suficientes (Fraser *et al*., 2007).

Dentro del rango mencionado para estas tasas, los procesos de recombinación actuarían como mecanismo de cohesión dentro de poblaciones de organismos (Adam *et al*., 2010). Los *clusters* generados, debido a la menor tasa de recombinación con el incremento de la divergencia de secuencias mencionada, podrían aproximarse a lo que se considera una especie. Sin embargo, y aunque con tasas mucho menores, se estima que la recombinación homóloga podría actuar entre especies cercanas. Un ejemplo es el género *Streptococcus*, donde especies como *S. pneumoniae* y *S. pseudoneumoniae* presentan una divergencia de secuencia del 3%, lo que

reduciría sólo su tasa de recombinación 4 veces respecto a la intraespecífica detectada en cada una de ellas. Una segunda consecuencia derivada del proceso de homogenización sería la pérdida de clonalidad dentro de las especies y la incongruencia filogenética observada para determinados genes en comparaciones intraespecíficas pero no interespecíficas.

El modelo planteado por Fraser explicaría las evidencias presentadas en estudios poblacionales de MLSA con amplios tamaños muestrales en *Halorubrum* (Papke *et al.,* 2007) y *Sulfolobus islandicus* (Whitaker., 2005), que incluyen 150 y 60 cepas cercanas respectivamente. En ambos casos se apreció un elevado grado de recombinación homóloga, presentando tasas de recombinación de 2 y 1,2 respectivamente, detectándose incongruencias filogéneticas en diversos genes, derivadas de los procesos de recombinación. En el caso de *Halorubrum* se detectaron diferentes grados de diversidad en los genes analizados, lo cual soporta como hipótesis más probable la selección individual de cada uno de ellos tras el intercambio mediante recombinación homóloga.

El balance entre recombinación como fuerza cohesiva capaz de generar unidades taxonómicas diferenciables y las barreras que impidan dicha transferencia marcarán la diversificación de estos taxones. En el caso de especies de vida libre en ambientes fragmentado como son los halófilos extremos (apartado 1.1.1), las barreras geográficas podrían reducir suficientemente los tránsitos permitiendo la diferenciación de filogrupos o especies (Fuller *et al*., 2014).

1.2.3- Fuentes de microdiversidad: plásmidos, elementos móviles e islas genómicas.

Los eventos de HGT y mutaciones puntuales son responsables de generar gran parte de la diversidad génica, favoreciendo los procesos de adaptación a diferentes ambientes y nuevos nichos ecológicos (Boucher et *al*., 2003, Fernández-Gómez *et al*., 2012). Cada ambiente presenta una microdiversidad y *pool* génico accesible particulares que determinará los *clusters* o grupos poblacionales de intercambio de material genético. El tipo de genes transferido horizontalmente por los miembros de una comunidad varía enormemente dado que la presión selectiva y requerimientos específicos de cada ambiente son diferentes y por tanto también las ventajas

adaptativas derivadas de determinado evento de HGT. El ambiente no sólo determina las condiciones selectivas sino también las adaptaciones génicas disponibles, seleccionando el *pool* génico y la diversidad microbiana, y por tanto las fuentes de microdiversidad, siendo la HGT el mecanismo más común para el acceso a este *pool* (Pallen y Wren 2007). El tamaño y composición del *pool* génico disponible para la incorporación al genoma de una especie se conoce como supergenoma (Norman *et al*., 2009). El repertorio de MGE y su impacto sobre el genoma de una especie a menudo está íntimamente relacionado con capacidades celulares como la competencia natural o la capacidad de conjugación de la especie. Entre las principales fuentes de microdiversidad procariótica y que forman parte del *pool* génico flexible encontramos partículas virales y plásmidos además de elementos conjugativos integrativos (incluyendo transposones, plásmidos integrativos e islas genómicas conjugativas), inserciones génicas e islas genómicas y transposones simples (Hacker y Carniel., 2001) **(figura I10**). Junto a los procesos de mutación puntual y recombinación homóloga que regulan la diversidad alélica poblacional, estos mecanismos de adquisición génica introducen elementos génicos desde el *pool* ambiental, formando parte del genoma accesorio una vez estabilizados. Su estabilización en la célula tendría lugar por mecanismos de recombinación no homóloga en regiones no sujetas a recombinación homóloga, como son las islas genómicas, o por la incorporación de plásmidos autorreplicativos. Algunos de estos MGE interaccionan entre sí e incluso albergan mecanismos barrera que dificultan la transferencia de DNA (apartado 1.2.4), generando una compleja red que controla las tasas de entrada neta y recombinación de DNA en la célula.

Los elementos conjugativos integrativos contienen gran cantidad de elementos móviles que contienen la información génica necesaria para su integración en el genoma y para su transferencia entre células por conjugación. Junto a los plásmidos conjugativos, constituyen los principales elementos móviles transferidos mediante este tipo de mecanismo (Burrus *et al*., 2002). Su escisión e integración tiene lugar mediante recombinación específica de sitio. El incremento de secuencias de genomas completos en las bases de datos muestra como este tipo de elementos, que incluyen transposones conjugativos e islas genómicas conjugativas análogas a transposones, se encuentran ampliamente distribuidas en el dominio *Bacteria*. Junto con los plásmidos y fagos contribuyen notablemente a la HGT.

Los plásmidos son elementos móviles ampliamente distribuidos en diferentes taxones del dominio *Bacteria* y *Archaea*. En muchos casos forman parte del genoma accesorio de la especie, detectándose una elevada microdiversidad plasmídica en diversos taxones. La principal barrera para su perpetuación es la autorreplicación dado que en ocasiones las proteínas encargadas de reconocer los orígenes de replicación no están codificadas en genes del genoma receptor o no interaccionan con el mismo de manera productiva, limitando el rango de hospedador (Caspi *et*

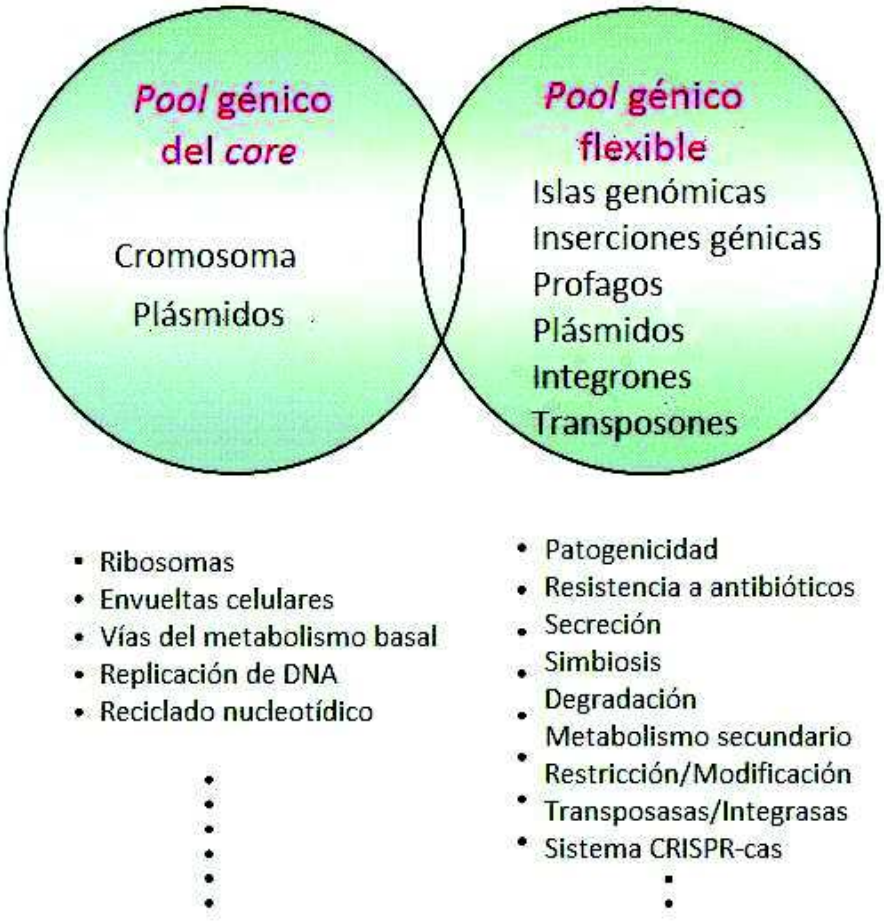


Figura 10I. Modelo de distribución del pool génico de una especie procariota. El diagrama representa las fracción del pool génico del *core* y flexible, detallando en su parte inferior las principales categorías de genes que incluiría cada una de ellas. (adaptación de Hacker y Carniel., 2001).

al., 2001). La integración mediante recombinación aditiva es un fenómeno que permite estabilizar los plásmidos o parte de los mismos en replicones preexistentes. Entre el contenido génico de los plásmidos es posible encontrar elementos como sistemas de restricción modificación (MR), genes de resistencia o sistemas CRISPRs-Cas (Mojica *et al*., 2005, Fricke *et al*., 2011), los cuales afectan a las tasas de recombinación homóloga y a los procesos de estabilización de DNA foráneo mediante HGT como se comentará más adelante.

Las islas genómicas forman parte del *pool* genético flexible, constituyendo regiones genómicas de entre 10-100kb de longitud que difieren en su contenido de G+C respecto al del resto del genoma. Contienen habitualmente secuencias y genes transferidos horizontalmente derivados de fagos o plásmidos, incluyendo transposones e integrasas (**figura I11**), por lo que constituyen regiones de elevado intercambio génico que facilitan el acceso al *pool* génico flexible. Respecto a las pequeñas inserciones genómicas, típicamente menores de 10 kb, las islas

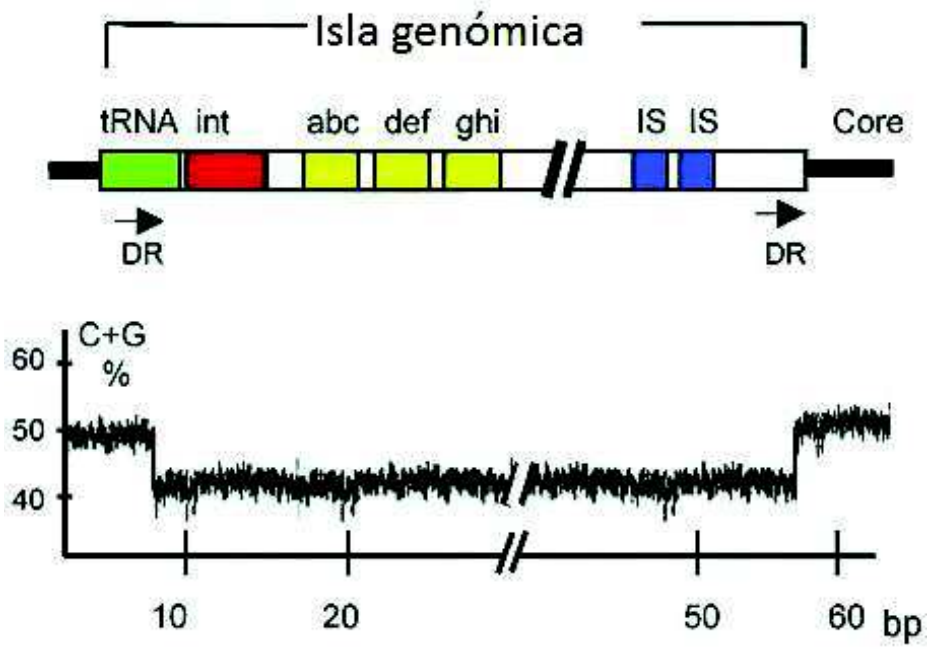


Figura 11I. Modelo esquemático de la estructura de una isla genómica procariota. La figura muestra en detalle algunos elementos génicos que pueden encontrarse en las mismas y que contribuyen a su detección: tRNAs, repeticiones directas flanqueantes (DR), genes con diversas funciones (abc, def, ghi) (transportadores, envolturas y pared celular, patogenicidad), integrasas de fago (int), elementos transponibles (IS). El contenido en G+C de estas regiones es diferente al promedio del genoma *core* (adaptación de Hacker y Carniel, 2001).

genómicas ofrecen la ventaja adaptativa de poder albergar operones completos transferidos mediante procesos de recombinación únicos permitiendo en ocasiones una rápida adaptación frente a cambios ambientales en periodos breves (Hacker y Carniel., 2001). De entre todas ellas, las islas de patogenicidad son las mejor estudiadas por sus implicaciones clínicas, aunque posteriormente se han caracterizado islas genómicas en organismos muy diversos ecológicamente, destacando su papel en procesos de evolución adaptativa mediante HGT de organismos comensales, simbióticos y de vida libre (Dobrind *et al*., 2004; Fernández-Gómez *et al*., 2012). Los genes albergados en estas regiones son muy variables, hallándose islas genómicas enriquecidas en elementos transponibles y sistemas de defensa como factores de virulencia o sistemas CRISPR-cas, confiriendo inmunidad frente a fagos y la entrada de elementos plasmídicos (Fernández-Gómez *et al*., 2012; Jo Sui., *et al*., 2009).

1.2.4- Mecanismos barrera y factores que afectan a la HGT en procariotas.

Existen diversos mecanismos que impiden que una molécula o secuencia de DNA se estabilice en una célula procariota y se replique. Estos se suceden a lo largo de las diferentes etapas implicadas en el proceso de transferencia horizontal desde su entrada a la célula, fase en la cual las envueltas celulares constituyen la principal barrera salvada por mecanismos de entrada como la transfección, transformación y conjugación (apartado 1.2.1), hasta su integración en un replicón autónomo. La estabilidad del DNA en el ambiente extracelular así como la diversidad intra e interespecífica son factores externos a la célula que afectarán notablemente a la tasa incorporación de DNA (Thomas y Nielsen., 2005). Una vez dentro de la célula, la estabilización de las moléculas y su replicación se verá afectada por mecanismos y maquinaria específica: sistemas de recombinación y reparación, similitud taxonómica entre el microorganismo donador y receptor, y mecanismos barrera tales como los sistemas de restricción modificación (MR) y CRISPR-Cas. La presión selectiva del ambiente y la expresión de los genes incorporados determinará su incorporación definitiva o no en un linaje de la especie. Paralelamente, las redes y mecanismos de transferencia horizontal suelen evolucionar constantemente dentro de las poblaciones e individuos como resultado de los múltiples procesos capaces de aliviar o modificar

las distintas barreras, cambios continuos en la presión selectiva ambiental o la enorme plasticidad de los elementos móviles.

Los sistemas MR constituyen uno de los principales mecanismos barrera que afectan la permanencia de DNA foráneo dentro de la célula. Estos sistemas son ubicuos dentro de los procariotas, pudiendo encontrar miles de sistemas MR distintos con un amplio abanico de especificidad (Bayliss *et al*., 2006). Todos ellos se agrupan en 3 tipos principales, y la mayoría comparten la presencia de una metiltransferasa de DNA y una endonucleasa. En los sistemas MR de tipo I existe una única subunidad enzimática con dos dominios de unión a DNA que determina la especificidad de unión a la secuencia. En el caso de los sistemas de tipo II, la unión específica está marcada por la presencia de un domino de unión a diana, presente en ambas enzimas habilitándolas de manera independiente para el reconocimiento y unión a sitios específicos. Los sistemas de MR tipo III son los menos caracterizados, aunque las regiones implicadas en el reconocimiento de secuencia presentan presión selectiva positiva y están sometidas a procesos de HGT entre géneros (Bayliss *et al*., 2006). Los sistemas MR reducen notablemente la frecuencia de HGT entre especies e incluso en ocasiones dentro de una misma especie entre cepas distintas (Tock and Dryden., 2005, Hoskisson y Smith, 2007). Dado que el sistema actúa sobre DNA de doble cadena con un patrón de metilación distinto al de la célula receptora, algunos mecanismos de HGT y entrada de DNA como la conjugación o la transducción, en ocasiones, no son susceptibles al efecto de este sistema hasta la síntesis intracelular de la hebra complementaria (Thomas y Nielsen., 2005). El DNA incorporado por transformación por el contrario será susceptible de restricción desde su entrada. En especies como *N. meningitidis* se ha demostrado que la presencia de diferentes sistemas de MR en distintas cepas y la transformación afectan notablemente a las estructuras poblacionales, asociando diferentes clados con sistemas MR distintos y correlacionando las secuencias transferidas en diferentes regiones genómicas con las tasas de recombinación (Budroni *et al*., 2011). La capacidad de los sistemas MR a la hora de restringir la transformación intra e interespecifica se ha demostrado en especies como *Staphylococcus aureus*, cuyas cepas contienen diferentes tipos de sistemas (Corvaglia *et al*., 2010). Pese a la fuerte presión selectiva a favor del mantenimiento de estos sistemas, en algunas especies como *S.aureus* es posible

encontrar cepas con una elevada tasa de transformación y ausencia de sistemas MR y mecanismos que favorecen dicha variabilidad (Corvaglia *et al*., 2010). En ocasiones los sistemas MR en otras cepas de *S.aureus* se encuentran flanqueados por secuencias repetitivas que favorecen su escisión o en codificados en plásmidos. Por su parte los elementos móviles como fagos o plásmidos muestran una clara coevolución modificando los sitios diana donde actúan estos sistemas.

Otro mecanismo que afecta al proceso de estabilización del DNA tras su entrada en la célula es el sistema CRISPRs-Cas (**figura I12**). Se trata de un sofisticado sistema de defensa contra virus, plásmidos y DNA foráneo formado por secuencias repetitivas cortas regularmente espaciadas denominadas repeticiones entre las que se intercalan otras exógenas denominadas espaciadores (Mojica *et al*., 2005, Samson *et al*., 2013; Koonin y Wolf 2015). Entre los principales mecanismos empleados por los virus para la evasión de los sistemas CRISPR-Cas se encuentran las elevadas tasas de mutación en secuencias protoespaciadoras y de mosaicismo génico (Tyson y Banfield, 2008). En ocasiones los sistemas CRISPR-Cas aparecen en elementos móviles como plásmidos pudiendo afectar a la estructura poblacional fruto del importante papel de los fagos y plásmidos en la especialización bacteriana y la microevolución de sublinajes (Fricke *et al*., 2011).

Tal y como se mencionó con anterioridad al describir los mecanismos de recombinación homóloga, la distancia filogenética entre el organismo receptor y donador es la principal barrera en los procesos de transferencia horizontal. Influye no sólo en la eficiencia de mecanismos de transferencia y barrera mencionados anteriormente sino en los propios procesos de recombinación homóloga que estabilizan los fragmentos incorporados. La recombinación homóloga es el mecanismo principal de estabilización de DNA dentro de un replicón autónomo una vez superados mecanismos de evasión como los sistemas MR y CRISPR. Dependiendo del sistema, para que la recombinación homóloga tenga lugar de manera eficiente, el fragmento a integrar ha de contener extremos homólogos de entre 25-200 pb (Thomas y Nielsen., 2005), demostrándose en diversos grupo taxonómicos que la eficiencia de los sistemas de recombinación homóloga decae notablemente con la divergencia entre donador y receptor (Meier y Wackernagel., 2005), limitándose claramente a fragmentos con una divergencia menor del 25%

(Thomas y Nielsen., 2005). De iniciarse la recombinación entre secuencias con un elevado grado de disimilitud, sistemas de control como el de reparación de errores entre heteroduplex dirigido por metilación (MMRS, del inglés *Mismatch Repair System*) detectarían los *mismatches* o bases desapareadas abortando el proceso de recombinación homóloga (González-Candelas y Francino., 2011).

Otro factor que afecta al proceso de recombinación homóloga es la acción del sistema SOS, inducido en condiciones de estrés que generen daño en el DNA o interfieran con los procesos de replicación. Se cree que el DNA de cadena sencilla (ssDNA) es la señal inductora,

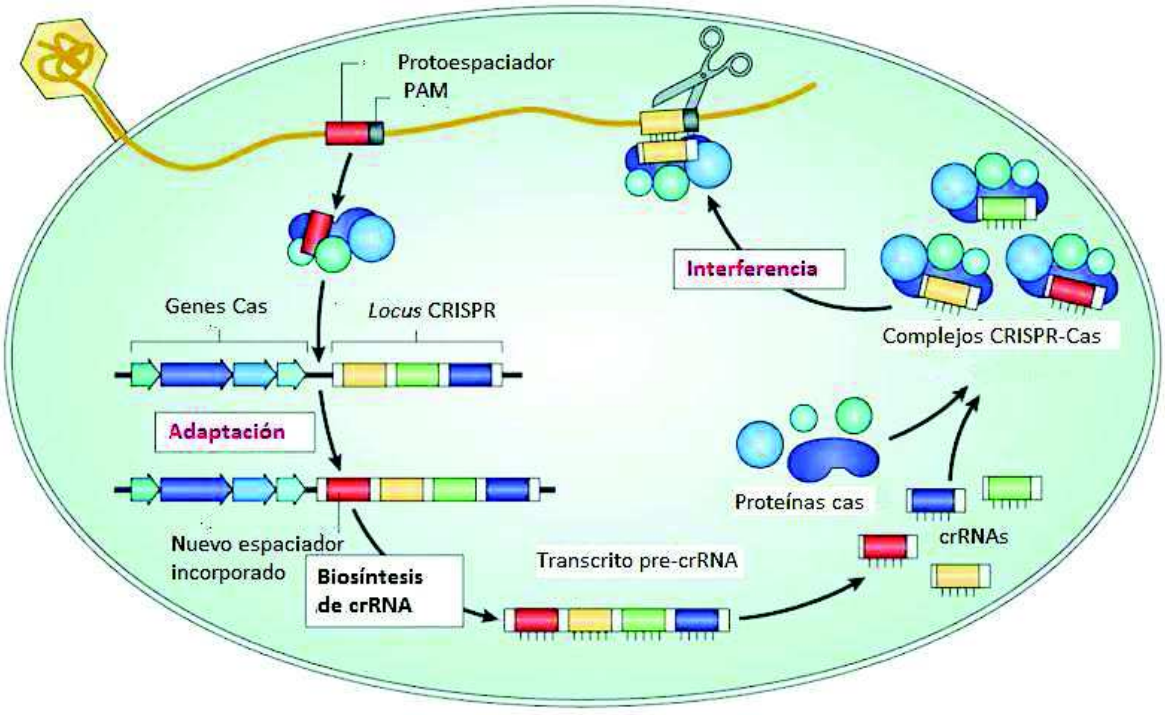


Figura I12. Modelo esquemático de la estructura típica de un sistema CRISPR-Cas en procariotas y el proceso de adaptación e interferencia tras la entrada de DNA foráneo por infección viral. La figura muestra en detalle algunos elementos génicos que pueden encontrarse en el CRISPR *array* como los genes *cas* y las repeticiones con espaciadores intercalados. Tras la primera infección se produce el proceso de adaptación, por el cual se incorpora un nuevo espaciador. Ante una nueva infección viral, la síntesis del crRNA yla formación del complejo CRISPR-cas mediará el proceso de interferencia frente al DNA del virus (adaptada de Samson *et al*., 2013).

por lo que este sistema se activa durante los procesos de conjugación, transformación, transposición, restricción de DNA foráneo y frente a la presencia de DNA plasmídico o viral en forma de cadena sencilla. Durante la transformación y la conjugación estimula la sobreproducción de enzimas de recombinación. La inducción del sistema SOS es mayor en transferencias interespecíficas, dado que la acción del sistema MMRS impide la recombinación de este ssDNA, que queda expuesto en el citoplasma. Sin embargo prevalece el efecto del sistema MMRS sobre el SOS en proceso de recombinación, por lo que la frecuencia de intercambio por recombinación homóloga disminuye notablemente con la distancia filogenética (Meier y Wackernagel., 2005).

Por último, tras la integración, la perpetuación de un gen adquirido depende en gran parte de su expresión efectiva y de que aporte un beneficio que compense los costes derivados de su replicación y expresión. A nivel transcripcional, la expresión de genes incorporados recientemente puede verse comprometida por que la maquinaria de la célula receptora, RNA polimerasa o factores de transcripción, reconozcan adecuadamente las secuencias promotoras reguladoras. Además, muchas de las secuencias heterólogas transferidas horizontalmente son ricas en AT. El sistema proteico de estructura nucleoide en bacterias, análogo al de histonas, se une preferentemente a este tipo de secuencias reprimiendo su expresión mediante un proceso conocido como silenciamiento xenogénico (Navarre *et al*., 2007). A nivel traduccional, diferencias en la frecuencia de uso de codones entre la bacteria donadora y receptora puede afectar a la expresión génica, especialmente en aquellos genes con un elevado nivel de expresión en la célula de origen ya que emplean los tRNAs más abundantes, y estos varían en las diferentes especies (Tuller., 2011).

1.2.5- Análisis de recombinación *in silico* con genomas completos.

En los últimos 15 años se han publicado numerosos estudios de recombinación homóloga en organismos procariotas, ya sea entre cepas de una misma especie o entre especies filogenéticamente cercanas (Didelot y Maiden 2010, Vos y Didelot 2009, Caro-Quintero et al., 2009). La mayoría de ellos, por su impacto clínico, se centran en organismos patógenos. Este

tipo de análisis se emplea frecuentemente en investigaciones epidemiológica a diferentes escalas y en estudios de biología, evolución y patogenicidad bacteriana. Entre las diferentes aproximaciones llevadas a cabo para estudiar recombinación génica en el dominio *Bacteria*, una de las más extendidas a lo largo de este periodo ha sido el MLSA (Maiden *et al*; 2006 Glaeser y Kampfer., 2015).

Aunque los estudios con MLSA han proporcionado una información valiosa en diversas especies *de Bacteria* y *Archaea,* en ocasiones empleando numerosos aislados (Perez-Losada *et al;* 2006) (Vos y Didelot., 2009*),* esta metodología utiliza normalmente unas pocas secuencias o fragmentos genómicos por individuo y un número de cepas reducidos. Normalmente se emplean entre 6 y 10 secuencias de genes *housekeeping* cada una de alrededor de 450 pb de longitud distribuidos a lo largo del genoma de tal manera que sea improbable que un evento de recombinación incluya a dos de ellas (Didelot y Maiden 2010). El tipado puede llevarse a cabo de varios modos, y durante el mismo se realizan test de estructuras poblacionales clonales. En ausencia de cualquier intercambio genético, toda la variabilidad adquirida sería mediante mutaciones puntuales pudiendo caracterizarse por: 1) Desequilibrios de unión; 2) Árboles filogenéticos basados en cada *housekeeping* y 3) Congruencia (capacidad de mostrar la misma filogenia para cada *loci* empleado) (Didelot y Maiden, 2010).

El empleo de un número bajo de *loci* conlleva importantes restricciones en cuanto a la cobertura genómica analizada ya que no permiten analizar el efecto de la recombinación homóloga a nivel de genoma completo, y por tanto acceder a la información inferida sobre la naturaleza de los eventos de recombinación observados y acerca del impacto global de la recombinación homóloga en dichos genomas y sus repercusiones en la evolución de la especie (Didelot *et al;* 2010). Sin embargo, el constante aumento del número de genomas secuenciados completamente ha permitido realizar cada vez más estudios comparativos, habitualmente empleando menos de 15 genomas, que revelan mucha más información acerca de los procesos de recombinación, distribución de los fragmentos recombinados y sus propiedades, patrones de flujo génico y genealogía clonal en comparación a los MLSA (Martin *et al*., 2010).

Según el tipo de datos disponibles para un estudio, genomas completos o genes *housekeeping*, existen numerosas y distintas herramientas bioinformáticas que permiten

aproximarse al estudio del impacto de la recombinación, analizando las tasas de recombinación a las que se ve sometida una especie o los eventos de recombinación acontecidos en ella (**figura I14**). La resolución de estos programas y su precisión a la hora de detectar señales de recombinación depende del tamaño y número de secuencias disponibles y de la representatividad de las mismas respecto del total del genoma. En general, desde un punto de vista práctico, para la detección de un evento de recombinación en particular, la secuencia analizada debe estar presente en al menos una cepa que cumpla estas dos condiciones: contener la región a analizar sin que esté involucrada en el proceso de recombinación y posicionarse filogenéticamente cercana a una de las dos cepas parentales recombinantes entre las que tiene lugar la recombinación (Martin *et al.,* 2011).

A la hora de determinar las frecuencias de recombinación en procariotas los estudios más comunes hasta la fecha emplean MLSA (Vos y Didelot., 2009) y dos parámetros o tasas: la tasa de recombinación mutación (r/m) y la relación rho/theta (σ/θ). La tasa de recombinación respecto a la de mutación puntal (r/m), que muestra de manera directa la contribución de la recombinación respecto a la mutación sobre la diversificación de la especie. La relación rho/theta (σ/θ) indica el impacto relativo de la recombinación y la mutación sobre la filogenia de la especie y su evolución. Las tasas relativas de mutación (r/m) se estiman típicamente mediante algoritmos que resumen los cambios apreciados en las secuencias nucleotídicas, entre los que tenemos el programa LDHat (McVean *et al*., 2002), o con algoritmos de reconstrucción, como el implementado por programas como ClonalFrame (Didelot *et al;* 2010). A diferencia de los primeros, estos algoritmos reconstruyen la genealogía clonal completa no perdiendo información contenida en el conjunto de secuencias.

La mayoría de programas que detectan eventos de recombinación individuales siguen dos estrategias conocidas como esquemas particionales y esquemas test (Martin *et al.,* 2011):

i) Los *esquemas particionales* detectan eventos de recombinación individuales tras dividir las secuencias alineadas en 2 o más fragmentos. Dentro de los esquemas particionales a su vez tenemos 3 categorías de programas: los más simples o estáticos, como por ejemplo LARD (Holmes *et al*.,1999), SMARTIE (Blooquist y Suchard., 2010) y los complejos o dinámicos, entre ellos los más usados son MAXCHI (Mainard Smith., 1992) y CHIMAERA (Posada y

Crandall., 2001). Los segundos utilizan ventanas móviles y son mucho más efectivos. El tercer grupo dentro de los esquemas particionales no emplea ventanas sino algoritmos mucho más sofisticados como los implementados en los programas GARD (Kosakovsky P. *et al*., 2006), rec HMM (Westesson y Holmes., 2009) o jpHMM (Schultz *et al*., 2006).

ii) La segunda estrategia, *esquemas tipo test*, es la opción que incorporan muchísimos métodos de detección de recombinación y consta de dos etapas secuenciales. En la primera fase se detectan cambios en las relaciones entre secuencias y en la segunda se evalúan estadísticamente la significancia de tales diferencias. Dentro de estos métodos que usan esquemas tipo test, durante la primera fase unos analizan la similitud de secuencias alineadas y otros tienen en cuenta el soporte filogenético entre ellas. Entre estos últimos, mucho más precisos que los basados en similitud, encontramos los programas SISCAN (Gibbs *et al*., 2000) y BOOTSCAN (Salmien *et al*., 1995). Una vez detectados los eventos de recombinación, existen numerosos métodos que evalúan la probabilidad de hallar estas señales de recombinación en ausencia de la misma, entre los que encontramos GENECONV (Padidam 1999), RDP (Martin *et al*., 2010) y MAXCHI.

Una vez se ha decidido el grupo de programas a utilizar en función del tipo de información que se pretende obtener, se debe elegir la herramienta que mejor se adapte a los datos de partida teniendo en cuenta el número de secuencias y longitud disponibles (**figura I13**). En los últimos años, en consonancia con el incremento de datos producidos mediante técnicas de secuenciación de nueva generación (NGS, del inglés *next-generation sequencing*), se han desarrollado programas que permiten aproximarse a la detección de eventos de recombinación y tasas de recombinación con genomas completos (Martin *et al.,* 2011). Entre ellos destacan dos programas que trabajan específicamente con este tipo de datos: *Recombination Detection Program* (RDP4.15v) (Martin *et al*; 2010) y ClonalFrame (Didelot *et al;* 2010). La principal cualidad del programa RDP4 es que habilita y aplica simultáneamente un amplio rango de métodos para detectar y caracterizar los eventos de recombinación dentro de un alineamiento sin la necesidad de incluir un subgrupo de secuencias no recombinantes. Se ha empleado con éxito en diversas especies de microorganismos con diferentes tamaños muestrales, tales como los estudios llevados a cabo en *E.coli* (Mau *et al*., 2006). Muchos programas y métodos de detección

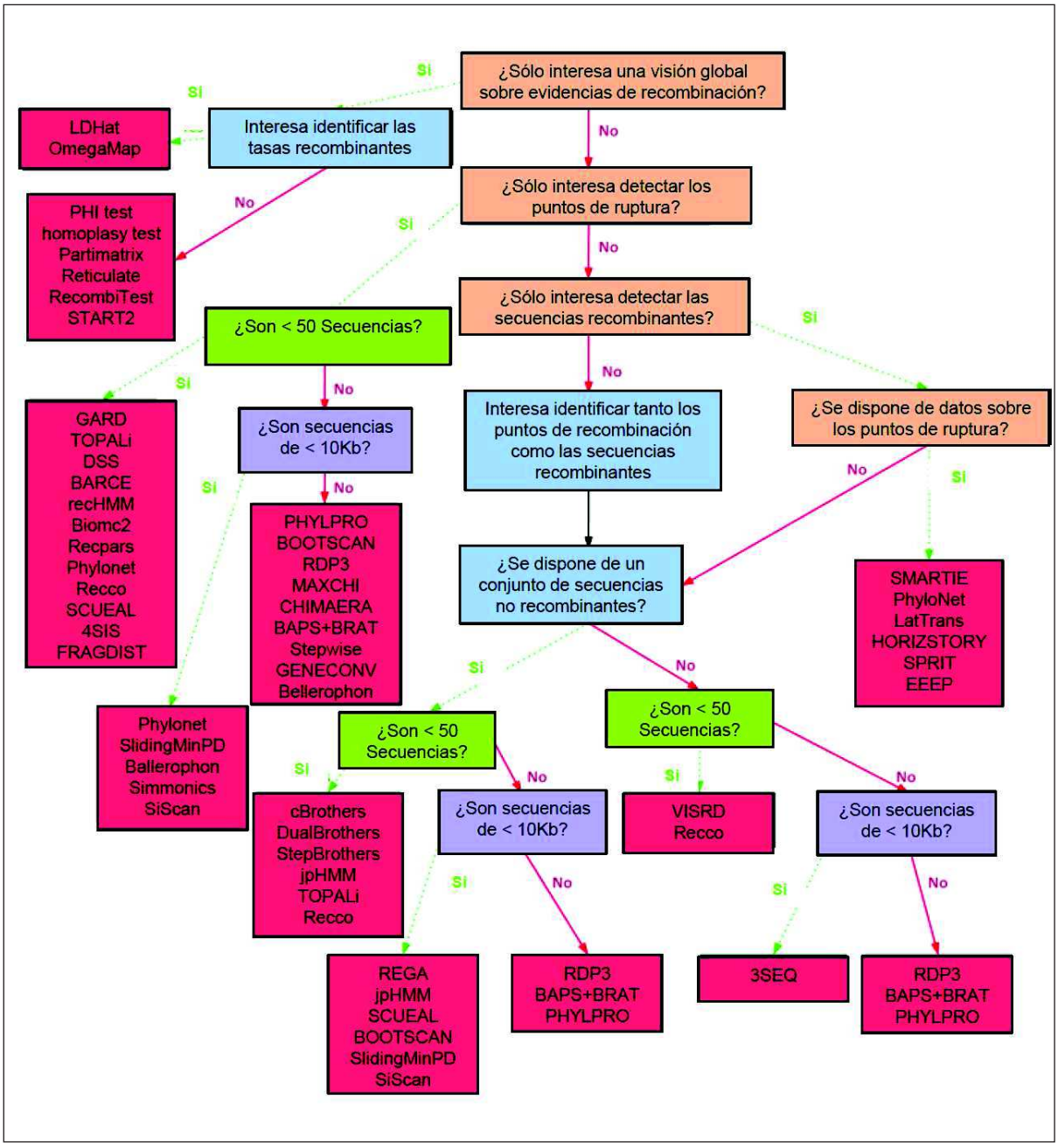


Figura 13I. Principales metodologías y programas para el análisis de recombinación según la naturaleza de las secuencias analizadas (adaptada de Martin *et al*., 2010). En rojo se muestran los diferentes programas una vez seleccionados el tipo de información que se pretende obtener (cuadros azules) y el tipo de secuencias que se emplearán, según su número (cuadros verdes) y tamaño (cuadros morados).

de eventos recombinantes sólo son capaces de detectarlos cuando los descendientes de la cepa donadora y receptor están presentes en la muestra. Una excepción es el programa ClonalFrame (Didelot y Falish 2007), muy útil a la hora de reconstruir la genealogía de la especie e identificar los patrones de recombinación. Normalmente, tal y como muestran publicaciones previas con *Chlamydia trachomatis* (Sandeep *et al*; 2011), *Listeria monocytogenes* (Orsi *et al*; 2008) y *Francisella thularensis* (Larsson *et al*; 2009), se obtienen muy buenos resultados empleando ClonalFrame con un número de cepas limitado (entre 4 y 20 de la misma especie).

1.3- Evolución en las estrategias de secuenciación y ensamblaje de genomas.

En los últimos años las mejoras experimentadas en las tecnologías en secuenciación han abaratado notablemente los costes y reducido significativamente los tiempos de espera respecto a la secuenciación tradicional de Sanger (Metzker 2010). El auge de las denominadas tecnologías de secuenciación de nueva generación (NGS), ha permitido su utilización por parte de muchos grupos de investigación, revolucionando el campo de la genómica (Schuster 2008; Paszkiewicz y Studholme 2010). Muchas de las cuestiones o problemas en el estudio de organismos procariotas pueden abarcarse de manera más informativa y directa mediante la secuenciación de genomas completos. Existen numerosos trabajos desarrollados en este sentido con el objetivo de entender procesos y capacidades de patogénesis y de adaptación a diferentes ambientes, interacción patógeno-hospedador o profundizar en los mecanismos de evolución de las mismas (Hackel y Carniel., 2001., Scott y Ely., 2014). En la última década se ha incrementado de manera notable el número de genomas procariotas secuenciados en las bases de datos (Jackman *et al*., 2010; Pagani *et al*., 2012), y se espera que continúe esta progresión ya que hasta el momento solo se dispone de los genomas completos de una pequeña fracción del total de especies procariotas (**figura I14**).

Las nuevas tecnologías de secuenciación proporcionan un número mayor de datos de modo más directo, más preciso y rápido (**tabla I4**) (Quail *et al*., 2012; Liu *et al*., 2012). Existen diferentes tecnologías disponibles en el mercado por lo que, a la hora de escoger entre una u otra, han de considerarse aspectos como el tamaño de *reads*, la precisión en las mismas, el tiempo

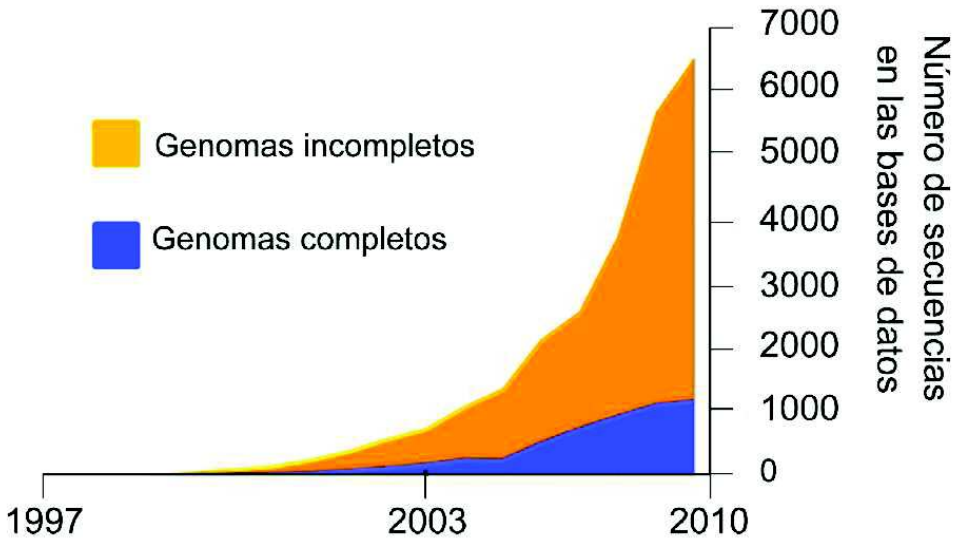


Figura I14. Progresión en el número de genomas secuenciados completamente o en proceso de secuenciación depositados en las bases de datos (Liolios *et al*., 2010).

Tabla I4. Características y comparativa de las principales plataformas de secuenciación en el mercado.

Tecnología de secuenciación	Longitud de lecturas (pb)	Precisión	Reads por Carrera (máximos)	Tiempo por carrera	Costes por millón de bases
Pacific Biosciences RsII	10.000-20.000	99.9999%	50.000 (500-1000Mb)	30 min-4 h	0,13-0,60 $
Ion Torrent	>400pb	98%	>80 millones	2h	1$
Pirosecuenciación Roche 454	700	99.9%	1 millones	24h	10$
Illumina Hiseq/Miseq	50 a 300 pb x2	98%	>3 billones (600Gb)	8 h-10 días	0,05-0,15 $
Secuenciación por ligación (SOLiD)	50+35pb o 50+50pb	99.9%	1,2-1,4 billones	1-2 semanas	0,13$
Secuenciación Sanger	400-900	99.9%	N/A	20 min-3h	2400 $

necesario y el coste de cada carrera. Entre ellas encontramos Roche 454, Pacific Biosciences RII (PacBio) e IlluminaMiseq **(tabla I4**). Con la entrada de las nuevas estrategias de secuenciación se ha tendido a proporcionar un mayor número de secuencias de longitudes menores, alcanzando coberturas de secuenciación y pares de bases secuenciados por carrera enormes.

Los cambios acontecidos en las tecnologías de secuenciación y en las características de las secuencias obtenidas, en términos de cobertura y longitud, ha motivado la necesidad de desarrollar algoritmos nuevos para el ensamblaje de las mismas. En los últimos años han surgido numerosos algoritmos capaces de trabajar cada vez con un mayor número de secuencias de tamaño más corto o de combinar el ensamblaje procedente de diferentes tecnologías (**Tabla I5**).

Además de adaptarse a los cambios de cobertura y longitud de secuencia, los nuevos ensambladores han de ser capaces de resolver situaciones conflictivas, algunas de ellas derivadas del empleo de secuencias de corta longitud. Entre los principales problemas apreciados en estudios comparativos de metodología de ensamblaje *de novo* destacan 4: el trabajo con secuencias de elevado contenido en GC o heterogeneidad en el mismo, la presencia de locus o regiones idénticas a lo largo del genoma, la cobertura heterogénea a lo largo del genoma a ensamblar y la presencia de regiones largas repetitivas (Scott y Ely., 2014).

En la actualidad los ensambladores de nueva generación son capaces de procesar archivos de secuenciación de más de 20 Gb con un menor gasto de tiempo y recursos bioinformáticos, automatizando enormemente los procesos de **pruebas *multi-kmer***. Ensambladores empleados de manera habitual hace tan solo 5-10 años, como SoapdeNOVO o Velvet (tabla 5I), han sido superados paulatinamente por nuevos algoritmos que resultan mucho más eficientes en el consumo de recursos además de proporcionar mejores resultados tal como muestran estudios comparativos y revisiones continuas (Narcisi *et al*., 2011; Quail *et al*., 2012; Schatz *et al*., 2011). Varias de estas publicaciones muestran que ensambladores como IDBA o SPAdes como los más eficientes del mercado (Peng, *et al*., 2010; Magoc *et al*., 2013; Gurevich *et al*., 2013) al solucionar los problemas derivados de la heterogeneidad de cobertura.

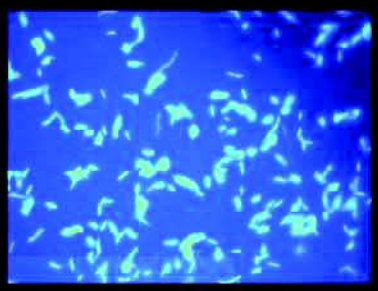
La comparación de ensamblajes requiere del establecimiento de unos criterios generales y objetivos. Para ello se consideran parámetros estándar que evalúan la métrica de los productos generados, entre ellos el número de *contigs* totales, el número de *contigs* de más de 1Kb, el

Tabla I5. Principales ensambladores empleados durante la última década en el ensamblaje de genomas procariotas y tipo de secuencias capaces de ensamblar.

Ensamblador	Tipo de ensamblado	Tecnología	Referencia/Ultima actualización
ABySS	Genomas grandes	Solexa, SOLiD	(Simpson *et al*., 2009) /2011
AMOS	Genomas	Sanger, 454	(Salzberg *et al*. 2002) /2008
Celera WGA Assembler / CABOG	Genomas grandes	Sanger, 454, Solexa	(Myers *et al*.; 2004) /2010
CLC Genomics Workbench & CLC Assembly Cell	Genomas	Sanger, 454, Solexa, SOLiD	CLC bio (2008)/2011
Cortex	Genomas	Solexa, SOLiD	(Iqbal *et al*.,.2012)
Euler	Genomas	Sanger, 454, Solexa	(Pevzner *et al*.,2001) /2006
Euler-sr	Genomas	454, Solexa	(Chaisson *et al*., 2009)
Geneiouis	Genomas	Sanger, 454, Solexa, Ion Torrent, Complete Genomics, PacBio, Oxford Nanopore, Illumina	Biomatters Ltd (2009) /2013
IDBA (Iterative De Bruijn graph short read Assembler)	Genomas grandes	Sanger,454,Solexa	(Peng, *et al*., 2010)
LIGR Assembler (derived from TIGR Assembler)	Genómica	Sanger	2009/2012
MaSuRCA (Maryland Super Read - Celera Assembler)	Genomas grandes	Sanger, Illumina, 454	(Zimin *et al*., 2013)/2013
MIRA (Mimicking Intelligent Read Assembly)	Genomas, ESTs	Sanger, 454, Solexa	Chevreux (2004) /2011
Newbler	Genomas, ESTs	454, Sanger	454/Roche (2009)
JRAssembler	Genomas	Solexa	(Chu *et al*., 2013)
PANDAseq	Genomas grandes	Solexa	(Maesella *et al*., 2012)
SGA	Genomas grandes	Illumina, Sanger (Roche 454?, Ion Torrent?)	(Simpson *et al*. 2011) /2012
SOAPdenovo	Genomas	Solexa	(Li *et al*., 2009)
SPAdes	Genomas reducidos, SCG	Illumina, Solexa	(Bankevich *et al*., 2012) /2013
Velvet	Genomas reducidos	Sanger, 454, Solexa, SOLiD	(Zerbino y Birney 2008) /2009

tamaño total de la secuencia ensamblada, el número de regiones indeterminadas dentro de los ensamblados (Ns), y el valor de **N50*** (tras ordenar los *contigs* de mayor a menor, tamaño de aquel que completa el 50% del ensamblado). Actualmente, dado el abanico de tecnologías disponibles, los estudios genómicos tratan de optimizar combinaciones adecuadas de estas y determinar el esfuerzo de secuenciación necesario para generar genomas procariotas completos salvando los problemas derivados de las regiones con repeticiones en tándem o duplicaciones génicas.repetitivas. Ya que los mayores costes de secuenciación derivan de la generación de librerías principalmente, el reto actual reside en generar ensamblajes completos económicos empleando secuencias cortas de gran cobertura. Los últimos ensambladores producen ensamblajes *de novo* de muy buena calidad empleando librerías sencillas, pero todavía incompletos. El reto en el futuro reside en la generación de algoritmos que resuelvan estas regiones repetitivas conflictivas que actualmente sólo pueden ensamblarse con la ayuda de librerías de secuencias de mayor longitud (Scott y Ely., 2014).

* Se destacan en negrita a lo largo de la tesis algunos términos definidos en el glosario (véase anexos).



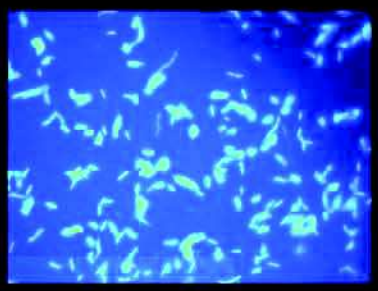
Este trabajo presenta el análisis transcriptómico y genómico comparativo de la bacteria halófila extrema *Salinibacter ruber*. El principal objetivo planteado en esta tesis es profundizar en la descripción de la microdiversidad genómica y funcional de esta especie y en los mecanismos evolutivos que la dirigen, explorando en este último caso el alcance de las estrategias evolutivas de los genomas *core* en especies de la misma comunidad y finalmente en especies procariotas con diversas estrategias de vida.

La consecución de estos objetivos generales se llevó a cabo a lo largo de los tres capítulos en que se estructura esta tesis, abordando en cada uno de ellos los siguientes objetivos específicos:

El primer capítulo explora la microdiversidad funcional existente mediante un análisis transcriptómico comparativo de las cepas M8 y M31 mediante RNAseq, observando los efectos derivados de diferencias genómicas sutiles. En segundo lugar, y objetivo principal del capítulo, se describen en las diferencias de expresión derivadas de la interacción de cepas cercanas en cultivo mixto para elucidar si ambas se comportan como la adición individual de cada una de ellas o si modifican sus actividades, planteando las implicaciones microevolutivas tiene este tipo de interacción.

A lo largo del capítulo 2 se lleva a cabo un análisis genómico extenso de la especie *S. ruber*. Como primer objetivo se describe la diversidad genómica y patrones de arquitectura de 8 aislados caracterizando los genomas *core* y accesorio de la especie. En segundo lugar se analizan los mecanismos evolutivos que actúan sobre estas dos fracciones del genoma y sus implicaciones evolutivas.

Por último en el tercer capítulo se evalúa el impacto de la recombinación homóloga en 54 especies procariotas con tres objetivos: el primero evaluar si, como sucede en *S. ruber,* este mecanismo resulta determinante en la evolución de los genomas *core* de alguna otra especie; el segundo explorar que factores determinan grado de incidencia de la recombinación homóloga sobre los genomas *core* y por último elucidar que efecto ejerce la recombinación homóloga sobre la microdiversidad y *clusters* poblacionales de cada una de estas especies y sus mecanismos de adaptación.



1. TÉCNICAS EXPERIMENTALES “WET LAB”.

1.1- Diseño experimental del análisis transcriptómico y del estudio de los mecanismos de microdiversidad de cepas de *S.ruber*.

El análisis transcriptómico de las cepas M8 y M31 de *S. ruber*, correspondiente al capítulo 1 de resultados y discusión, consta de una primera etapa de procesamiento de muestras en el laboratorio y una segunda de análisis bionformático *in silico*. La **figura M1** muestra un esquema general de la metodología del trabajo realizado en el laboratorio. En primer lugar se procedió al diseño *in silico* de cebadores específicos para las cepas M8 y M31 de *Salinibacter ruber*. Estos se probaron empleando DNA extraído de cultivos líquidos puros de las cepas M8 y M31 de *S. ruber*. Antes de la extracción, se estimó el volumen a emplear tras fijar muestras de estos cultivos y determinar su densidad celular mediante recuento por DAPI. Los cebadores diseñados se validaron mediante PCR en gradiente y qPCR, estimando su eficiencia, sensibilidad y Tm óptima seleccionando aquellos que presentaron un mejor comportamiento.

Una vez validados, los cebadores se emplearon para cuantificar de modo absoluto mediante qPCR el número de células en cada uno de los puntos muestreados a lo largo de las curva de crecimiento de los cultivos puros de las cepas M8, M31 y el cultivo mixto. Para cada punto se llevaron a cabo recuentos celulares por DAPI (4',6-diamidino-2fenilindol) y se comprobó la pureza de los cultivos por PCR convencional.

El estudio de la microdiversidad de la especie *S. ruber* y sus mecanismos de diversificación (capítulo 2) involucró una primera etapa de trabajo experimental en el laboratorio en la cual se cultivaron 6 cepas de *S.ruber (***figura M1)**, de distinta procedencia y aisladas en diferentes periodos de tiempo, con el objetivo de secuenciar sus genomas. Posteriormente, ya en la etapa de trabajo *in silico* se procedió a su ensamblaje y anotación. Se llevó a cabo un análisis microevolutivo con estas 6 cepas y las depositadas en el NCBI, M8 y M31, con el fin de discernir los mecanismos que afectan a la microdiversidad y diversificación de la especie.

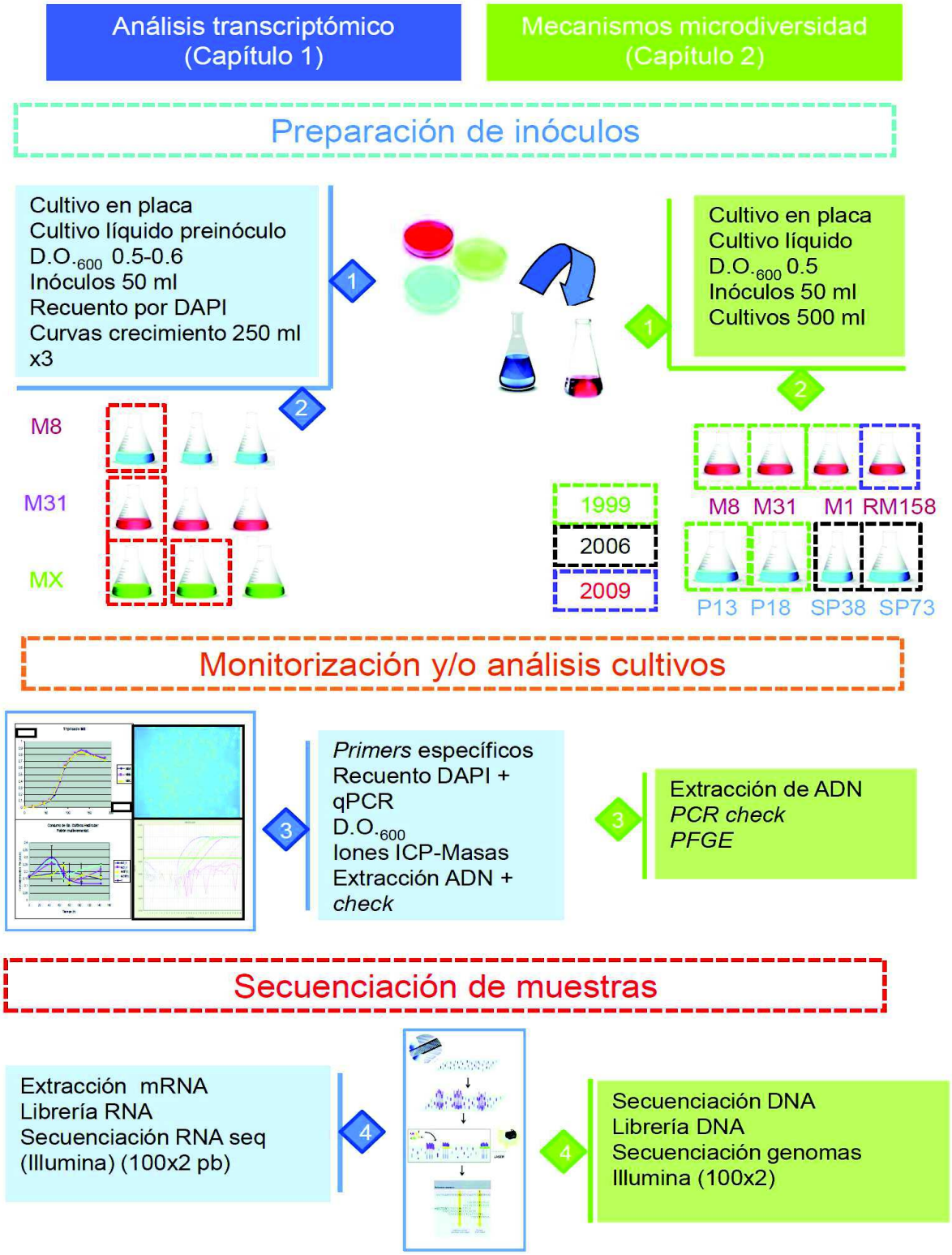


Figura M1. Esquema de trabajo experimental de los capítulos 1 y 2. La numeración refleja la secuencia temporal de los análisis realizados

1.2- Cultivo de *S. ruber*.

Los análisis transcriptómicos se realizaron con cultivos puros y mixtos de las cepas M8 y la cepa tipo M31 (DSM-13855) de *S. ruber*, aisladas simultáneamente de las salinas de Campos de Mallorca en 2002. En el caso del estudio genómico comparativo se emplearon cepas aisladas en diversos años y localizaciones geográficas (véase **tabla M1**).

Tabla M1. Cepas empleadas en este estudio.

Año de aislamiento	Cepa	Localización	Capítulo de resultados
1999	*S. ruber* P13	SP	3
1999	*S. ruber* P18	SP	3
1999	*S. ruber* M8	MLL	1,3
1999	*S. ruber* M31	MLL	1,3
1999	*S. ruber* M1	MLL	3
2006	*S. ruber* SP73	SP	3
2006	*S. ruber* SP38	SP	3
2009	*S. ruber* RM158	MLL	3

SP. Salinas de Bras del Port de Santa Pola. Cristalizador CR30.
MLL. Salinas de Campos de Mallorca.

Para el aislamiento de colonias se procedió a resembrar aislados en nuevas placas de agua de sales (SW) al 25% (Rodríguez-Valera *et al.,* 1985) con 20 g/l de agar-agar y 1 g/l de extracto de levadura. Las placas se incubaron a 37ºC durante al menos 15 días. Una vez crecidas las colonias, se procedió a inocular una de cada cepa, M8 y M31, en tubos de 10 ml que contenían 1 ml de agua de sales SW25% suplementado con 2g/l de extracto de levaduras. Se incubaron a 37ºC con agitación, 170 rpm, durante aproximadamente 10 días. Tras el crecimiento de los preinóculos se procedió a la inoculación al 5% de 20 ml del mismo medio líquido en tubos de 50 ml. Se incubaron aproximadamente durante 5 días a 37ºC, a 170 rpm, hasta alcanzar el punto medio de la fase exponencial ($D.O._{600.}$ = 0,5).

Con la finalidad de obtener suficiente biomasa para la extracción de ácidos nucleicos y secuenciación de los genomas se cultivaron inóculos de cada cepa en matraces de 500 ml conteniendo SW25% suplementado con extracto de levadura (0,2%).

Durante el estudio de la interacción existente entre las cepas M8 y M31 de *S. ruber* en cultivo mixto se cultivaron dos inóculos, uno de cada cepa, en SW25% suplementado con extracto de levadura (0,2%). Cuando la $D.O._{600}$ alcanzó un valor de 0.6 se determinó mediante DAPI la cantidad de células y se inocularon, en matraces de 500 ml conteniendo 250 ml de medio, cultivos puros de M8 y M31 y mixtos de ambas cepas, todos por triplicado. Los cultivos puros se inocularon con 10^9 células y para el mixto se utilizaron $5x10^8$ células de cada cepa.

Se monitorizó el crecimiento de todos los cultivos mediante la medida de $D.O._{600..}$ En los puntos iniciales, con $D.O._{600.}$ menores de 0,3, se tomaron:

- 1 ml para medir D.O.
- 1 ml para tinción con DAPI (apartado 1.3).
- 2 ml para extracción de DNA mediante *boiling* (para comprobar la ausencia de contaminación mediante PCR convencional) (apartado1.4.1).
- 5 ml para la extracción con el *kit* Dneasy Blood and Tissue (QIAGEN) y determinación del número de copias de DNA mediante PCR cuantitativa (qPCR) y re-comprobar la ausencia de contaminación (apartado).

Para valores de $D.O._{600}$ superiores a 0,3 se tomaron:

1. 1 ml para D.O.
2. 100 µl para recuento mediante DAPI (apartado 1.3).
3. 100 µl para *boiling* (apartado1.4.1).
4. 1 ml para qPCR (apartado 1.7).
5. 10 ml para la extracción de RNA para el futuro análisis del transcriptoma.

La toma de muestras se llevó a cabo cada 24 horas hasta alcanzar una $D.O._{600..}$ de 0,3, a partir de la cual se realizó cada 12 horas hasta fase estacionaria.

Las muestras para PCR cuantitativa o extracción de RNA se centrifugaron (3900 xg, 15 minutos), lavando el pellet celular con SW25% para conservarlo a -80ºC.

1.3- Recuento de células.

Se fijaron las muestras tomadas en los sucesivos puntos de las curvas de los cultivos analizados (M8, M31 y mixto) con formaldehído (Sigma Life Science), a una concentración final del 7% v/v, durante 16 horas a 4ºC. Para cultivos cuya $D.O._{600}$ fuese superior a 0,3 se emplearon 100 µl. Para aquellos puntos que presentaron una $D.O._{600}$ menor a 0,3 se fijó 1 ml. Una vez fijadas, las muestras se llevaron a 10 ml con PBS 1X (véase anexo 1) y se almacenaron a 4ºC hasta su filtración y posterior recuento.

Para realizar los recuentos celulares en cada punto de las curvas de crecimiento, se filtró un volumen determinado de muestra fijada a través de filtros GTTP Isopore (Millipore, Billercia, MA, USA) de 0,22 µm de diámetro de poro. El volumen filtrado varió según la densidad celular, siendo menor a mayores concentraciones de células, para obtener un número de células estadísticamente significativo (entre 30-300 células por campo). Los filtros se tiñeron con 4' 6'-diamidino- 2- fenilindol- dihdrocloruto (DAPI) (Sigma Life Sciences) a una concentración de 1 µg/ml durante 5 minutos. Posteriormente se lavaron con agua destilada estéril durante 1 minuto, se pasaron por etanol y se dejaron secar en oscuridad sobre papel de filtro. El recuento de células se llevó a cabo en un microscopio de epifluorescencia Leica DMLA.

1.4- Extracción de ácidos nucleicos.

1.4.1- Extracción de ácidos nucleicos para PCR (*boiling*).

Durante la monitorización de curvas de crecimiento para su posterior análisis transcriptómico, se realizaron extracciones de ácidos nucleicos mediante *boiling* a partir de 2ml de cultivo líquido, en el caso de los puntos con una $D.O._{600}$ menor a 0,3 y 100 µl cuando la $D.O._{600}$ fue superior a 0,3. Tras centrifugar dicho volumen, a 13.000 rpm durante 10 minutos, se descartó el sobrenadante y se resuspendió el *pellet* en 100 µl de agua mQ. Tras someter la muestra a 10 minutos de ebullición y de nuevo a centrifugación (16.000 xg, 10 minutos), se recuperó el sobrenadante, que finalmente se empleó como molde en la PCR.

1.4.2- Extracción de ácidos nucleicos para qPCR

La extracción de DNA genómico para su utilización en qPCR se llevó a cabo mediante el *kit* de extracción Dneasy Blood and Tissue (QIAGEN), eluyendo en un volumen de 100 µl. Una vez realizada la extracción, el RNA se digirió con 1 µl de RNasa (RNase DNase-free Roche). Como la columna empleada en la extracción se satura a partir de $2x10^9$ células, antes ha de determinarse mediante tinción la densidad celular de la muestra empleada.

La calidad del DNA extraído se determinó por medio de una electroforesis en gel de agarosa Seakem LE al 1% (FMC Bioproducts) en tampón TAE 1X, a 5V/cm, empleando 500 ng de DNA por calle. Previamente se obtuvieron los datos de concentración del DNA extraído, así como la calidad de la extracción (relación A260/A230) y la pureza del mismo (relación A260/A280) utilizando un espectrofotómetro ND-1000 (Nanodrop). Las muestras cargadas se mezclaron con tampón de carga 6X (Sambrook *et al*., 1989). Como marcador de peso molecular se utilizó λHind III (Gene Ruler, Fermentas). Los geles se visualizaron en un transiluminador de luz UV a 312 nm tras haberlos teñido con bromuro de etidio (1 µg/ml) y su posterior lavado en agua destilada. Las imágenes se tomaron con el sistema fotográfico Uvidoc (Uvitec).

El DNA extraído se diluyó para poder cuantificar la densidad celular mediante qPCR empleando la curva estándar, cuyo rango comprende 0,3-0,03 ng/µl (apartado 1.7).

1.4.3- Extracción de ácidos nucleicos para la secuenciación de genomas.

Para el estudio de microdiversidad y mecanismos de diversificación de *S.ruber* (Capítulo 2) se crecieron cultivos puros de 6 cepas de *S. ruber* M1, RM158, SP73, SP38, Po13, P18 Cuando la $D.O._{600}$ de los cultivos alcanzó un valor de 0,5 se centrifugaron 20 ml (14.000 rmp, 10 min) y el pellet se resuspendió en 1 ml de SW 25% estéril. La extracción de DNA genómico para su secuenciación y la estimación de su calidad se llevó a cabo tal como se detalló en el anterior apartado.

1.5- Diseño de cebadores específicos de las cepas M8 y M31 de *S.ruber* para su estudio transcriptómico.

Se diseñaron cebadores específicos de cepa con el objetivo de detectar y cuantificar la abundancia relativa de las cepas M8 y M31 de *S.ruber* en los cultivos puros y mixtos a lo largo de las curvas de crecimiento. Se usaron secuencias de genes específicos de las cepas M8 y M31, y por lo tanto ausentes en cualquier otro microorganismo de la base de datos de nucleótidos del NCBI (nt/nr). Estas secuencias génicas están disponibles en la base de datos del NCBI, con los números de acceso NC_014032 para el genoma de M8 y NC_007677 para el de M31. Se realizó un *BLASTn* (http://blast.ncbi.nlm.nih.gov/Blast.cgi) (Altschul *et al*., 1990) de los genes candidatos contra la secuencia completa del genoma de ambas cepas. De este modo se descartó la presencia en los genomas de regiones homólogas a secuencias internas de las ORF seleccionadas que afectaran a la eficiencia de los cebadores por hibridación inespecífica. Tras el proceso de diseño, con un segundo *BLASTn* de las secuencias de los cebadores contra la base de datos de nucleótidos (nucleotide collection) (nr/nt) se descartó la posibilidad de amplificación de secuencias de DNA procedente de cualquier otro microorganismo halófilo del mismo ambiente.

El diseño de cebadores específicos se llevó a cabo mediante la herramienta Primer-Blast (Ye *et al.*, 2012) del NCBI (http://www.ncbi.nlm.nih.gov/tools/primer-blast/). Durante el proceso, esta herramienta tiene en cuenta la temperatura de hibridación (Tm), longitud de los cebadores, preferentemente 20 pb, y su contenido en GC. Como criterio de búsqueda se consideró una longitud del amplicón entre 200-300 pb, temperaturas de hibridación similares para ambos oligonucleótidos, no difiriendo en más de 5ºC para el par, y entre 55º-60ºC. Entre el listado de parejas candidatas, se seleccionaron manualmente aquellas que no presentasen un contenido en A+T mayor de 3 para las 5 entre las 5 primeras bases de su extremo 5', ni un G+C superior a 3 para las 5 bases situadas en 3', conteniendo en su región central, los 10 nucleótidos restantes, un elevado % C+G (**Tabla M2**).

Una vez diseñados los cebadores, se caracterizó su estabilidad termodinámica con el *software* Net Primer (Premier Biosoft International, http://premierbiosoft.com). Se estimó la probabilidad de formación de dúplex, heterodúplex u horquillas así como la estabilidad de los

fragmentos 5' y 3' de cada cebador, temperatura de hibridación (Tm) y la eficiencia teórica para cada pareja de cebadores. Se realizó una selección de cebadores para cada una de las ORFs escogidas. Estas estimaciones se obtuvieron para las concentraciones de nucleótidos, iones monovalentes y divalentes establecidas por defecto en dicho *software*: 250 pM de dNTPs, 50,0 mM de iones monovalentes, 1,5 mM de Mg^{2+} libre y 204,92 mM de Na^{+}. Los cebadores preseleccionados se probaron *in silico* a las temperaturas de 25ºC y 60ºC, la primera de las cuales constituye la temperatura estándar termodinámica, empleada como referencia a la hora de comparar la estabilidad molecular, y la segunda aquella a la cual se utilizarán los cebadores en la qPCR. El porcentaje de eficiencia individual por cebador se calculó mediante la fórmula: % Eficiencia= 100 + (ΔG dímeros) x 1,8 + (ΔG horquillas).

Tabla M2. Características de los 7 cebadores diseñados obtenidas mediante el programa Primer-Blast y las ORFs de los genes específicos consideradas para su diseño.

Anotación del gen y locus tag	Cebador	Secuencia (5'-3')	Tm (ºC)	Posición inicio/fin en el gen	G+C	Tamaño de amplicón
Ribonucleasa H SRU_1109	1109_2-F	GAGGGTCGCTATCGCATCTC	59,6	499/518	60%	261
	1109_2-R	ACGGTTCTCACTGGCATTCC	59,61	759/740	55%	
Ribonucleasa H SRU_1109	1109_4-F	CCACTCCTGAGGGTCGCTAT	58,97	491/510	60%	269
	1109_4-R	ACGGTTCTCACTGGCATTCC	59,6	759/740	50%	
Proteína hipotética SRU_614	614-F	CATGCTGAGCCGAGCAGTATT	60,51	271/291	52%	272
	614-R	GCCACGATCAGAAGCCAAGA	60,89	543/524	55%	
UDP-N-acetilglucosamina 2-epimerasa SRU_602	602-F	GTAGCCGCAGAGCCATATCG	60,85	457/476	60%	220
	602-R	AGACCTCACTGAGCGTGACA	55,7	658/639	55%	
Proteína hipotética SRM_00707	707-R	GGAGAGGAGGCTGAGGAGTATG	59,62	28/49	59%	220
	707-F	GACACATCCCACCCCAACAC	60,08	300/320	60%	

1.6- Amplificación de DNA mediante la reacción en cadena de la polimerasa (PCR).

1.6.1- PCR en gradiente.

Mediante PCR en gradiente se comprobó el rango de temperatura de hibridación para el cual los cebadores diseñados resultaron funcionales. Las amplificaciones se llevaron a cabo en un volumen de 10 µl, conteniendo cada tubo de reacción 1.5 mM de MgCl, 10mM de Tris-HCl, 50 mM de KCl, 200 µM de cada dNTP, 0,2 µM de cada cebador y 1 unidad de Taq DNA polimerasa (Invitrogen). En cada reacción se emplearon 40 ng de DNA extraído mediante *boiling*. La PCR se llevó a cabo en un termociclador Mastercicler Gradient (Eppendorf) con el siguiente programa de amplificación: desnaturalización inicial (94ºC, 5 minutos), 34 ciclos de desnaturalización (94ºC 30 segundos), hibridación (se probaron 8 temperaturas distintas, 30 segundos) y elongación (72ºC durante 1 minuto). Por último se llevó a cabo una extensión final (72ºC, 10 minutos). El gradiente considerado incluyó 8 valores diferentes de temperatura de hibridación entre 52,8ºC y 64,0ºC (52,8ºC; 55,3ºC; 56,9ºC; 58,5ºC; 60,2ºC; 61,7ºC; 63,0ºC; 64,0ºC).

1.6.2- Evaluación del rango de amplificación.

Una vez escogidos los cebadores para qPCR se procedió a cuantificar el rango de concentración de DNA en el cual rendían un producto detectable en gel de agarosa. Para ello se empleó como molde DNA de las cepas M8 y M31 extraído mediante el *kit* de extracción Dneasy Blood and Tissue (QIAGEN) (véase apartado 1.4.2). La reacción se llevó a cabo en un volumen total de 10 µl, conteniendo: 5 mM de MgCl, 10 mM de Tris-HCl, 50 mM de KCl, 0,2 µM de cada cebador, 200 µM de cada dNTP, 1 unidad de Taq DNA polimerasa (Invitrogen). Se emplearon las siguientes concentraciones de DNA molde: 30 ng/µl, 3 ng/µl, 0,3 ng/µl, 0,003 ng/µl y 0,0003 ng/µl. El programa de amplificación empleado comprendió una etapa de desnaturalización inicial (94ºC, 3 minutos); 35 ciclos de desnaturalización (94ºC, 30 segundos), hibridación (60ºC, 1 minuto), elongación (72ºC, 2 minutos). Por último se añadió una extensión

final (72ºC, 10 minutos). Las amplificaciones se realizaron en un termociclador Mastercycler Gradient (Eppendorf).

1.6.3- Comprobación de la ausencia de contaminación en cultivos puros.

Se analizó mediante PCR con los cebadores específicos la pureza de los cultivos de las cepas M8 y M31 en los experimentos de transcriptómica y de cultivos puros del resto de cepas empleadas en los estudios de microdiversidad y mecanismos de diversificación (**figura M1**). En el caso de los experimentos de transcriptómica, se usaron como molde los DNAs de *boiling* de los diferentes puntos de las curva (véase apartado 1.8) comprobando la ausencia de contaminación en los cultivos puros y la presencia de ambas cepas en el mixto. Las reacciones se llevaron a cabo en un volumen final de 25 µl empleando las mismas proporciones de reactivos y programa de amplificación detalladas en el apartado 1.6.2 y los cebadores específicos de cepa 1109_4F/R (M31) y 308_2F/R (M8). Se emplearon 45 ng de DNA molde. Los productos de amplificación se sometieron a electroforesis en geles de agarosa Seakem LE (FMC Bioproducts) al 2% en tampón TAE 1X (véase apartado 1.4.2). Se cargaron aproximadamente 20 µl de producto de amplificación. Como marcador de peso molecular se utilizó DNA 1 Kb DNA Ladder Plus (Invitrogen). Los geles se visualizaron en un transiluminador de luz UV (véase apartado 1.4.2).

Por otro lado, parte del DNA extraído para la secuenciación de genomas y análisis microevolutivo (capítulo 3), se empleó para comprobar la ausencia de contaminación con DNA de organismos del dominio *Archaea* y de las cepas de *Salinibacter* más empleadas en el laboratorio, M8 y M31. Las reacciones de amplificación se llevaron a cabo en un volumen final de 50µl. Para descartar la contaminación por M8 y M31 se emplearon los cebadores específicos de cepa diseñados para el experimiento de transcriptómicas y el mismo programa. Además se realizó una segunda reacción de PCR empleando esta vez cebadores universales para los dominios *Bacteria* y *Archaea* (**tabla M3**), con el objetivo de descartar la presencia de contaminación por microorganismos del dominio *Archaea* en las distintas fases de crecimiento. Las condiciones de amplificación para los dominios *Bacteria* y *Archaea* fueron: una etapa de

desnaturalización inicial (94ºC, 3 minutos); 34 ciclos de desnaturalización (94ºC, 15 segundos), hibridación (55ºC, 30 segundos), elongación (72ºC, 2 minutos). Por último se añadió una extensión final (72ºC, 10 minutos). Las amplificaciones se realizaron en un termociclador Mastercycler Gradient (Eppendorf).

1.7- PCR cuantitativa (qPCR).

Se determinó la abundancia de cada cepa en cultivo puro y mixto a lo largo de las curvas de crecimiento del estudio transcriptómico mediante qPCR (capítulo 1, véase **figura M1**). Se realizaron en placas de 96 pocillos empleando el sistema ABI PRISM 7000 Sequence Detection System (Applied Biosystems). El programa consistió en 40 ciclos en los cuales se sucedieron desnaturalización (95ºC, 15 segundos), hibridación (60ºC, 1 minuto) y elongación (72ºC, 15 segundos). Por último se incluyó una etapa de disociación final (95ºC, 15 segundos; 60ºC, 15 segundos; 95ºC, 15 segundos).

Con el objetivo de valorar la **eficiencia, sensibilidad** y **especificidad** de las parejas de cebadores diseñadas, específicas de M8 (707 y 308_2) y de M31 (1109_4 y 614), se preparó un banco de diluciones decimales empleando DNA de las cepas M8 y M31 extraído desde cultivos puros con el *kit* Dneasy Blood and Tissue (QIAGEN) tras cuantificar la densidad celular de los mismos mediante tinción con DAPI y posterior recuento. El banco incluyó concentraciones seriadas de DNA en el rango 10^5 a 10^2 copias genómicas, correspondientes a concentraciones de DNA del rango 3-0,003 ng/µl. El DNA de cada una de estas diluciones se amplificó, y el análisis gráfico de los resultados permitió evaluar las **curvas de disociación, amplificación** y **estándar o de calibración** para cada uno de los cebadores diseñados. Las amplificaciones se llevaron a cabo en un volumen de 30 µl, conteniendo cada pocillo 10 µl de *mix* de amplificación SYBR Green (Applied biosystems), 0,4 µl de cada cebador, 4,2 µl de agua milliQ y 5 µl de DNA molde. Se seleccionaron las dos parejas que presentaron una eficiencia y curva de amplificación similares. Los resultados se analizaron con el programa qPCR 700 system SDS Software (Applied Biosystems) que permite visualizar los datos representados en forma de curvas de amplificación, disociación y estándar.

Una vez seleccionadas las dos parejas de cebadores 1109_4 y 308_2, se comparó la similitud de eficiencia y los datos cuantitativos obtenidos con respecto a los cebadores 338f y 500r (**tabla M3),** que amplifican una región del gen del rRNA 16S. Para esta comparación se utilizaron sus respectivas curvas de calibración y las diferencias en los valores de c_t para las muestras estándar.

Por último, se comprobó la especificidad de estas dos parejas de cebadores, (1109_4 y 308_2), mediante una segunda qPCR incorporando a la placa de un triplicado de reacciones cruzadas. En estas reacciones se empleó DNA molde de la cepa contraria para la cual amplifican los cebadores específicos. Además se incluyeron controles negativos NTC (*No Template Control*), que contienen todos los elementos y componentes de la reacción a excepción de DNA molde. Finalmente, la cuantificación absoluta de los puntos de las curvas de crecimiento de los cultivos de M8 y M31, tanto puros como mixtos se llevó a cabo mediante las parejas de cebadores 1109_4 y 308_2, empleando los parámetros descritos anteriormente en este mismo apartado.

Tabla M3 Cebadores del gen rRNA 16S empleados en las reacciones de PCR y qPCR

Cebador	*Secuencia 5′ → 3′*	Posiciones	Especificidad	Referencia
338f	ACT CCT ACG GGA GGC AGC	338-355*[1]	*Bacteria*	(Amman, 1995)
500r	TTA CGC GGC TGC TGG CAC G	*[1]	*Bacteria*	(Amman, 1995)
21F	TTC CGG TTAGA GTT TGA TCA	2-21	*Archaea*	(De Long, 1992)
27f	TGG CTC AGG ATC CTG CCG GA	8-27	*Bacteria*	(Lane, 1991)
1492r	GGT TAC CTT GTT ACG ACT T	1510-1492	*Archaea y Bacteria*	(Lane, 1991)

*[1] Posiciones de *Escherichia coli* (Bacteria y Archaea).

1.8- Extracción de RNA y eliminación de rRNA. "Librería" y secuenciación del RNA.

Se extrajo el RNA de dos réplicas de los cultivos mixtos y una de cada cultivo puro. Se seleccionó el mismo punto en mitad de fase exponencial (82 horas de crecimiento) en todos los casos. Se recuperaron las células por centrifugación (15 min, 3900xg, 4ºC) y se resuspendió el pellet en TE (Tris-HCl 100 Mm, pH8; EDTA 100 mM, pH8) añadiendo posteriormente 10 µl lisozima (300 mg/ml). La extracción de RNA total se realizó con el *kit* RNeasy Mini Kit, de QIAGEN. Los RNAs totales se resuspendieron en 100 µl de agua libre de nucleasa. El pellet de células contenía cerca de 3 $x10^9$ células en los 4 casos. Los extractos se digirieron con 2 µl de TURBO DNase (Ambion), incubando durante 1 hora a 37ºC. Tras la inactivación de la enzima, los RNAs totales se sometieron a una nueva electroforesis para visualizar la eliminación del ADN genómico. Los RNAs se precipitaron y concentraron en un volumen de 17 µl, y se emplearon 2 µl para la cuantificación y el análisis de la calidad del RNA mediante bioanalizador (Agilent). A continuación se eliminaron de los rRNAs y tRNAs en dos fases secuenciales: la eliminación de los rRNA 16S y 23S con el *kit* MICROBE Express (Ambion) y la del rRNA 5S y tRNAs con el *kit* MEGA Clear (Ambion). La eficiencia en la eliminación de los rRNAs 16S y 23S (1ª fase) y del rRNA 5S y los tRNAs (2ª fase) se comprobó mediante bioanalizador (Agilent).

La preparación de la librería y posterior secuenciación se llevó a cabo por el servicio de Genómica del Centre for Genomic Regulation (CRG) del Parc de Recerca Biomédica de Barcelona (PRBB). La fragmentación del RNA se realizó a 94ºC durante 2 min con el tampón de fragmentación del *kit* "Illumina mRNA sequencing Sample Prep". Los perfiles de las muestras fragmentadas se visualizaron en un QC con el bionalalizador seleccionando para las librerías los fragmentos de entre 215-315 pb. Después de la fragmentación del mRNA se retrotranscribió el RNA a cDNA con la enzima SuperScript II. En cada carrera de secuenciación mediante Illumina Hiseq se obtuvieron entre 39 y 44 millones de secuencias pareadas, *pair end reads*, de 100pb con un tamaño de inserto, *insert size*, de 200 pb.

1.9- Análisis de la composición iónica del medio extracelular.

Se repitieron las curvas de crecimiento del estudio de interacción entre las cepas M8 y M31 de *S. ruber* monitorizándolas del mismo (apartado 1.2), con el objetivo de analizar la composición iónica del medio durante la fase media exponencial. Se tomaron 10 ml del sobrenadante de cultivos puros y mixtos en 5 puntos tiempos y se centrifugaron (15 min, 3900xg, 4ºC). Los sobrenadantes se filtraron por filtros de 0.45 μm (Millipore). Posteriormente se diluyeron 1:10 con agua mQ para la cuantificación iónica de los elementos Fe, Ni, Cu, Mg dentro del rango 0.5-10 ppb y Mg dentro de los límites 0.5-50 ppb mediante Espectrometría de Masas con fuente de Plasma de Acoplamiento Inductivo (ICP-MS) (unidad de Genómica y Proteómica, SSTTI UA).

2. ANÁLISIS Y ESTUDIOS *IN SILICO*.

Los análisis *in silico* desarrollados en esta tesis comprenden el análisis de las secuencias de los transcriptomas de *S. ruber* M8 y M31 en cultivo puro y mixto (capítulo 1), el estudio de los mecanismos de microevolución de *S. ruber*, incluyendo el ensamblaje y anotación de cepas completas (capítulo 2) y la totalidad análisis del efecto de la recombinación homóloga en la microevolución de 54 especies procariotas y los factores que influyen en la misma (capítulo 3).

2.1- Análisis de datos de expresión obtenidos mediante RNA seq. Estudio de los transcriptomas puros y mixtos.

2.1.1- Análisis de expresión y detección de ortólogos.

Los archivos de secuenciación de las cuatro muestras procesadas, dos réplicas del cultivo mixto y un cultivo puro de cada cepa, se procesaron inicialmente para eliminar los *reads* de baja calidad. Se recortaron todas aquellas secuencias con valores de **calidad PHRED** inferiores a 10. Esto equivale a cortar secuencias con un error cada 10 bases, es decir con una precisión en secuenciación por base inferior al 90%. Tras el recorte, las parejas con longitudes en alguno de los *reads* menores de 31 bases se descartaron. Las secuencias restantes se mapearon empleando el alineador BWA v0.6.1-r104 (Li y Durbin 2010) contra los genomas, cromosomas y plásmidos, de M8 (NC_014032.1; NC_014028.1; NC_014157.1; NC_014026.1; NC_014030.1) y M31 (NC_007677.1; NC_007678.1).

Los perfiles de expresión en cultivo puro se normalizaron teniendo en cuenta el tamaño de la muestra, millones de secuencias y la longitud de los transcritos, obteniendo los valores de expresión para cada gen en RPKM (*Reads* por Kilobase por Millón de *reads* mapeados) con el programa Cufflinks 1.1.0 (Tranpnell *et al*., 2010).

Durante los análisis de expresión diferencial entre cultivos puros y mixtos para cada cepa se alinearon de nuevo las secuencias pero esta vez contra un genoma de referencia concatenado que contenía los replicones de ambas cepas. Se eliminaron aquellas parejas de lecturas que

mapeaban en genes con secuencia idéntica para ambas cepas, ya que resulta imposible asignar de modo inequívoco estos a una u otra cepa en cultivo mixto. Para ello se signaron los pares de ortólogos entre los genomas de ambas cepas mediante Blast bidireccional, empleando un E-value de corte de 1e-05 (Gabaldón *et al.*, 2008). Se normalizaron las secuencias mapeadas por el tamaño de la muestra y la longitud de los genes ortólogos. La detección de los genes expresados diferencialmente se llevó a cabo mediante los programas Deseq (Anders y Huber., 2010) y Cufflinks, empleando un E-value de corte de $1x10^{-5}$.

2.1.2- Reanotación de genomas, mapeo de genes a vías metabólicas del KEGG y análisis de enriquecimeinto mediante test de Fisher.

Se reanotaron los genes codificados en los genomas de ambas cepas mediante la herramienta Blast2GO (Gotz et al., 2008), recuperando la **anotación en GO *terms*** y los **números EC** para los genes que mapeaban en vías metabólicas del KEGG. Se mapearon contra las vías del **KEGG** (Kanehisa y Goto 1999) los genes que presentaron expresión diferencial tanto al comparar los cultivos puros entre sí como al comparar los cultivos puros con los mixtos con el objetivo de observar patrones de cambio o diferencias en varios genes dentro de una misma ruta. Se realizaron test de enriquecimiento (**Fisher's Exact Test**) con estos subconjuntos de genes, tomando como referencia los genomas completos, y empleando tanto anotación en **términos GO** (GO *terms, Gene Ontology*) como *clusters of othologous groups* **(COG)**. Se consideraron significativos enriquecimientos con un E-value inferior $1x10^{-5}$.

Por medio de la herramienta ECF Finder (http://ecf.g2l.bio.uni-goettingen.de:8080/ECFfinder/) se confirmaron los factores sigma extracitoplasmáticos anotados previamente en ambos genomas. Por medio de los programas SignalP 4.1v Server (Petersen *et al*., 2011) and TAT Find 1.4v (Brueser *et al*., 2005) se identificaron los genes que cofidican proteínas con péptido señal o péptidos de translocación *twin-arginine translocation* (TAT), respectivamente.

2.2- Mecanismos de microdiversidad de cepas de *S.ruber*.

El estudio microevolutivo de las cepas aisladas de *S.ruber* constituye un trabajo de desarrollo mayoritariamente *in silico* (véase **figura M3**) que comprende una primera etapa en la cual se secuenciaron los genomas completos de 6 cepas aisladas de *S.ruber* (**Tabla M1**). Una vez completos los replicones, se procedió a la reanotación semiautomática de los mismos, incluyendo datos procedentes de análisis transcriptómicos (RNAseq capítulo 1). En una segunda etapa se analizaron los mecanismos de diversificación que determinan la elevada microdiversidad de la especie, evaluando el efecto de la recombinación homóloga y los factores que influyen en ésta empleando una metodología similar a la del capítulo 2, y profundizando en el contenido y diversidad de plásmidos, caracterización y dinámica de HRVs y el papel de los fagos y sistemas CRISPR-Cas.

2.2.1- Ensamblaje de genomas: combinación de ensamblajes, identificación de plásmidos y cierre de regiones no ensambladas.

Se analizó la calidad de las secuencias obtenidas mediante el programa FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc) (Andrews 2010), evaluando la distribución de calidades y G+C por posición, los promedios de calidad (*PHRED quality score*) y contenido en G+C por secuencia, distribución de longitud de secuencias y niveles de duplicación. Las secuencias se procesaron con un *script* en Python a diferentes umbrales de calidad PHRED: q10, q20, q30.

La estrategia de ensamblaje combinó los resultados de 3 ensambladores: SOAPdenovo (Luo *et al*., 2012), IDBA-UD 1.1.0 (Peng Y *et al*., 2012), JR-Asembler (Chu *et al*., 2013). Se emplearon, en pruebas independientes con cada ensamblador, los archivos de secuencias filtrados a diferentes niveles de calidad. En el caso de SOAPdenovo se ensayaron múltiples **intervalos de k-mer**. Tras cada ensamblaje se obtuvieron los datos estadísticos que permitieron seleccionar los mejores ensamblajes en cada caso y compararlos entre sí: número total de *contigs*, *contigs* mayores de 1kb, **N50**, **N90**, número de indeterminaciones (Ns), longitud del *contig* mayor y el

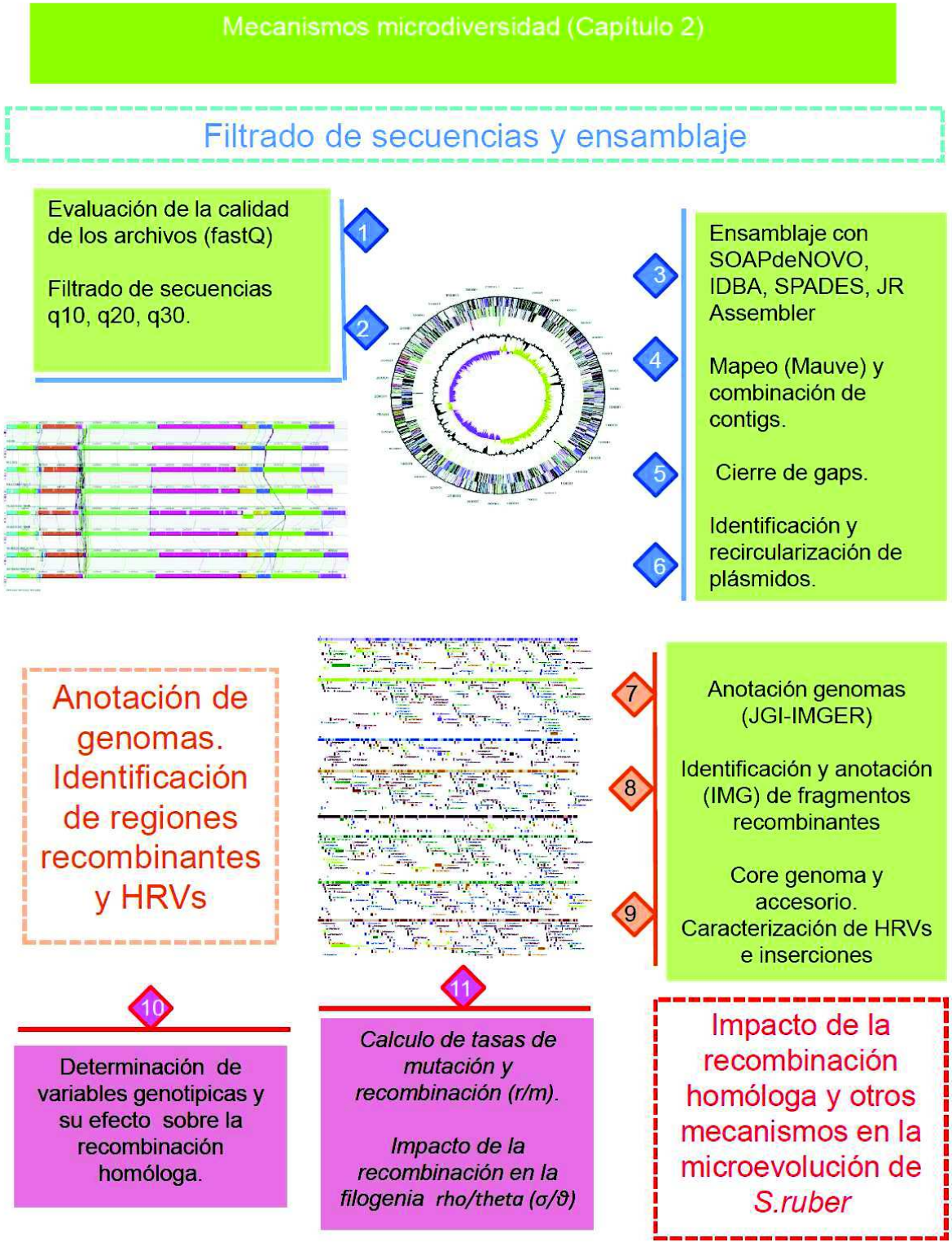


Figura M3. Esquema del diseño experimental *in silico* del capítulo 2. Los números indican la secuencialidad y orden de los análisis realizados.

empleó el programa GapCloser del paquete SOAPdenovo2 (Luo *et al*., 2012).

A continuación se mapearon los *contigs* contra los genomas de referencia de las cepas M8 y M31 de *S.ruber* empleando el programa progressive Mauve del paquete MAUVE 2.3.1v (Darling., *et al* 2010) con el objetivo de reorientar los *contigs*, por medio de la secuencia reversa complementaria, en la misma dirección y posicionarlos. El reordenamiento y reorientación se llevó a cabo con *scripts* en Python. Por último se combinaron aquellos *contigs* solapantes de diferentes ensamblajes con el objetivo de fusionar los alineamientos en uno final con el menor número de *contigs* posible.

Para cerrar las regiones restantes no ensambladas, se extrajeron las secuencias de los extremos de los *contigs* resultantes y se extendieron con la herramienta de mapeo en referencia de la plataforma *Geneious* 7.1.7. Con los extremos resultantes de la extensión se generó un archivo fasta para comparar todos contra todos usando la herramienta BLAST 2 sequences (Tatusova y Madden, 1999). Tomando las coordenadas se combinaron los extremos de los *contigs* colindantes cerrando los huecos y completando los genomas.

En el caso de los *contigs* pertenecientes a plásmidos, se trataron del mismo modo paralelamente. Entre las comprobaciones y requisitos a la hora de identificarlos se tuvieron en cuenta los siguientes:

1. Recircularización con la molécula y no mapeo con extremos de *contigs* vecinos.
2. Ausencia de mapeo o hit con los cromosomas de las cepas de referencia M8 y M31.
3. Presencia de un gen codificante para una proteína implicada en la replicación de plásmidos de bajo número de copia (RebB/ParA).
4. Presencia de *contigs* del mismo tamaño o muy similar en diferentes ensamblajes con las características anteriores.

2.2.2- Predicción de ORFs y reanotación con RNAseq.

La anotación de los genomas ensamblados se llevó a cabo en la plataforma Integrate Microbial Genomes (IMG) (Joint Genomes Institute) (http://www.jgi.doe.gov/) (Markowitz *et*

al., 2014) empleando el algoritmo habitual en anotación de genomas aislados (IMG ER). La predicción y anotación de genes se completó con la disponible hasta la fecha para el genoma de la cepa M8 depositada en el NCBI (NC 1432.1 a NC1426.1) y con la identificación de nuevos genes no considerados por ninguna de las dos anteriores anotaciones. Por comparación manual empleando el programa IGV (Integrative Genome Browser) (Robinson *et al*., 2012), se seleccionaron aquellos genes predichos y expresados significativamente de la anterior anotación que no se incluyeron por la plataforma IMG.. Se confirmó la expresión de nuevas ORFs predichas en la nueva anotación del IMG y no incluidas en la antigua en base a los datos de cobertura para las posiciones que abarcaban estos genes. Posteriormente se establecieron las coordenadas posicionales en los nuevos genes por medio del programa Exonerate 2.2 (Slater y Birney 2005). Las secuencias de los genes identificados se incorporaron a los nuevos archivos GeneBank y fasta de aminoácidos y nucleótidos para las secuencias codificantes. Una vez completo el genoma de M8 combinando la anotación nueva, la antigua y los datos de expresión, se utilizaron las secuencias de sus genes para completar la anotación del resto de genomas.

2.2.3- Identificación de regiones recombinantes y enriquecimientos. Puntos calientes de inserción.

Para la identificación de eventos recombinantes se empleó el programa RDP4 v4.15 (Martin *et al;* 2010) y los parámetros empleados en el estudio de recombinación homóloga (capítulo 2, véase apartado 2.3.4). Una vez identificados los puntos de ruptura, se extrajeron las secuencias de los eventos de cada cepa mediante un *script* en Python, y se anotaron del mismo modo que en el análisis in sílico desarrollado en el capítulo 2 (véase apartado 2.3.6) por medio de la plataforma IMG (http://www.jgi.doe.gov/). Una vez anotados los genomas completos y los eventos recombinantes, se llevaron a cabo los test de enriquecimientos (*Fisher's exact test*), con la anotación COG y GO *terms (*véase apartado 2.3.7*).*

2.2.4- Sintenia, identificación de 5' UTR y conservación de operones.

Se llevó a cabo una delimitación manual de operones e identificación de genes con región promotora 5' UTR (Untranslated Region) empleando los datos transcriptómicos de las cepas M8 y M31 y el programa IGV (Integrative Genome Viewer) (Robinson *et al*., 2012). Se optó por esta estrategia al tratarse de una librería no específica de hebra. Antes de la delimitación manual se estimó un umbral de cobertura por encima del cual considerar que una región intergénica se expresa significativamente. Este umbral se calculó en base a los criterios descritos en estudios anteriores (Kumar *et al*; 2012), considerando el valor correspondiente al primer decil, siendo el que representa al menos el 10% de las posiciones de genoma. Los valores de cobertura por posición se calcularon con el programa Bedtools (Quinlan *et al*., 2010).

Los datos obtenidos se compararon con las predicciones *in silico* depositadas en la base de datos pública MicrobesOnline (Dehal *et al;* 2010) para ambas cepas. Estas predicciones se obtuvieron mediante un algoritmo que tiene en cuenta distancias intergénicas, relación funcional GO entre genes colindantes y que pertenezcan a la misma COG.

Una vez determinados los ortólogos posicionales, se llevó a cabo un estudio sinténico de los operones de todos los genomas, valorando la conservación del orden de los genes en cada caso e inserciones o pérdida de ORFs si las hubiera.

2.2.5- Genomas *core* y accesorio, dN/dS y CAI.

Se recuperó la secuencia de aminoácidos desde los archivos GeneBank para los ortólogos posicionales detectados previamente. Se llevaron a cabo alineamientos 2 a 2 entre las cepas con el programa MUSCLE v3.0 (Edgar *et a*l; 2004). Los alineamientos de proteínas se convirtieron en alineamientos basados en codones con el programa trimAl v1.3 (Capella-Gutierrez *et al*; 2009) obteniendo las correspondientes secuencias codificantes. Finalmente los valores de dN/dS se calcularon con el programa CodeML (modo *pairwise*, modelo 1 NS sites, 0 parámetros) del paquete PaML v4.4 (Yang *et al*; 2007).

2.2.6- Caracterización de las zonas hipervariables y plásmidos. HGT.

Se identificaron y delimitaron las zonas hipervariables (HRVs) de cada cepa tras recuperar los ortólogos posicionales y alinear los cromosomas completos con el programa Mauve 2,3,1v (Darling *et al.,* 2010). Se observó la ruptura de sintenia en estas regiones mediante los visualizadores implementados en la platforma JGI (IMG), incluyendo el programa ACT Artemis (Carver *et al*., 2011). El origen filogenético de las secuencias codificadas dentro de las HRVs se determinó con la herramienta de análisis de perfil filogenético dentro de la plataforma JGI(IMG) (Makowitz *et al*., 2012).

2.2.7- Caracterización de variables genómicas, barrera y movilidad y correlación con niveles de recombinación homóloga. Impacto de la recombinación homóloga en la filogenia.

Empleando los mismos criterios que en el análisis de recombinación homóloga desarrollado en el capítulo 3 de esta tesis, se recuperó la información genómica y ecológica de *S. ruber*, construyendo una base de dato análoga (véase apartado 2.3.7, tabla 4M). De nuevo se estimaron las tasas de recombinación y mutación mediante el software ClonalFrame (Didelot y Falush 2007) empleando los parámetros descritos anteriormente (véase apartado 2.3.5).

2.2.8- Identificación y caracterización de sistemas CRISPR-Cas y sus espaciadores.

Se empleó la herramienta online CRISPRfinder (Grissa *et al*., 2007) para identificar las agrupaciones CRISPR, espaciadores y repeticiones, así como las proteínas cas presentes en los genomas de las cepas. La anotación de las secuencias de proteínas Cas se completó realizando un Blastn contra la base de datos del NCBI y contra una base de datos de genes *cas* (Grissa *et al*., 2007). Con el objetivo de encontrar el origen de las secuencias protoespaciadoras, se generó un fasta con las secuencias de espaciadores para realizar reclutamientos de secuencia locales contra metaviromas y secuencias de halofagos que infectan a *S.ruber* y H*aloquadratum walsbyi*.

2.3- Análisis de la incidencia de la recombinación homóloga en genomas completos.

2.3.1- Diseño experimental.

El análisis de recombinación con genomas completos constituye un trabajo realizado en su totalidad *in silico* **(figura M2).** Durante el estudio se trabajó con genomas completos de 54 especies procariotas diferentes que presentasen al menos 3 cepas secuenciadas completamente. Los de genomas completos permitieron la identificación de bloques sinténicos y genes colineares. A partir de estos bloques sinténicos se identificaron *in silico* regiones candidatas a haberse transferido mediante recombinación homóloga entre cepas cercanas.

La anotación de estas regiones permitió interpretar posibles estrategias microevolutivas asociadas al ambiente. Se reunieron datos de variables genómicas (tamaño de islas genómicas, tamaño genómico, número de operones ribosomales y tRNAs) desde diferentes bases de datos (véase **tabla M4**), variables de homología (ANIb, % genoma recombinable) , variables de recombinación (eventos recombinativos por cepa) de relevancia ecológica (modo de vida), variables que favorecen el intercambio de DNA (contenido en elementos de transposición, genes de transferencia horizontal (HGT), y genes relacionados con los mecanismos de competencia y conjugación). Por último se recopiló información sobre variables que dificultan dicho intercambio (contenido y tipos de sistemas de restricción-modificación y CRISPR-Cas). Además se analizó el impacto de la recombinación homóloga sobre la filogenia de cada especie y la influencia de la recombinación y la mutación en la microdiversidad observable en cada grupo (véase **tabla 4M**).

2.3.2- Construcción de la base de datos. Especies consideradas en el estudio.

Para el análisis de recombinación en genomas procariotas se seleccionaron todos los genomas completamente ensamblados y depositados en el directorio FTP del NCBI (http://www.ncbi.nlm.nih.gov/genomes) en febrero de 2012. Entre todos ellos se escogieron

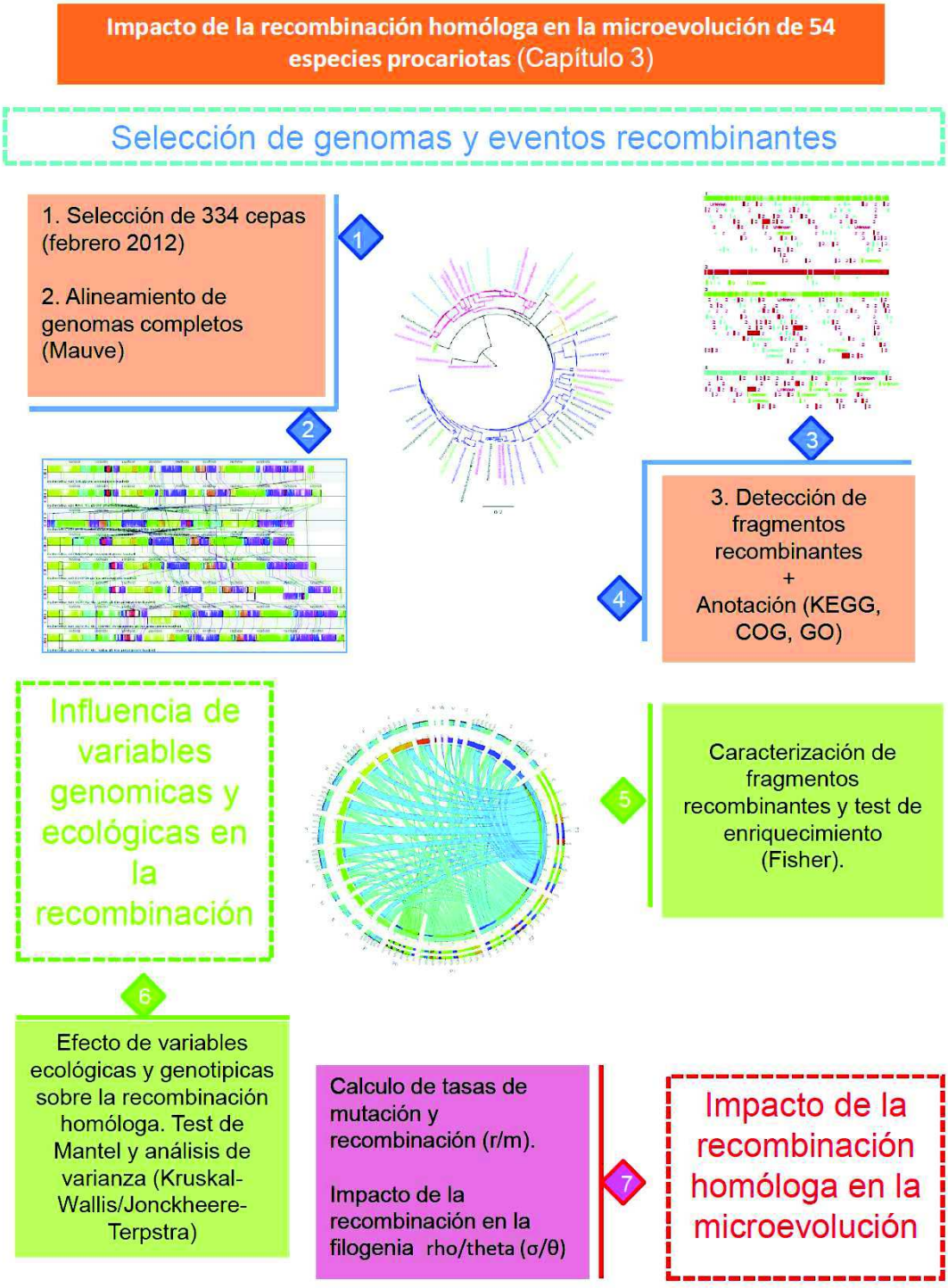


Figura M4. Esquema de trabajo experimental del capítulo 3. La numeración refleja la secuencia temporal de los análisis realizados.

aquellos casos en los que se encontraban al menos 3 cepas para una misma especie, recuperando entre los anteriores un total de 338 genomas, 325 del dominio *Bacteria* y 13 del dominio *Archaea* pertenecientes a 54 especies. Además se incluyeron 33 genomas de organismos aislados no caracterizados pertenecientes a 6 grupos taxonómicos distintos con identidad a nivel de género. Para confirmar que las cepas consideradas para las 54 especies pertenecían en cada caso a la misma especie se comprobó la identidad a nivel de RNA 16S alineando las secuencias con el programa Silva (Pruesse *et al*., 2007). Se consideraron cepas de la misma especie aquellas con una identidad rRNA 16S > 98.7%.

2.3.3- Alineamiento de genomas completos, identificación de ortólogos posicionales y ANIb.

Los cromosomas de las cepas de cada una de las especies se alinearon entre sí con el programa Progressive Mauve del paquete MAUVE 2.3.1v (Darling *et al.,* 2010). De este modo se obtuvieron los bloques colineares entre cepas, que son precisamente los candidatos a presentar eventos de recombinación homóloga, en forma de alineamiento en formato multifasta extendido (*xmfa). Para ello se emplearon los archivos GeneBank de cada una de ellas. Además el programa proporciona el listado de los ortólogos posicionales localizados dentro de dichos bloques. Mediante un *script* en Python se seleccionaron los genes compartidos por todas las cepas de la especie (**genoma *core***) y aquellos compartidos por algunas de ellas o específicos de cepa (**genoma accesorio**) (Medini *et al*., 2005).

Por otro lado, los genomas de cada especie se alinearon 2 a 2 con el programa Nucmer del paquete MUMmer (Kurtz *et al*; 2004). Los archivos de coordenadas generados (*.coords) se procesaron con un *script* en Python para calcular el promedio de identidad de nucleótidos (ANIb) entre parejas. Al final se obtuvo un valor de ANIb para cada par de genomas promediando la identidad de cada fragmento homólogo. Se ponderó el peso de cada fragmento empleado en el cálculo de ANIb en base su longitud respecto del cromosoma.

2.3.4- Detección y caracterización de eventos recombinantes.

Con el fin de identificar los eventos de recombinación homóloga acontecidos entre cepas de una misma especie se realizaron los alineamientos genómicos con el programa RDP4 v4.15 (Martin *et al;* 2010), que combina varios métodos de detección de este tipo de eventos.

Se utilizaron los archivos de los alineamientos realizados con progressive Mauve. Con el programa RDP4 se realizó una primera fase exploratoria en la que se emplearon cuatro métodos de detección de eventos recombinantes: RDP (Martin, 2000), GENECONV (Padidam, 1999), MaxCHI (Maynard Smith, 1992), Chimaera (Posada y Crandell 2001). En la etapa de detección secundaria el programa reescaneó de manera automática cada elemento detectado con 5 programas: RDP (Martin *et al*., 2010), GENECONV, MaxCHI, Chimaera y 3Seq (Boni *et al*., 2007). Sólo las señales de recombinación confirmadas por al menos 3 de los 5 métodos se aceptaron como evidencias de recombinación, considerando como valor de corte $p<0.001$.

Los parámetros de tamaño de ventana (*Window Size*) para el programa RDP se ajustaron a 90 nucleótidos y un valor de sitios variables y ventana de 210 en el caso de MaxCHI. Para el resto de programas implementados se emplearon los parámetros por defecto. Los eventos detectados se procesaron manualmente, considerando en todos los casos como las **posiciones de ruptura** más probables aquellas inferidas mediante el método MaxCHI (que es considerado el método que detecta con mayor precisión las posiciones de ruptura entre los 5 métodos de detección no paramétricos implementados en el RDP4). A partir de estos datos se obtuvieron el porcentaje de genoma recombinado y los eventos por cepa, que constituirán las dos variables de recombinación en los análisis estadísticos posteriores. Además se calculó el %G+C y la distribución de eventos de recombinación dentro de 4 intervalos discretos de tamaño (0-2Kb; 2-10Kb; 10-50Kb; >50Kb).

2.3.5- Análisis evolutivo: Cálculo de tasas de mutación y recombinación.

Se reconstruyó la genealogía clonal de cada especie individualmente mediante el programa ClonalFrame (Didelot *et al*, 2007). Usando los alineamientos completos de los genomas generados con la herramienta progressive Mauve (Darling *et al*; 2010), se extrajo el

core de los alineamientos, definido como aquellos bloques o regiones alineadas con al menos 500 pb de longitud. Se llevaron a cabo tres carreras independientes con Clonal Frame, cada una de 40.000 iteraciones. La primera mitad de las mismas se descarta en el proceso conocido como ***burn-in***. La convergencia de los triplicados se comprobó manualmente, asegurando la consistencia de las estimaciones para los parámetros globales **r/m** (donde r representa la tasa de recombinación y m la de mutación) y **rho/theta** (σ/θ) con un intervalo de confianza del 95% y la genealogía clonal.

2.3.6- Anotación de genomas completos y de secuencias de las regiones recombinantes.

Se descargaron los archivos de anotación en formato GeneBank para las 338 cepas desde la base de datos Integrate Microbial Genomes (IMG) (Joint Genome Institute) (http://www.jgi.doe.gov/). Esta plataforma proporciona acceso público a los genomas depositados en el NCBI, algunos de ellos reanotados y actualizados. Mediante un *script* escrito en Python se recuperaron las secuencias afectadas por eventos de recombinación en cada genoma teniendo en cuanta las coordenadas proporcionadas por el programa RDP4. Se hizo una predicción de genes *de novo* para estas secuencias. En el caso de secuencias menores de 10 Kb la predicción y anotación de ORFs se realizó con el algoritmo implementado para secuencias metagenómicas (IMG/M ER), y para las mayores de 10Kb se empleó el habitual en anotación de genomas aislados (IMG ER).

La anotación de los fragmentos recombinados desde la misma plataforma permitió realizar posteriormente análisis de enriquecimientos empleando como referencia los genomas completos y sin sesgos derivados de anotación. Además esta plataforma estandariza los criterios de anotación entre cepas y especies pudiendo comparar los resultados obtenidos entre ellas. Desde la plataforma se obtuvo la anotación en COG para las regiones recombinadas y genomas completos además de datos como el tamaño de los fragmentos recombinados, su distribución, contenido en genes y el %G+C.

2.3.7- Análisis estadístico.

Variables genómicas y ecológicas consideradas en el estudio. Construcción de bases de datos.

Para las 338 cepas incluidas en el estudio se obtuvieron datos genómicos, ecológicos, de presencia o impacto de la recombinación, mobiloma o variables que representan intercambio genético y variables que dificultan el mismo o barrera (véase **tabla M4**). Se exploraron variables de tipo ecológicas, genómicas, de homología, de recombinación, variables referentes a mecanismos de movilidad y barrera, entendidas estas dos últimas como aquellas que favorecen/dificultan el intercambio génico.

Se recuperaron datos ecológicos para las cepas analizadas desde la base de datos Integrated Microbial Genomes (IMG) del Joint Genome Institute (JGI) y revisiones en las que se evalúan los niveles de recombinación obtenidos en estudios por MLSA (Vos y Didelot 2009; Didelot *et al*., 2010). Las especies se agruparon en 4 clases ecológicas: (0) simbiontes y patógenos intracelulares, (1) no patógenos (comensales y de vida libre), (2) patógenos obligados y (3) patógenos oportunistas.

Además, para cada genoma incluido se obtuvo desde los archivos GeneBank con un *script* en Python la distribución de número de operones ribosómicos, transposones, tRNAs, y proteínas implicadas en procesos de reparación, recombinación y competencia. Como genes implicados en funciones de reparación se consideraron los 21 descritos y caracterizados a lo largo de 900 genomas bacterianos en un estudio previo (García-González *et al*; 2013). Dentro de los genes implicados en conjugación se tuvieron en consideración los 35 incluidos en los operones pili tipo IV *pil, tad* y *com* (Inam *et al*; 2011) y entre los necesarios para la transformación se incluyeron 22 genes pertenecientes al regulón *com* (Claverys *et al;* 2006).

Tests de enriquecimientos para las regiones recombinantes (*Fisher's Exact Test*).

Se realizaron tests de enriquecimiento (***Fisher's Exact Test***) (FT) con la anotación COG y GO *terms*. Como referencia se emplearon los recuentos totales para los genomas completos de

Tabla M4. Variables ecológicas, genómicas, de homología, de recombinación, variables referentes a mecanismos de movilidad y barrera exploradas en el análisis del efecto de variables sobre la recombinación homóloga.

Clase de variable	Variable de estudio	Base de datos/origen	Test y análisis estadístico
Ecológicas	Modo de vida	JGI/IMG (Markowitz *et al*., 2012)	Kruskal-Wallis/Jonckheere-Terpstra
Anotación genomas y fragmentos recombinantes	COG KEGG GO	JGI/IMG (Markowitz *et al.,* 2012) JGI/IMG (Markowitz *et al*., 2012) Blas2GO (Gotz et al., 2008)	Enriquecimiento Fisher Enriquecimiento Fisher Enriquecimiento Fisher
Genómicas	Nº operones ribosómicos Nº Transposasas % Islas genómicas Tamaño genoma	rRNAdb (Klappenbach, *et al*; 2001) Archivo Genebank (JGI-IMG) Archivo Genebank (JGI-IMG) Archivo Genebank (JGI-IMG)	Kruskal-Wallis/Jonckheere-Terpstra Mantel-Haenszel
Barrera	Nº proteínas RM Nº Proteínas sistemas CRISPR-Cas	REBASE (Roberts *et al;* 2010). JGI/IMG (Markowitz *et al*., 2012)	Kruskal-Wallis/Jonckheere-Terpstra Mantel-Haenszel
Movilidad	% HGT Nº Transposones Nº Genes reparación Nº Proteínas conjugatuvas Competencia	JGI/IMG (Markowitz *et al*., 2012) (González *et al*; 2013) (Saheeed Inam *et al*; 2011)	Kruskal-Wallis/Jonckheere-Terpstra Mantel-Haenszel
Homología	*Core* genoma %Genoma recombinable	Alineamientos MAUVE (Darling AE., *et al* 2010) Alineamiento Nucmer MUMmer (S. Kurtz et al; 2004)	Kruskal-Wallis/Jonckheere-Terpstra Mantel-Haenszel
Recombinación	Nº Eventos recombinación/cepa % Genoma recombinado (ANIb)	Identificación RDP4 v4.15 (Martin *et al;* 2010)	Kruskal-Wallis/Jonckheere-Terpstra Mantel-Haenszel

cada una de las especies del estudio.

En el caso de la anotación en COG, los datos se procesaron con un *script* en Python construyendo las **tablas de contingencia 2x2**. Se realizaron test de Fisher para cada una de las categorías funcionales en cada especie aplicando una corrección FDR (*False Discovery Rate*). Se consideraron significativos p-valores menor s de 0,05.

Además se realizó una anotación de los genomas en **términos GO** (*GO terms*) mediante Blast2GO (B2G) (Conesa *et al;* 2008; Gotz *et al*; 2008) Se obtuvo la anotación en términos GO para las tres principales categorías funcionales: **Componentes Celulares** (CC), **Procesos Biológicos** (BP) y **Funciones Moleculares** (MF). Se anotaron las 54 especies mediante la función de anotación implementada en el programa empleando los archivos fasta de secuencias aminoacídicas de cada uno de los genes albergados. Se consideraron significativas a la hora de asignar anotación aquellos *hits* con un E-value de corte $1x10^{-20}$ y una identidad del 55% a nivel de secuencia aminoacídica. Además se recuperaron los términos GO (Ashburner *et al*; 2000) y los números EC de las vías de KEGG (Kanehisa y Goto 2000) para mejorar la asignación de términos GO. Para estimar qué funciones estaban sobrerrepresentadas o infrarrepresentadas en los fragmentos recombinantes se usó el paquete Gossip implementado en B2G. Este paquete emplea FT con **corrección FDR**. Se consideraron significativos p-valores inferiores a 0,05.

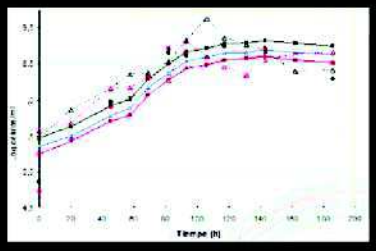
Análisis de varianza y correlaciones.

Se llevaron a cabo análisis de varianza de tipo **Kruskal-Wallis/Jonckheere-Terpstra** con el paquete estadístico SPSSv22, comparando la distribución de las variables descritas anteriormente en las 4 clases ecológicas definidas. Los análisis se acompañaron de comparaciones múltiples o ajustes de Bonferroni. Se analizó la distribución de variables de recombinación (porcentaje de genoma recombinado y eventos por cepa), tamaño de los fragmentos recombinados, genes codificantes de tRNAs, rRNAs, genes pertenecientes a los sistemas de modificación-restricción (RM) e islas genómicas.

Se emplearon correlaciones parciales y bivariadas para explorar relaciones entre variables cuantitativas genómicas, de motilidad, de recombinación, barrera y referidas a fragmentos

recombinantes, empleando el paquete estadístico SPSSv22. Se realizaron test paramétricos y no paramétricos considerando los coeficientes de correlación de Pearson y Spearman respectivamente para la significancia bilateral (marginalmente significativos $p<0,1^{\cdot}$, $p<0,05^{*}$, $p<0,01^{**}$, $p<0,001^{***}$).

Se aplicaron **tests de Mantel-Haenszel** y test parciales de Mantel-Haenszel a matrices de distancia euclídeas generadas desde las variables de recombinación (eventos/cepa y %genoma recombinado), variables de motilidad (competencia y RM), variables barrera (CRISPR-Cas y HGT), filogenia, genotipo y clase fenotípica. Se consideraron significativos p valores menores de 0,05. Se construyó un modelo general para los factores involucrados en la distribución de la variable recombinación homóloga a partir de los resultados de los test de Mantel-Haenszel por medio de un ***path analysis***.



Introducción

Objetivos

Materiales y métodos

Resultados y discusión

Capítulo 1

Análisis de las diferencias transcripcionales e interacción de cepas cercanas de *S. ruber* mediante RNAseq.

Capítulo 2

Estudio de los mecanismos y estrategias de diversificación genómica en *S. ruber*

Capítulo 3

Impacto de la recombinación homóloga sobre la evolución de genomas *core* procariotas

Conclusiones

Bibliografía

Anexos

Resumen

Los estudios genómicos comparativos, metagenómicos y basados en SCG muestran que las poblaciones de especies bacterianas engloban cepas estrechamente relacionadas. Esto plantea la cuestión de si células de determinadas cepas responden a la presencia de células de otras cepas de su mismo linaje, modificando sus niveles de expresión, o por el contrario la expresión del conjunto es la simple suma aritmética de sus componentes individuales. En este capítulo se emplea la versatilidad y precisión del RNAseq para abordar esta cuestión, mediante el análisis de cultivos puros y mixtos de las cepas M8 y M31 de *Salinibacter ruber* en fase exponencial.

En general, los patrones transcripcionales de las cepas M8 y M31 en cultivo puro fueron muy parecidos entre sí, detectándose niveles de expresión significativos, aunque en muchos casos bajos, en el 98% de los genes, prácticamente a lo largo de todo el genoma. Un subconjunto de 64 genes presentaron niveles de expresión muy elevados en ambas cepas. Entre ellos encontramos reguladores transcripcionales, destacando la presencia de factores sigma alternativos; genes relacionados con la elevada tasa metabólica en crecimiento exponencial como proteínas ribosómicas o chaperonas; genes implicados en respuesta a estrés, y el codificante para la xantorrodopsina. Tras comparar los niveles de expresión de ambas cepas, se encontraron diferencias significativas en 165 genes, relacionados con respuestas a estímulos ambientales.

Cuando se compararon los perfiles transcripcionales de las cepas al pasar de cultivo puro a mixto, un total de 354 y 446 genes presentaron expresión diferencial en M8 y M31, respectivamente, destacando la presencia de genes pertenecientes a sistemas de dos componentes. El genoma *core* mostró más cambios que el accesorio y la respuesta de cada cepa fue específica. Para M31 se apreció un aumento de expresión de genes relacionados con el crecimiento exponencial mientras que en M8 este aumento se dio en genes relacionados con respuesta a estrés y producción de antibióticos, entre otros. Estos resultados ponen de manifiesto que diferencias sutiles entre cepas a nivel genómico, como las que existen entre M8 y M31, pueden contribuir notablemente a la microdiversidad funcional de la especie. Dada la elevada diversidad metabólica previamente descrita en la fracción extracelular de cultivos de *S.ruber*, nuestros datos sugieren que compuestos como antibióticos u otras moléculas señal podrían estar implicados en la comunicación entre las cepas de esta especie.

1. Introducción

Las poblaciones de especies bacterianas en ambientes naturales incluyen una gran diversidad de cepas muy próximas con una heterogeneidad y microdiversidad génica extensas, tal como muestran estudios de genómica comparativa con cepas aisladas (Caro-Quintero *et al*., 2009), análisis metagenómicos (Luo *et al*., 2011; Caro-Quintero *et al*., 2011; Konstantinidis y De Long., 2008; Caro-Quintero y Konstantinidis 2012) y, más recientemente, estudios de SCG (Kasthan *et al*., 2014). Sin embargo los análisis genómicos de cepas aisladas no revelan de manera clara el significado ecológico y funcional de estas variaciones, más si cabe cuando pequeñas variaciones génicas pueden conducir a diferentes estrategias ecológicas (Denef *et al*., 2009; Wilmes 2011) y cuando la proporción de estudios con organismos de relevancia ecológica, en especial coaislados, es bastante menor que la de estudios con patógenos.

El estudio de una extensa colección de aislados de *S.ruber* de todo el mundo y análisis metagenómicos de ambientes hipersalinos han puesto de manifiesto que esta especie presenta una elevada homogeneidad filogenética a la par que una enorme microdiversidad, incluso para cepas que coexisten (Peña *et al*., 2010; Pasic *et al*., 2009). El análisis genómico comparativo completo entre las dos cepas de *S.ruber* coaisladas más próximas filogenéticamente, M8 y M31, permitió establecer los niveles de microdiversidad génica existente entre ambas (Peña *et al*., 2010). Ambas cepas presentan además diferentes perfiles metabólicos, que podrían atribuirse a las diferencias génicas observadas, así como diferencias fisiológicas tales como una distinta susceptibilidad a fagos. En conjunto estos resultados muestran que *S.ruber* constituye un ejemplo de microdiversidad funcional y genómica, e indican que las diferencias genómicas observadas podrían no ser neutrales desde una perspectiva ecológica (Peña *et al*., 2010).

En el primer capítulo de esta tesis se emplea la tecnología RNAseq para explorar en detalle las diferencias a nivel transcripcional entre las cepas M8 y M31 crecidas tanto en cultivo puro como en co-cultivo. La comparación entre transcriptomas en cultivo puro puede mostrar una relación directa del impacto de las diferencias en las secuencias y arquitectura genómicas entre ambas cepas contribuyendo a entender sus implicaciones ecológicas (Yoder-Himes *et al*., 2009; Scaria *et al*., 2013; Voigt *et al*., 2014; Kimes *et al*., 2014). De hecho, trabajos previos

basados en la observación de patrones de expresión específicos de cepa sugieren que las diferencias a nivel de mecanismos reguladores pueden tener un papel relevante en la divergencia y diversificación microbianas (Wilmes, 2011; Yoder-Himes *et al*., 2009). En este sentido las diferencias de expresión génicas son probablemente la primera manifestación de divergencia ya que los elementos regulatorios evolucionan más rápido que los elementos que controlan (Vicente y Mingorance, 2008).

Hasta la fecha se han realizado algunos análisis transcriptómicos intraespecíficos empleando RNAseq, en los que se analizan las diferencias transcripcionales de cepas cercanas:

i) El estudio de la respuesta específica de dos cepas de *Burkholderia cenocepacia* en diferentes nichos (Yoder-Himes *et al*., 2009).
ii) La caracterización de la respuesta individual de cepas coaisladas de la bacteria marina *Alternomonas macleodii* en diferentes condiciones de crecimiento (Kimes et al., 2014).
iii) La identificación de diferencias en los patrones y arquitectura transcripcionales (RNAs no codificantes y sitios de inicio de transcripción) de dos cepas de la cianobacteria marina *Prochlorococcus* (Voigt *et al*., 2014).
iv) La identificación de genes controlados por sistemas *quorum sensing* en diversas cepas de *Pseudomonas aeruginosa* (Chugani *et al*., 2012);
v) La caracterización de diferencias transcripcionales entre cepas hipervirulentas históricas y recientemente emergentes del patógeno entérico *Clostridium difficile* (Scaria *et al*., 2013).

Sin embargo, dada la enorme microdiversidad observada en cepas coaisladas de microorganismos de vida libre (Peña *et al*., 2014, Kasthan *et al*., 2014, Kimes *et al*., 2014, Salter *et al*., 2015, Takemura *et al*., 2014), resulta improbable que un linaje individual de *S. ruber* o de cualquiera otra especie *(Prochlorococcus*, *Alteromonas*, *Pelagibacter*, *Vibrio…)* crezca de manera aislada en la naturaleza y sin interactuar con otros. Por lo tanto, a la hora de interpretar los datos inferidos de cultivos puros, como los presentados en los trabajos mencionados, ha de considerarse este factor. Desde una perspectiva ecológica es por tanto interesante conocer si a

nivel funcional cepas que coexisten en la naturaleza actúan como la adición individual de cada una de ellas o si en presencia de otras cercanas modifican sus actividades. El objetivo principal de este capítulo es dar respuesta a esta pregunta analizando si existen diferencias de expresión entre cultivos puros y co-cultivos de las cepas M8 y M31 comparando sus perfiles transcripcionales mediante RNAseq.

2. Monitorización de los cultivos puros y mixtos de las cepas M8 y M31.

En este estudio seleccionamos las cepas M8 y M31 de *S.ruber* para su análisis transcriptómico empleando la tecnología RNAseq por su elevada resolución. Tanto los cultivos puros como mixtos de ambas cepas presentaron una elevada reproducibilidad en sus medidas de OD600, recuentos celulares y recuentos mediante qPCR (**figura C1.1; anexo, tablas S1.1**). Las tasas de crecimiento de ambas cepas en cultivo puro fueron similares (0.053±0.002 and 0.055±0.003 horass^{-1} para M8 y M31, respectivamente).

Cuantificación en cultivos mixtos y diseño de cebadores específicos.

Debido a que ambas cepas presentan una identidad del 100% a nivel del gen rRNA16S, hecho que impide el diseño de sondas específicas para FISH, se diseñaron cebadores específicos para poder cuantificar la abundancia relativa de cada una de ellas en co-cultivo mediante qPCR. Mediante el uso del *software Primer-Blast* (Yu *et al*., 2012) se obtuvieron de manera automática varios cebadores empleando las secuencias de genes específicos de las cepas M8 y M31 (**anexo, tabla 2M**). Se llevó a cabo una primera selección considerando su secuencia y posteriormente, con el software *Net Primer* (Premier Biosoft International), se realizó una caracterización de cada pareja de cebadores descartando aquellos que presentaban características termodinámicas que favoreciesen la formación de estructuras secundarias. Usando estos criterios se seleccionaron 7 parejas (**anexo, tabla S1.2**).

Una vez diseñados se evaluó la calidad de los cebadores mediante PCR. El óptimo de temperatura de hibridación se determinó mediante PCR en gradiente y la eficiencia se evaluó

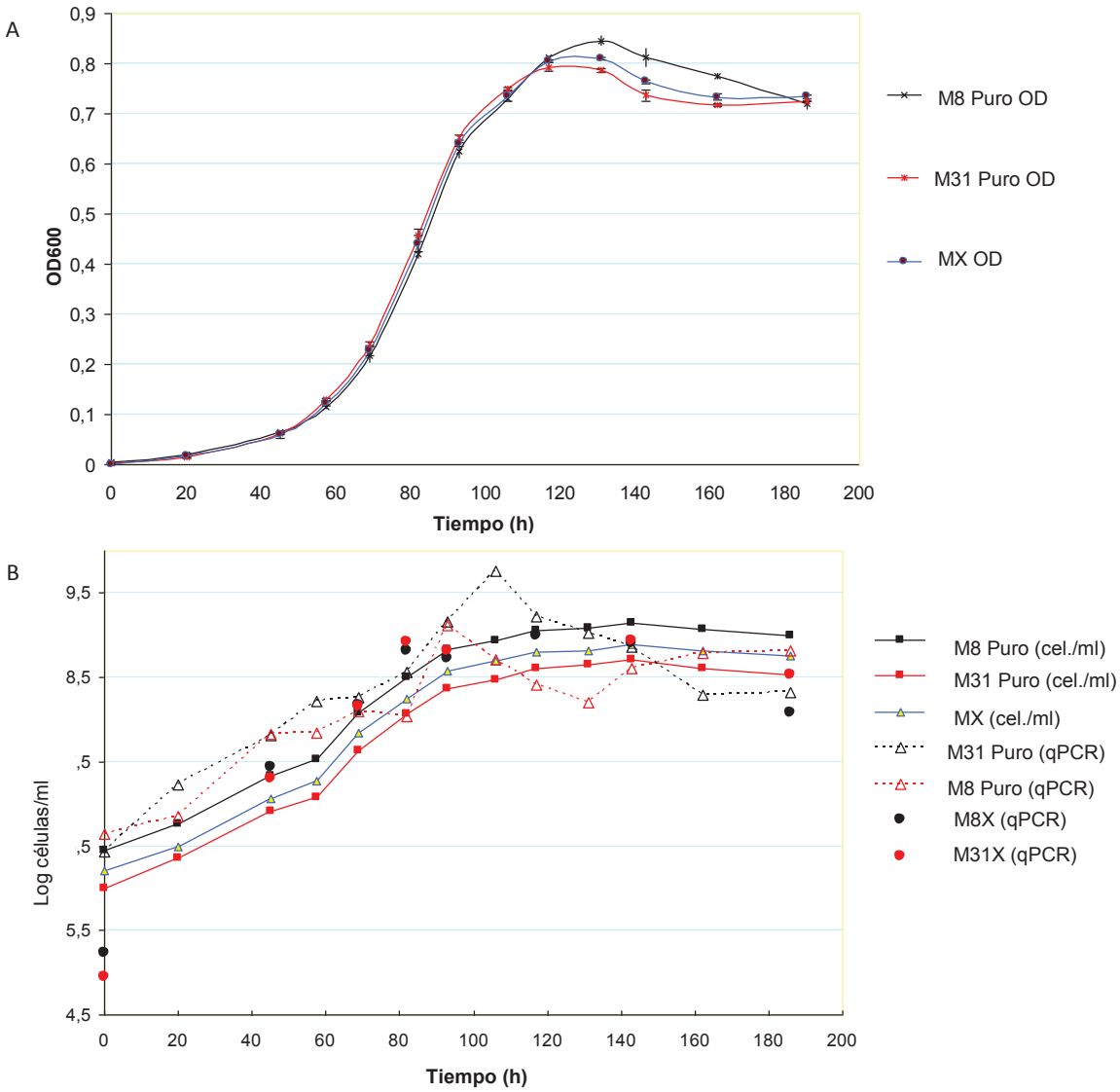


Figura C1.1. Curvas de crecimiento y de los cultivos puros (M8 en azul y M31 en rojo) y mixtos (negro). Figura A: Valores de OD_{600} promedio para los triplicados de cada tipo de cultivo tomados en los 13 puntos de monitorización. Figura B: Valores de densidad celular total (células/ml) obtenidos por DAPI (líneas continuas) y qPCR (líneas discontinuas para cultivos puros). En los cultivos mixtos (MX), la densidad celular de cada cepa obtenida por qPCR se representa mediante círculos (M8X en negro; M31X en rojo).

mediante qPCR. Para concluir la validación, se comprobó que las dos parejas de cebadores seleccionadas (1109_4F/R para M31 y 308_2F/R para M8) presentaban una eficiencia similar a la de los cebadores 338F/500R, obteniéndose datos cuantitativos acordes. Estos últimos, diseñados con anterioridad para amplificar un fragmento del operón ribosómico del 16S, ya mostraron empíricamente una elevada eficiencia.

Extracción de RNA y aislamiento de mRNA para su secuenciación por RNAseq.

Se seleccionó un punto en mitad de la fase exponencial (T5, tras 82 horas de crecimiento), **(tabla S1.1**), para su posterior análisis transcriptómico. En este punto la cantidad de células de ambas cepas en cultivo puro fue idéntica, mientras que la proporción de células en cultivos mixtos, determinada mediante qPCR, mostró una relación de 8:10 a favor de M31, acorde con las cantidades de mRNA obtenidas (**tabla C1.1**).

Una vez seleccionado el punto se extrajo el RNA total y, puesto que el kit de extracción y purificación de RNA no garantiza la ausencia de DNA en el extracto final (observable en el gel de electroforesis a la altura de las 23 kb), los extractos se digirieron una DNasa. En el anexo **figura S1.1,** se muestra un gel de electroforesis en agarosa al 0.8% donde se cargaron 2µl del extracto antes y después de la eliminación del DNA genómico. La calidad del RNA total extraído se evaluó mediante bioanalizador (**figura C1.2**). En esta figura los picos mayores corres ponden al rRNA 16S y 23S. Entre ambos se establece la ratio RIN (del inglés *RNA integrity number*) indicativa del estado de degradación del RNA (Schroeder *et al*., 2006). Valores de RIN por encima de 7 son propios de RNAs de buena calidad, empleándose idealmente muestras de RNA con un RIN al menos de 8, como es el caso de nuestras muestras, para secuenciación masiva de RNA (Sigurgeirsson *et al*., 2014).

A continuación se aisló el mRNA de un cultivo puro de cada capa y dos cultivos mixtos eliminando los rRNAs 16S y 23S con el kit MICROBE Express (Ambion) y el RNA 5S y tRNAs con el kir MEGA Clear (Ambion). La eficiencia en la eliminación de los rRNAs 16S y 23S (1ª fase) y del rRNA 5S y los tRNAs (2ª fase) se comprobó mediante bioanalizador (figura C1.1B).

Como puede observarse, el tratamiento de los rRNAs y tRNAs llevó a la eliminación de entre un 86% y un 95% del RNA total extraído. No obstante, en los cromatogramas obtenidos tras las dos fases de eliminación se puede observar la drástica disminución en la cantidad de rRNAs ribosómicos 16S y 23S (ver la altura de sus picos con respecto a la altura del *ladder*, representado por la barra de color naranja) así como la eliminación del rRNA 5S y tRNAs, señalados en los cromatogramas, excepto en la última fase de purificación, por un círculo negro.

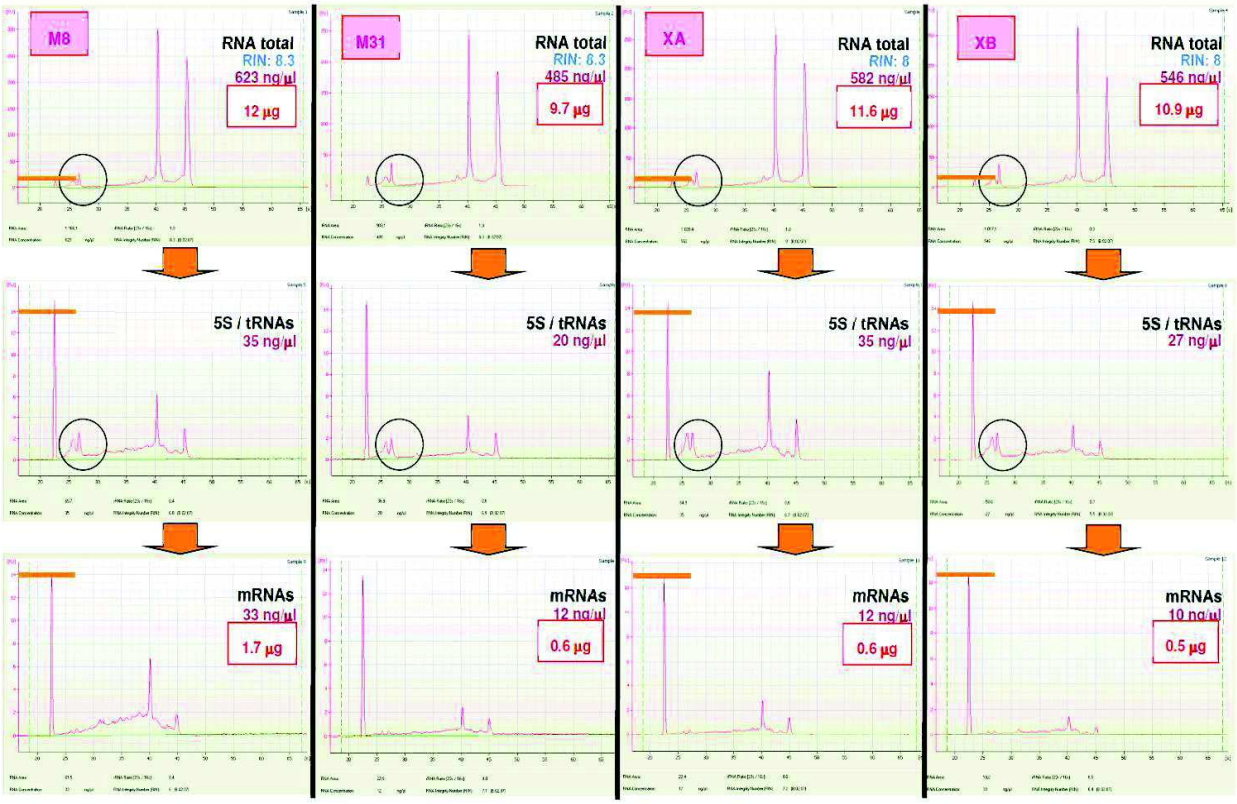


Figura C1.2. Cromatogramas del RNA obtenido tras las sucesivas fases de aislamiento de mRNA: RNA total (arriba), tras la fase1º de eliminaciónde rRNAs 16S y 23S (centro) y tras la eliminación de los tRNAs y rRNA 5S (abajo). Se muestra la concentración de RNA en los extractos y la cantidad de RNA total extraído para los cultivos puros (M8 y M31) y mixtos (XA y XB). El círculo negro indica la región del cromatograma donde se localizan los tRNAs y el rRNA 5S.La barra naranja indica la altura del marcador de tamaño que marca el comienzo del área del cromatograma que contiene el RNA.

Tabla C1.1- Resumen de los resultados de secuenciación con la plataforma Illumina para las muestras de *S.ruber*.

Muestra[a]	Pares de *reads* procesados	*Reads* filtrados	*Reads* filtrados (%)	*Reads* de rRNA	*Reads* de rRNA (%)	*Reads* finales útiles	Kb secuenciadas[b]	Cobertura útil[b]	*Reads* finales útiles (%)[b]	Copias qPCR[b]
M8-P	43071950	7088806	16,45	25838764	65,45	10144380	507219	132,33X	14.09	$3.63x10^8$
M31-P	42924442	6631194	15,44	25765713	70,05	10527535	526377	146,73X	14.5	$1.08x10^8$
MXA	39918009	54445077	13,64	25087962	72,74	M31: 5044060	M31: 252203	M31: 70,30X	M31: 7,32	406,16
						M8: 4340910	M8: 217046	M8: 56,62X	M8: 6,30	375,49 **
MXB	40131308	5433881	13,54	24885733	72.33	M31: 5549600	M31: 277480	M31: 77,35X	M31: 8,00	219,68
						M8: 4262094	M8: 213105	M8: 55,60X	M8: 6,14	309,12 **

[a]M8-P y M31-P: cepas M8 y M31 en cultivo puro, respectivamente. MXA y MXB: réplicas de los cultivos mixtos.
[ub]Contribución de cada cepa en cultivo mixto.
** Copias detectadas por Qpcr. Copias en 5µl de muestra eluída una vez hecha la extracción de RNA y tras la digestión de DNA.

Las cantidades de RNA total tras la última fase de purificación (entre 500 ng y 1,7 µg) estarán por tanto, muy enriquecidas en mRNAs. Tal como muestran los resultados de RNAseq tras el mapeo de secuencias leídas (*reads*) (**tabla C1.1**). Los porcentajes de rRNA obtenidos, alrededor del 70%, fueron menores a los observados en otros estudios con cepas cercanas como el llevado a cabo en *B. cenocepacia* (Yoder-Himes *et al*., 2009) o *A. macleodii* (Kimes *et al*., 2014), en donde más de un 90% de las secuencias mapearon en los operones ribosómicos en la mayoría de las muestras analizadas. Una proporción elevada de tRNAs (11/43) en M8 y (17/44) en M31 no presentaron expresión en cultivos puros, lo cual refleja la elevada eficiencia del proceso de eliminación de tRNAs, considerando además sus elevados niveles de transcripción.

3. Secuenciación, tratamiento de datos y validación.

Una vez secuenciados, y tras el filtrado de *reads*, los resultados obtenidos fueron similares para las cuatro muestras en términos de cobertura y calidad (**tabla C1.1**). La cobertura en todos los casos fue elevada y similar en las muestras secuenciadas, 10 veces mayores a las obtenidas en estudios previos en cepas cercanas de especies como *A. macleodii* (Kimes *et al*., 2014) o *B. cenocepacia* (Yoder-Himes *et al*., 2009). Además de un importante enriquecimiento en mRNAs, los porcentajes finales mapeados para las muestras secuenciadas de los cultivo puros y mixtos de *S.ruber* fueron similares entre sí (entre el 13,62% y 14,14%). Esto favorece la comparación de los niveles de expresión entre ellas y contribuye a la detección incluso de genes con niveles de expresión muy bajos tal como muestran trabajos realizados con diferentes cepas de *P. aeruginosa* (Chugani *et al*., 2012), en donde con un menor esfuerzo de secuenciación se detectó expresión en más del 90% de los genes del genoma. Como resultado se detectaron niveles de expresión significativos en más del 98% de los genes en ambas cepas.

Validación de los resultados.

Los datos de expresión para cada uno de los genes en las réplicas biológicas de los cultivos mixtos, MXA -B, presentaron muy buenas correlaciones (R=0,99 y R=1 para M8 y M31

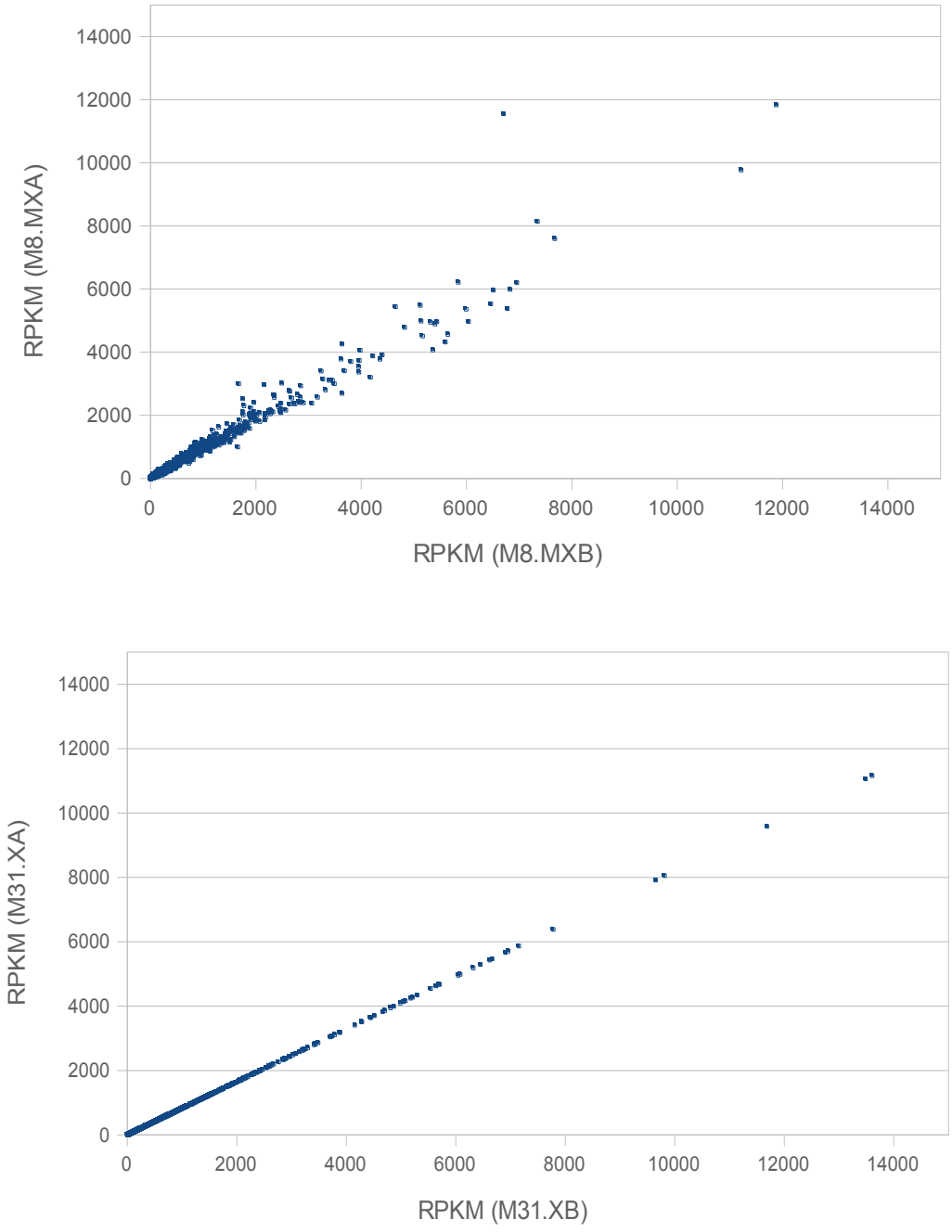


Figura C1.3. Reproducibilidad entre replicas biológicas de cultivos mixtos (MXA y MXB). Una vez asignadas las secuencias a cada cepa, se correlacionaron los niveles de expresión en M8 (figura superior), con R=0,9956 y M31 (figura inferior) con R=1.

respectivamente) (**figura C1.3**). Los coeficientes de correlación obtenidos fueron mayores que los mostrados en trabajos de RNAseq previos (0.95-0.98 para Li *et al*., 2014; 0.93-0.95 en Nagalakshimi *et al*, 2008; 0.947-0.977 en Scaria *et al*., 2013). Las relaciones de *reads* de RNAseq asignados mediante mapeo a cada una de las cepas (reads de M8/reads de M31, 0,86 para MXA y 0,70 para MXB) mayor por lo tanto en M31 en ambas réplicas (tabla C1.1). Como validación de los datos de RNAseq obtenidos, se obtuvo la abundancia relativa de cada una de las cepas en cultivo mixto mediante qPCR. En el caso de los cultivos mixtos la abundancia relativa de cada cepa, relación M8/M31, obtenida mediante qPCR fue de 0,92 y 0,70 para las réplicas MXA y MXB respectivamente (**tabla C.1.1**).

En el caso de los cultivos puros se realizó un mapeo cruzado de los *reads* de cada cepa contra el genoma contrario, es decir, los *reads* de M8 contra el genoma de M31 y viceversa. Este mapeo proporcionó el número de falsos positivos, entendiendo como tales todas aquellas secuencias procedentes de cultivo puro de una cepa que mapearon con mayor puntuación con el genoma de referencia de la otra cepa. La cifra fue inferior al 0.03% de las secuencias alineadas lo cual indica una alta pureza de estos cultivos y una alta calidad de las secuencias que pasaron el filtrado tras el proceso de secuenciación.

4. Análisis transcriptómico de cultivos puros de *S.ruber* M8 y M31.

Perfiles de expresión generales en los cultivos puros de M8 y M31.

Los análisis de expresión en cultivo puro mostraron que más del 98% de los genes se transcribieron en ambas cepas, valores superiores a los mostrados en otros estudios transcriptómicos de RNAseq en bacteria en los cuales también se determinó una elevada proporción de genes expresados: 64% en *Acidovorax avenae* (Li *et al*., 2014), 83% en *Listeria monocitogenes* (Pinto *et al*., 2011), 89% en *P. aeruginosa* (Chugani *et al*., 2008). Estas diferencias apreciadas entre diferentes transcriptomas pueden deberse tanto a las distintas profundidades de secuenciación (*sequencing depth*) empleadas, así como al umbral de expresión considerado, más si cabe cuando los niveles de transcripción para genes individuales pueden

diferir en varios ordenes de magnitud (Filiatrault *et al*., 2011). El criterio considerado a la hora de establecer un umbral de expresión resulta algo arbitrario. En este caso consideramos que presentaban expresión estadísticamente significativa aquellos genes con valores mayores a 0 RPKM. Tan sólo 55 y 30 genes (anexos, **tablas S1.3 y S1.4**) no presentaron niveles de expresión significativos en cultivo puro para las cepas M8 y M31 respectivamente. Entre estos genes, un elevado porcentaje codificaron para proteínas hipotéticas y específicas de cepa (87% y 67% en M8 y M31 respectivamente), y un 36% y 50% mapearon en HRVs y plásmidos. Estos datos sugieren que se trata de genes de reciente adquisición ya que además, presentan valores de GC y CAI inferiores al promedio. Estos resultados reflejan la dinámica de transferencia horizontal de genes no homólogos en las HRVs tal como se discutirá más adelante al analizar la dinámica génica, procesos de duplicación y la relación de los procesos de adaptación con los niveles de expresión observados.

En términos generales ambas cepas mostraron perfiles de expresión similares a lo largo del cromosoma y tal como se muestra en la **figura C1.4**. En esta misma figura es posible apreciar como la mayoría de genes (80%) se expresaron por debajo de las 200 RPKM y un elevado porcentaje por debajo de las 1000 RPKM (97%). En total se identificaron 64 y 65 genes en M8 y M31 altamente expresados (VHGE, del inglés *very higly expressed genes*), con niveles de expresión superiores a las 1000RPKM (**tabla S1.5).** El 68% de los VHEG presentaron un ortólogo altamente expresado en la otra cepa. Muchas de las proteínas codificadas por VHEG en ambos genomas presentaron relaciones funcionales, pudiendo agruparlas en proteínas relacionadas con la alta demanda de crecimiento en fase exponencial, reguladores transcripcionales y reguladores traduccionales, algunos de ellos vinculados a mecanismos de respuesta a estrés. Esto junto al hecho de que muchos de los VHGE se localicen en regiones cromosomales cercanas lleva a pensar que se trate de genes co-rregulados, ya sea por pertenecer a un mismo regulón o a un operón (**figura C1.5**). En el caso de las proteínas ribosómicas **(anexo, tabla S1.6),** sus niveles de expresión fueron muy similares, entre las 500-1200 RPKM, acordes con la relación estequiométrica presentada en la estructura ribosómica. Otro ejemplo de genes con elevados niveles de expresión son los situados en el extremo 5’ de la HRVII de ambos genomas como se comentará más adelante (anexo, **figura C1.9**).

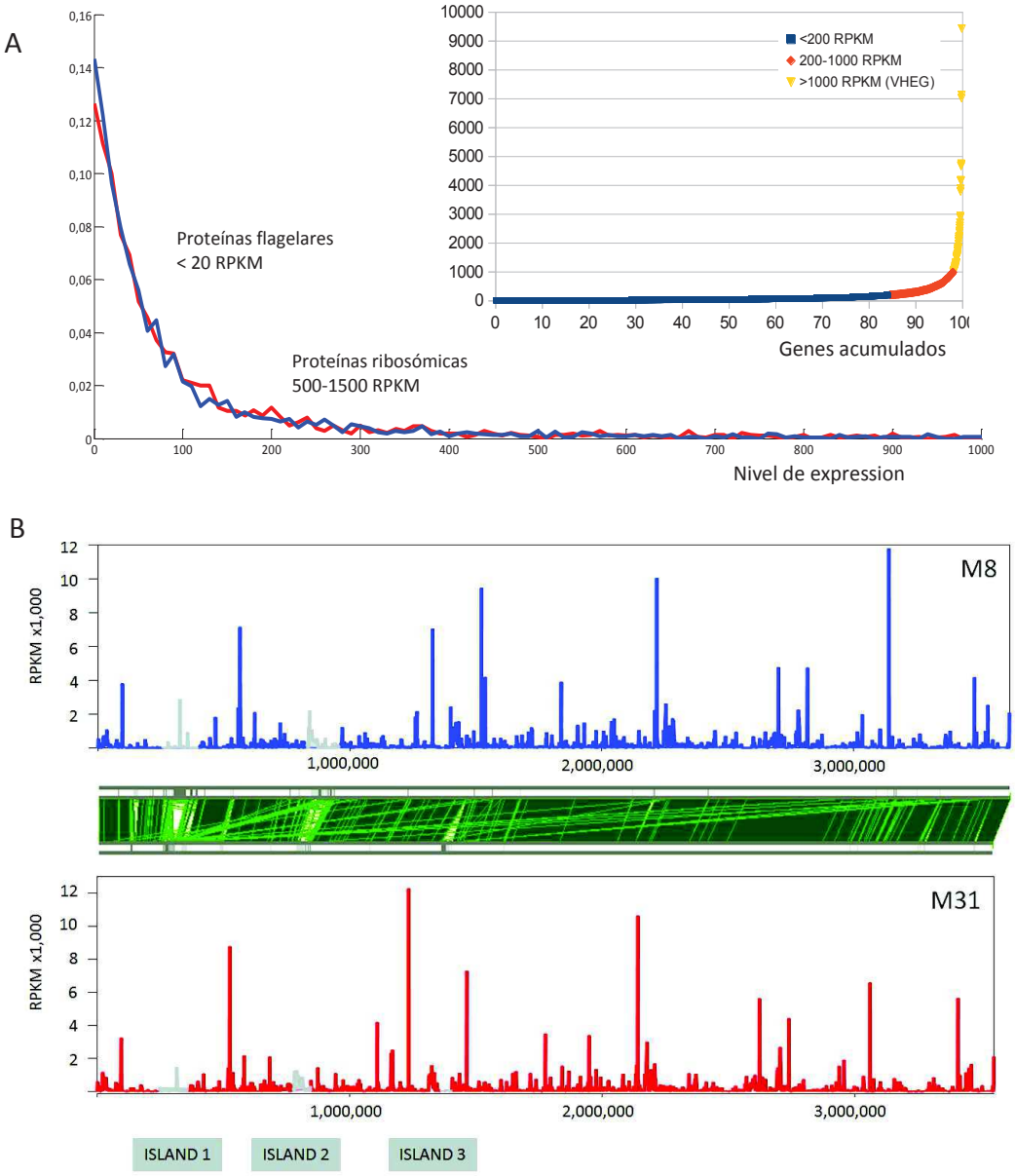
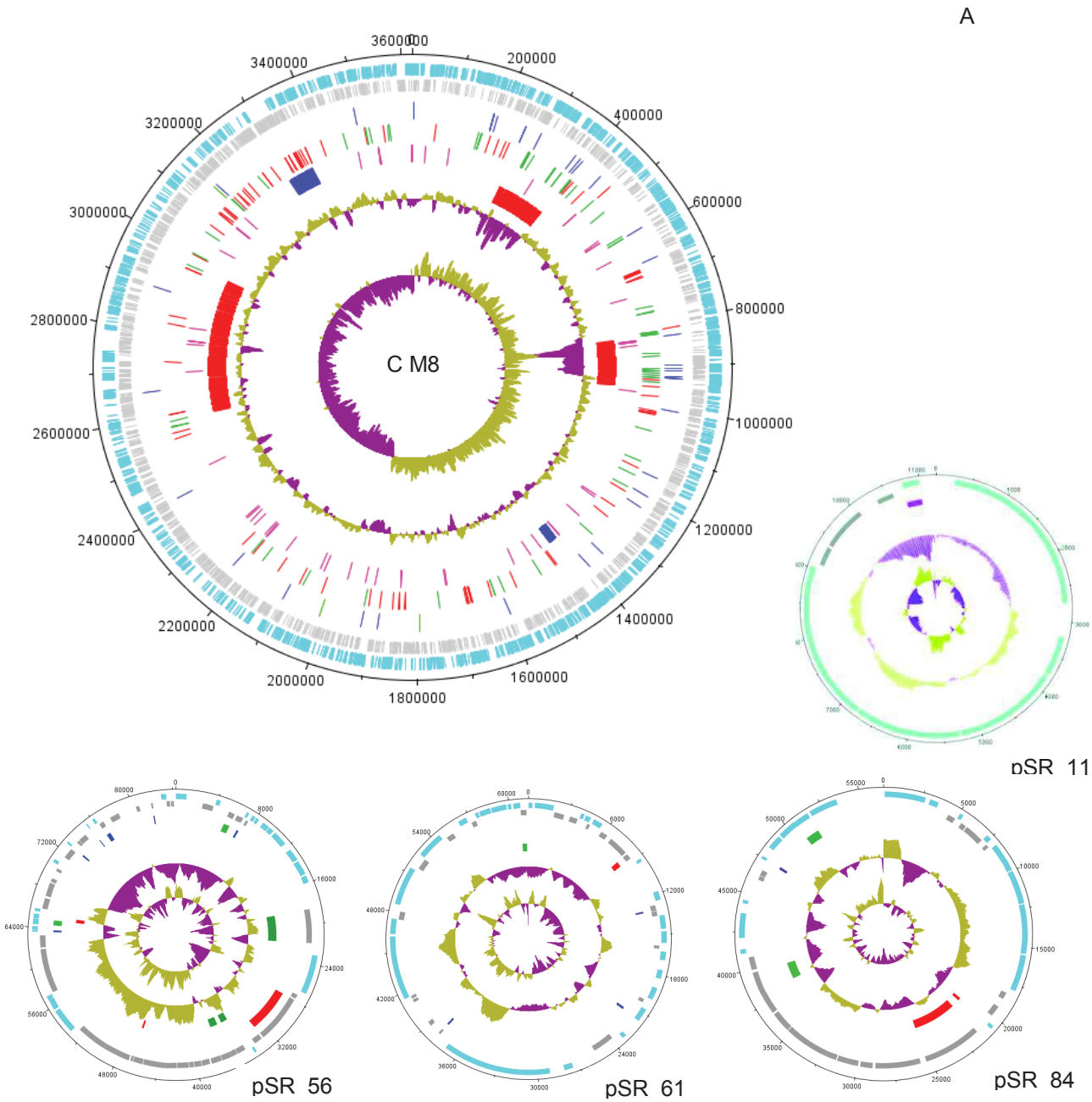


Figura C1.4. Perfiles de expresión de las cepas M8 y M31 de *S.ruber* en cultivo puro.
Figura A: Comparación de las abundancia de genes en base a su nivel de expresión M8 (azul) y M31 (rojo); porcentaje de genes acumulados para niveles menores de 200 RPKM (azul), 200-1000 RPKM (naranja) VHEG (>1000 RPKM).
Figura B: Perfiles de expresión de M8 (azul) y M31 (rojo), destacando los valores de expresión de las HRVs (gris).


A
C M8
pSR 11
pSR 56
pSR 61
pSR 84

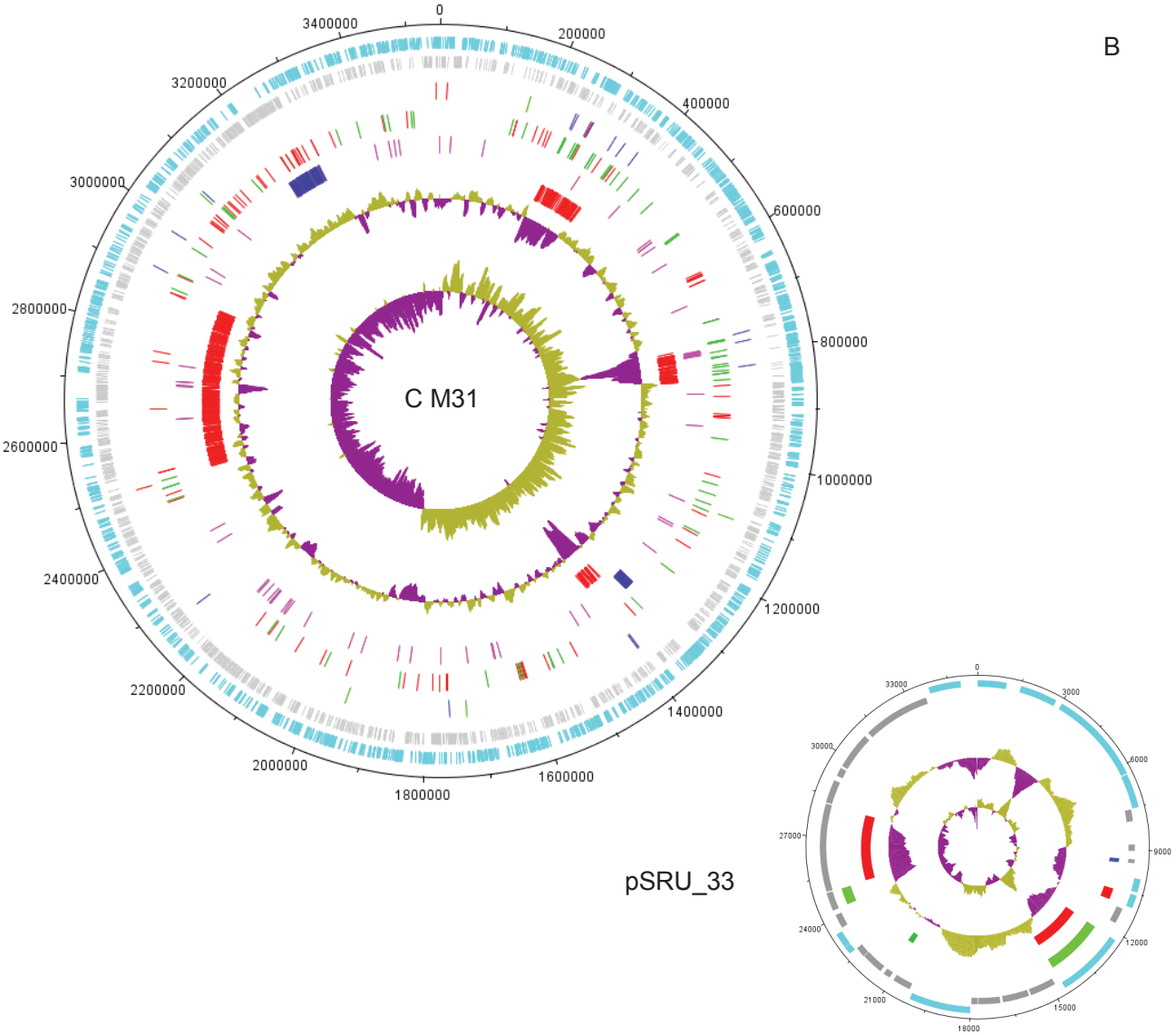


Figura C1.5. Principales eventos de expresión de los genomas de M8 (figura A) y M31 (figura B). De fuera hacia dentro en los cromosomas (CM31, CM8): Anillo1: genes de la cadena F (azul); Anillo2: Genes de la cadena R (gris); Anillo3: Genes no expresados específicos de cepa (azul), con ortólogo en la otra cepa (rojo), no expresado en ambas (verde); Anillo 4: 165 genes con expresión diferencial entre cultivos puros, y mayor expresión en M8(rojo) M31 (verde); Anillo5: VHEG (violeta), cluster genes ribosómicos/flagelo (azul); Anillo6: Posición de las HRVs y CR (rojo); Anillo 7: %GC; Anillo 8: GC Skew. En los plásmidos (pSRU33; pSRM11; Psrm56; pSRM61; pSRM84): Anillos 1; 2; 3 y 4 (idéntico al cromosoma); anillo 5: genes expresados diferencialmente, sobreexpresados (verde), reprimidos (rojo); anillo 6: %GC; anillo7: GC skew (GC Skew = (G - C)/(G + C)).

Dentro del grupo de los VHEG que codifican para proteínas relacionadas con la alta demanda en fase exponencial encontramos proteínas ribosómicas, transportadores TonB, superóxido dismutasa, ATPasas, subunidades de la RNA polimerasa, citocromo C oxidasa y chaperonas. Destaca la expresión elevada del gen codificante para la xantorrodopsina (SRM_01698/SRU_1500 en M8/M31 respectivamente). Esta proteína actúa como bomba de protones activada por luz generando un gradiente protón motriz empleando la salixantina como pigmento antena (Balashov *et al*., 2005).

Entre las proteínas implicadas en respuesta a estrés encontramos diversos factores transcripcionales y traduccionales. Entre ellos encontramos las 3 ORFs que codifican para la proteína de respuesta a estrés térmico análoga a *csp*C (SRM_01159/SRU_0966; SRM_01933/SRU_1727; SRM_02671; SRU_2449) y *csp*G, este último gen específico de M31 (SRU_1563). Esta proteína pertenece a la familia de proteínas de respuesta a estrés CspA (COG1278; K03704: factores de transcripción). Las proteínas CSP actúan como reguladores traduccionales uniéndose a moléculas de mRNA regulando su tasa de degradación, la fase de terminación de transcripción y la unión de los ribosomas (Dominy *et al*., 2002). En *C. difficile* se ha detectado la sobreexpresión de proteínas CSPs tras una situación de shock osmótico (Scaria *et al*., 2013). Dentro de esta respuesta a estrés se observaron valores de expresión elevados de algunos factores sigma y del gen *asn*C (SRM_0384/SRM_00452) que codifica para el regulador transcripcional asnC, que controla el metabolismo de ectoínas en *Halomonas* (Schwibbert *et al.,* 2011). Además se apreció la expresión elevada de genes implicados en regulación transcripcional tales como ncRNA riboswitches, factores sigma alternativos con función extracitoplasmática (ECF, del inglés *extracitoplasmatic function*), factores sigma flagelares y del citrato férrico. Algunos de ellos presentaron valores de expresión mayores que los del estándar sigma 70, que también presentó valores elevados. Dentro de cada uno de seis tipos de factores sigma codificados en los genomas de *S.ruber* observamos patrones de expresión similares en ambas cepas (anexo, **tabla S1.7**), siendo el factor ECF sigma E (sigma 24) (SRM_02973/SRU_2762), el que mostró los niveles de expresión más elevados. Este factor sigma pertenece a la familia ECF01, descrita en las bases de datos MiST y ECF *finder* como especializada en respuesta a estrés periplásmico y shock térmico. En conjunto, los niveles elevados de expresión para

diferentes genes involucrados en procesos de estrés ambiental podrían indicar que las condiciones de cultivo consideradas como ”estándar” en realidad sean estresantes para *S. ruber*.

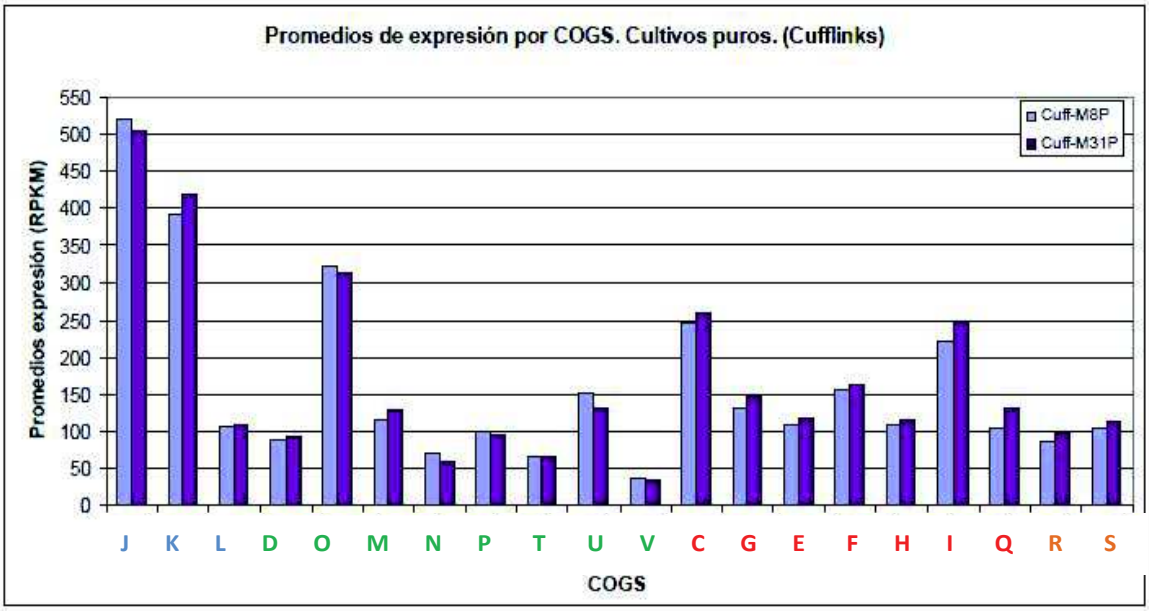
El análisis de los perfiles de transcripción de los genes de las rutas del KEGG (Kanehisa y Goto 1999) mostró para algunas de ellas varios genes con niveles de expresión elevados, (entre 200-100RPKM, anexo, **tabla S1.8**). Entre estas vías, marcadas en la tabla S1.8, destacan las implicadas directamente en la producción y conversión energética, el metabolismo de carbohidratos (glucólisis/gluconeogénesis, CAT y pentosas fosfato), metabolismo de nuecleótidos, de ácidos grasos y de algunos aminoácidos. Entre las categorías COG mas expresadas tenemos: J (traducción, estructura ribosómica y biogénesis), K (transcripción), O (modificaciones postraduccionales, reciclado de proteínas y chaperonas), M (envueltas celulares), U (transporte vesicular), C (producción y conversión de energía), F (transporte y metabolismo nucleotídico) e I (metabolismo de lípidos) (**figura C1.6**). Por lo tanto, se obtuvieron niveles elevados de expresión para todas las proteínas implicadas en vías del metabolismo central o funciones esenciales como la glucólisis.

En el caso de la glucólisis, se detectó la expresión de todos los genes codificantes para las enzimas implicadas en la vía Embden-Meyerhof, incluyendo la fructosa-1,6-bifosfato aldolasa cuya actividad no había sido medida previamente (Oren, 2013). La vía glucolítica alternativa, Entner Doudoroff mostró niveles de expresión decrecientes desde la primera reacción, corroborando la ausencia de actividad 2-ceto-3-desoxy-6-fosfoglucanato aldolasa descrita en publicaciones previas (Oren y Mana., 2003).

Arquitectura de los perfiles de transcripción.

Para cada uno de los genomas se obtuvieron los distintos perfiles de expresión de sus replicones, las diferencias significativas al compararlos y los patrones de distribución particulares (**figura C1.7**). En términos generales se apreciaron niveles de expresión bajos en plásmidos y HRVs **(figura C1.8**), que contuvieron una representación significativa de genes no expresados**.** En concreto un 40% (22/55) y un 47% (14/30) de los genes no expresados en M8 y M31 se localizan en estos elementos (anexos, **tablas S1.3 y S1.4).**

Estos datos, así como el hecho de que *S. ruber* presenta una elevada microdiversidad plasmídica (Peña, datos no publicados), sugieren que podría haber una elevada frecuencia de adquisición y pérdida de plásmidos y genes individuales tanto en plásmidos como en HRVs. Los datos obtenidos para M8 y M31, cepas muy próximas filogenéticamente y coaisladas, indican que gran parte del contenido génico presente en los plásmidos e islas genómicas de la especie



Información almacenamiento y procesamiento.
J: Traducción, estructuras ribosómicas y biogénesis
K: Transcripción.
L: Replicación, recombinación y reparación de DNA.

Procesos celulares.
D: División celular y cromosómica.
O: Modificaciones postraducionales, reciclado proteico.
M: Biosíntesis de envueltas celulares, membrana externa.
N: Motilidad celular y secreción.
P: Transporte de iones inorgánicos y metabolismo.
T: Mecanismos de transducción de señales.
U: Tráfico intracelular, secreción y transporte vesicular.
V: Mecanismos de defensa.

Metabolismo.
C: Conversión y producción de energía.
G: Metabolismo y transporte de carbohidratos.
E: Metabolismo y transporte de aminoácidos.
F: Metabolismo y transporte de nucleótidos.
H: Metabolismo de coenzimas.
I : Metabolismo lipídico.
Q: Biosíntesis, trasporte y catabolismo de metabolitos secundario.

Pobremente caracterizados.
R: Sólo función general predicha.
S: Función desconocida.

Figura C1.6 Promedios de expresión (RPKM) para los genes de *S.ruber* M8 y M31 en cultivo puro anotados en cada una de las categorías COG

podría ser de reciente adquisición. Estas regiones albergan la mayoría de genes específicos de cepa y pertenecientes al genoma accesorio. En conjunto, los datos indican que estos elementos genómicos, plásmidos e islas, podrían ser un importante mecanismo de plasticidad genómica, facilitando la incorporación continuada de genes del *pool* accesorio del ambiente.

En el caso de los plásmidos, gran parte de genes no se expresan significativamente, o lo hacen a muy bajo nivel, la mayoría de ellos se expresaron por debajo de las 200 RPKM **(figuras C1.7 y C1.8**). Recientemente también se han observado niveles bajos de cobertura en plásmidos de organismos acuáticos de vida libre como *A. macleodii* (Kimes *et al*., 2014), aunque son escasos los datos publicados hasta la fecha. Hay excepciones, como la del gen pSR_11013 contenido en el plásmido pSR11 clasificado como VHEG, que codifica para una proteína de unión a DNA similar aun regulador transcripcional, en sintonía con el elevado número de reguladores transcripcionales encontrados entre los genes VHEG (anexo, **tabla S1.5; figura C1.7**).

En cuanto a HRVI y HRV II, encontramos una gran cantidad de genes con niveles de expresión bajos para ambas cepas, muchos de ellos específicos de cepa. La HRVI es la región génica con una mayor proporción de genes con baja expresión en ambos genomas. En el caso de la HRVII encontramos niveles de expresión bajos en su región 3'. Ambas islas presentan niveles anómalos de %GC (Pasic *et al*., 2009) lo que soporta la hipótesis de que genes con un contenido en GC distinto del promedio el genoma derivan de procesos recientes de LGT y presentan niveles de expresión por debajo de los óptimos (Haecker y Carniel, 2001; Bellanger *et al*., 2014). En conjunto estos datos, %GC, niveles de expresión bajos y diferencias en contenido génico entre dos cepas tan próximas, sugieren que la adquisición horizontal de genes, incluso procedentes de especies lejanas filogenéticamente, resulta bastante frecuente en ambas cepas. Apoyando esto, se observó en ambos genomas niveles mayores de expresión para aquellos genes con un %GC cercano al promedio de cada uno de los genomas (66,12% en M8 y 66,29% en M31) **figura C1.10**), los cuales se distribuyen normalmente en torno a este promedio. Dentro de la HRVII de ambos genomas, se observa una clara polaridad de expresión en la cual los 30 primeros genes situados en 5' tienen un nivel de expresión mucho más elevado que el del resto de genes

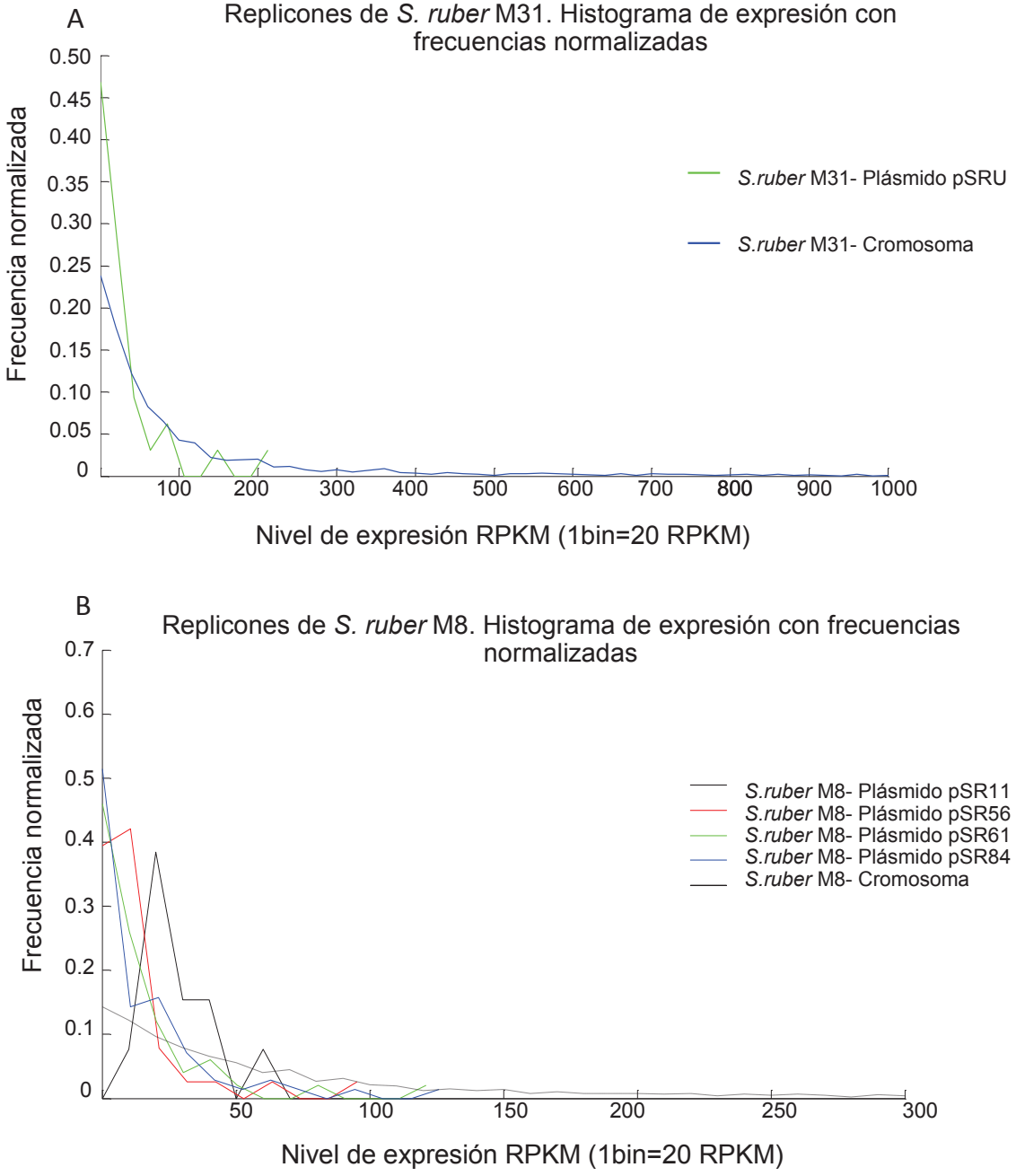


Figura C1.7. Distribución génica normalizada para M31 (figura A) y M8 (figura B) en base a sus niveles de expresión (RPKM). Todos los replicones se han incluído en cada caso.

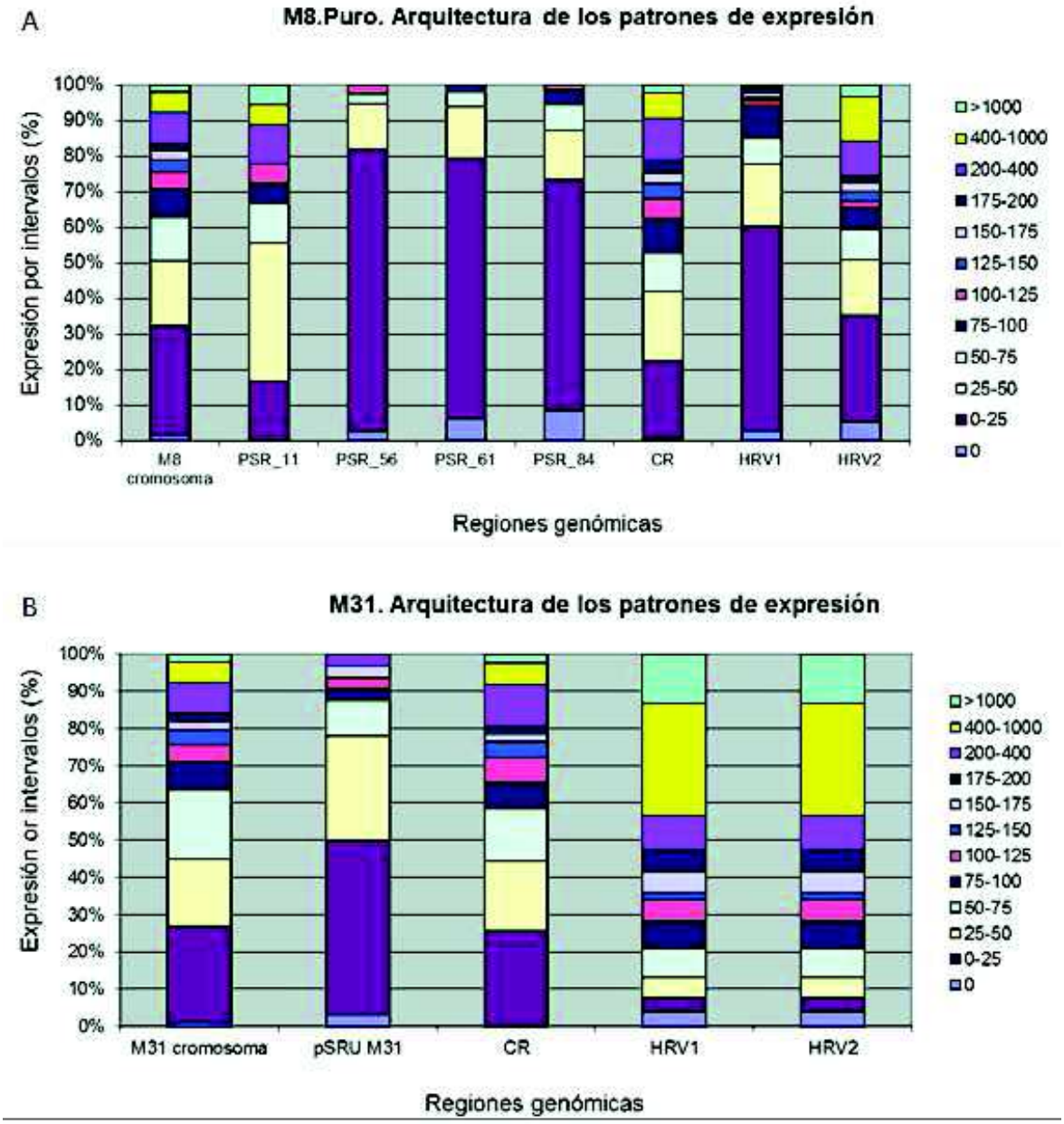


Figura C1.8. Distribución de los niveles de expresión por replicón en los genomas de M8 (figura A) y M31 (figura B). Distribución porcentual por niveles de expresión (RPKM) (leyenda superior derecha del gráfico) de los genes albergados en:
Figura A. El cromosoma, 4 plásmidos y regiones hipervariables (HRVI y II) y conservada (CR) (leyenda eje x) de la cepa M8.
Figura B. El cromosoma, plásmido pSRU, regiones hipervariables (HRVI y II) y conservada (CR) (leyenda eje x) de la cepa M31.

de estas islas (**figura C1.9**). Encabezando esta agrupación de 30 genes destaca la presencia de una integrasa de fago posicionada en 5',y en posición 3' respecto a la integrasa un gen involucrado en la síntesis de lipopolisacáridos, ambos VHGE. Muchos de estos 30 genes con elevados niveles de expresión están involucrados en la síntesis de envolturas celulares y frecuentemente en procesos de reconocimiento de hospedador (Peña *et al*., 2005), aunque en las condiciones del experimento no existe presión vírica.

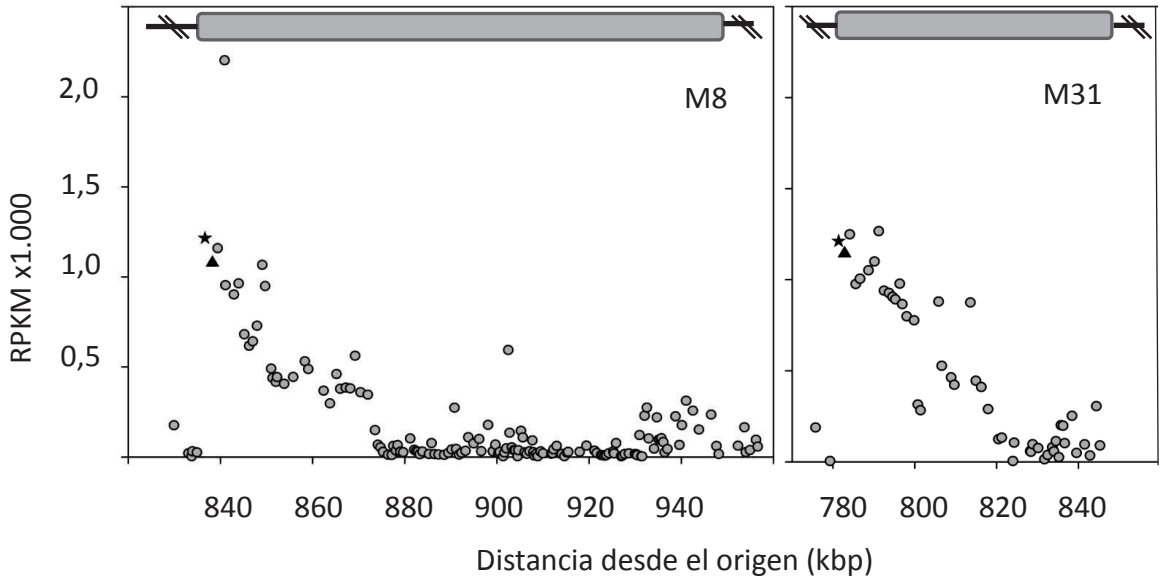


Figura C1.9. Valores de expresión (RPKMx1000) para los genes posicionados en la HRVII de M8 y M31 ordenados respecto al origen de replicación. Los bloques grises delimitan la extensión de las HRVs.

Al igual que en plásmidos e islas genómicas, en el cromosoma de ambas cepas encontramos regiones de genes cercanos entre sí con niveles de expresión similares, en ocasiones VHEG y otras muy poco expresados. Dentro del primer grupo sobresale la región contenida entre las ORFs SRM_01230-SRM_01030 en M8, SRU_1030-SRU_1061en M31. Muchos de estos genes codifican para proteínas ribosómicas y presentan niveles de expresión elevados y semejantes entre sí. El mapeo de los *reads* de RNAseq en esta región muestra como muchos de estos genes se coexpresan, formando parte de operones y regulones comunes, y favoreciendo la estequeometría molecular. Otros genes implicados en vías del metabolismo central, como los relacionados con la fosforilación oxidativa (srm00190), se agrupan de manera similar, lo que de

nuevo estaría indicando una co-regulación de genes implicados en vías metabólicas comunes. Además de estas agrupaciones, destaca la presencia de un grupo de alrededor de 60 genes prácticamente contiguos con niveles de expresión muy bajos en ambas cepas (**figura C1.5**) (SRM_2799 a SRM_2866 en M8 y de SRU_2582 a SRU_2649 en M31). La mayoría de estos genes codifican para proteínas involucradas en motilidad celular (COG N: biosíntesis de flagelo, estructura flagelar y factores quimiotácticos). Su relación funcional y la existencia de cobertura significativa de las zonas intergénicas indican que se trata de genes corregulados, algunos de ellos como parte del mismo operón tal y como muestran los mapeos de los *reads* secuenciados.

Dinámica genómica y análisis del proceso de adaptación e intercambio génico.

Otra tendencia global analizada fue la correlación entre los valores de expresión obtenidos en los transcriptomas en cultivo puro y el CAI (del inglés, *Codon Adaptation Index*), valor empleado en los estudios genómicos previos con estas cepas (Peña *et al*., 2010) como aproximación teórica a los niveles de expresión. En dichos estudios, los genes con valores de CAI más elevados fueron los involucrados en transporte y energía así como en metabolismo energético y de nucleótidos, mientras que la mayoría de genes con CAI bajos se anotaron como proteínas hipotéticas y transposasas. Gran parte de las proteínas ribosómicas mostraron unos valores de CAI típicos para microorganismos de crecimiento lento (Carbone *et al*., 2003). Aunque no se obtuvo una correlación clara entre los niveles de expresión y los valores de CAI (r= 0.0012), se observó que bajo las condiciones ensayadas había una mayor proporción de genes con niveles bajos de expresión entre los que presentaron un CAI inferior (**figura C1.10**), entre ellos un elevado porcentaje de proteínas hipotéticas. De acuerdo con los estudios previos mencionados, las categorías COG C (producción y conversión de energía) y F (transporte y metabolismo nucleotídico) fueron de las más expresadas (**figura C1.6**). La ausencia de una correlación numérica clara entre los valores de CAI y los niveles de expresión obtenidos sugiere que el valor CAI no proporciona una estimación cuantitativa acertada de los niveles de expresión cuantitativas. Otros factores como el entorno génico próximo y la función biológica de los genes

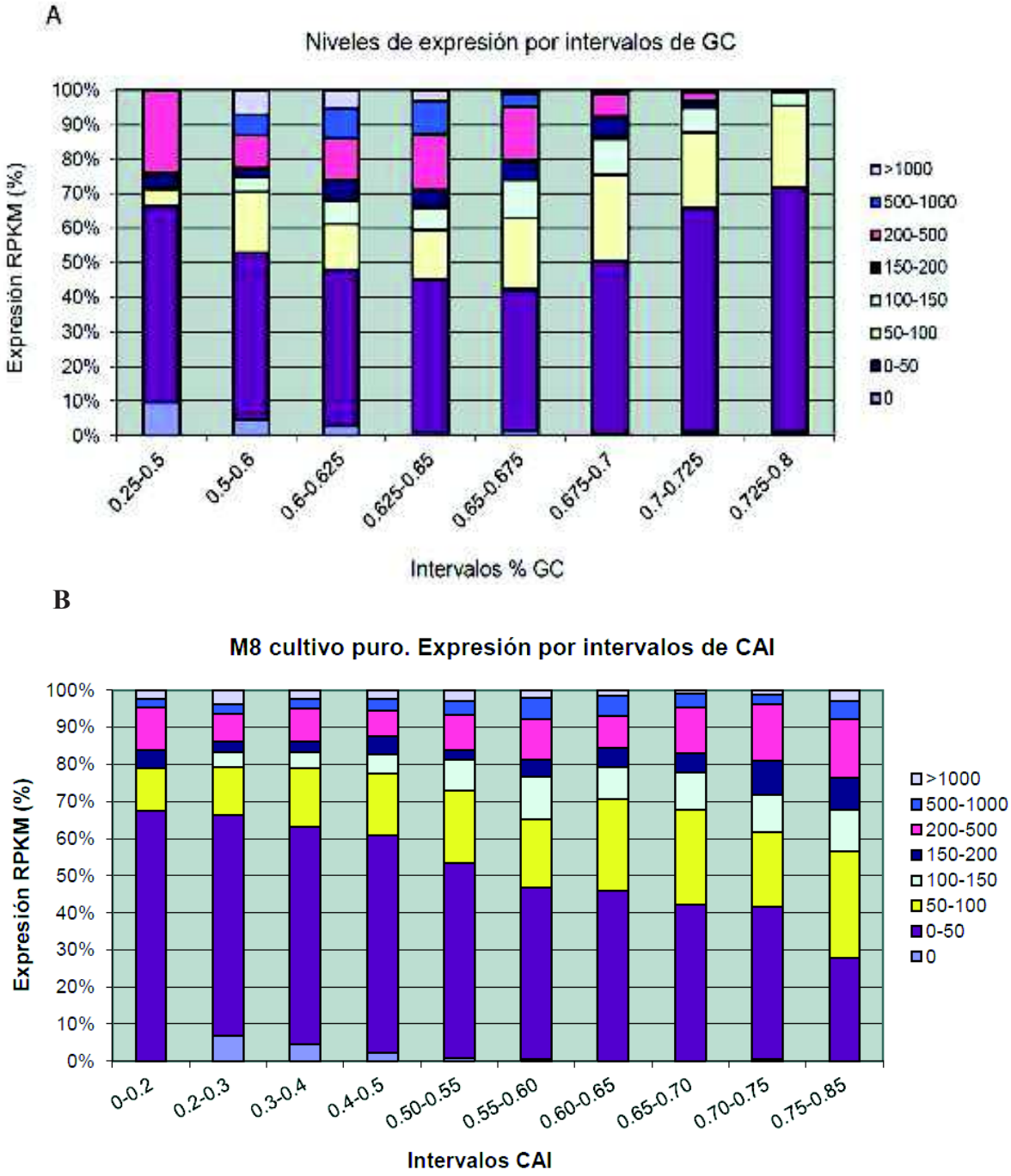


Figura C1.10. Distribución porcentual de genes en base a sus valores de GC y expresión (figura A) y CAI y expresión (figura B). El tamaño de los intervalos se incrementa con los valores de expresión (RPKM) a que niveles superiores contienen un menor número de genes.

podrían condicionar sus niveles de expresión. Sin embargo, a nivel cualitativo, genes con expresiones bajas tienden a poseer CAI bajos. En conjunto, los datos observados sugieren que aquellos genes mejor adaptados, cuyas secuencias reguladoras y codificantes se habrían sometido a adaptación durante más tiempo en su contexto génico y condiciones ambientales particulares, se expresan más eficientemente. El hecho de que se observe un enriquecimiento en proteínas hipotéticas y transposones entre aquellas que presentan un menor CAI junto a su localización dentro de islas genómicas y plásmidos apoya esta idea.

Con el objetivo de valorar el efecto del proceso adaptativo tras la adquisición de nuevos genes o tras su movilidad dentro del genoma se analizaron los niveles y diferencias de expresión entre parálogos contenidos en los genomas de M8 y M31. Se observó con este fin la expresión de parálogos procedentes de eventos de duplicación detectados y clasificados en un estudio anterior (Peña *et al*., 2010), en el que se diferenciaron 182 eventos de duplicación previos a la divergencia de las cepas M8 y M31(**tabla S1.9**), denominados ”eventos ancestrales”, y 23 posteriores o “eventos recientes” (**tabla S1.10**). Se analizaron las diferencias de expresión entre los parálogos dentro de cada uno de estos dos grupos, encontrando que las diferencias fueron mayores en el caso de los parálogos derivados de los eventos de duplicación ancestrales. Estos datos sugieren que tras su duplicación, los parálogos evolucionarían de manera independiente, acumulando cambios en su secuencia y modificando los niveles de expresión.

La mayoría (87%) de duplicaciones de eventos recientes involucraron a las islas genómicas, frente a tan sólo un 25.3% en las ancestrales (**figura C1.11**), y ninguna de ellas involucró a la CR. La mayoría de eventos recientes afectaron a genes de la categoría L (replicación, recombinación y reparación de DNA), en su mayoría transposasas, una de las más enriquecidas en las islas genómicas.

Entre las duplicaciones ancestrales, casi todos los eventos que involucraron HRVs se anotaron como elementos transponibles en el caso de la HRVII, aunque en los que afectaron a la HRVI encontramos además genes de las categorías P y T, en su mayoría transportadores iónicos, proteínas kinasas, elementos involucrados en sistemas de dos componentes y proteínas hipotéticas. Casi un 75% de las duplicaciones previas a la divergencia de ambas cepas no afectaron a las islas genómicas, un 10% incluyendo genes de la CR y el 65% restante de regiones

no incluidas en islas ni en CR (no CR no HRV). Dentro de este 75% encontramos gran cantidad de genes pertenecientes a las categorías COG T, P y N, incluyendo gran cantidad de transportadores de membrana y proteínas de respuesta a estrés. Además los análisis indican que 62 de los 182 eventos ancestrales habían perdido al menos uno de los parálogos en una de las dos cepa. El 50% (23/46) de las duplicaciones ancestrales en HRVs presentaron pérdidas de parálogos en una de las dos cepas mientras que en el caso de las ubicadas en la CR y el resto de regiones el porcentaje fue menor con un 17.6% (3/17) y 30.3% (36/119) respectivamente.

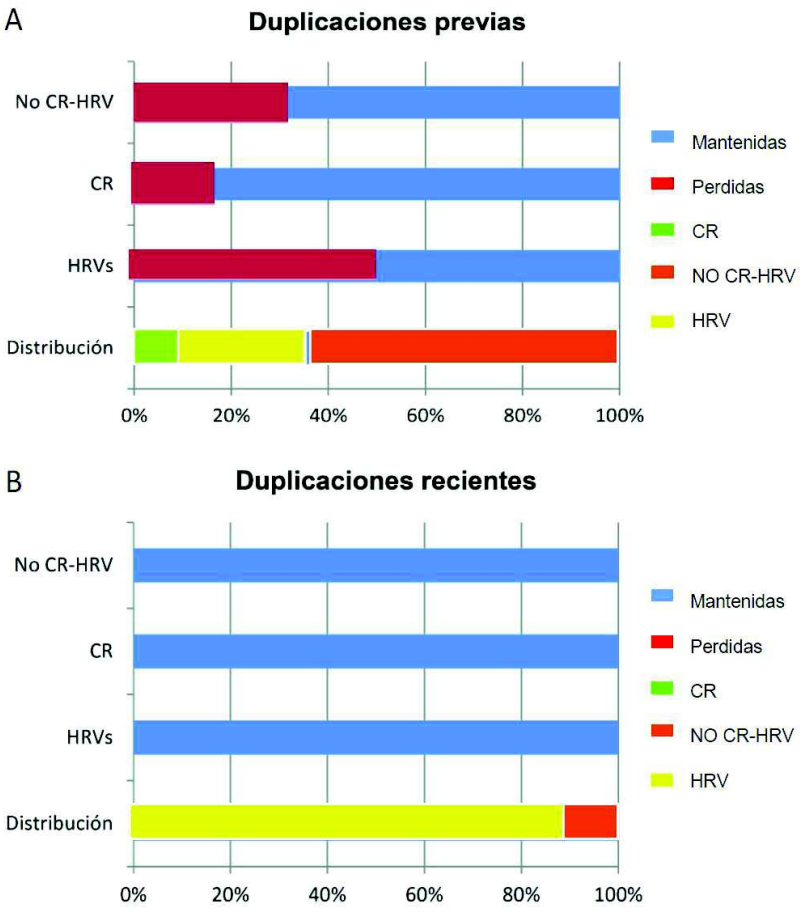


Figura C1.11. Distribución genómica y dinámica de los eventos de duplicación previos a la divergencia de M8 y M31 (figura A) y posteriores a la misma (figura B). Se detalla la distribución en CR, HRVs y el resto del genoma y el porcentaje de duplicaciones perdidas en cada una de estas fracciones.

En conjunto, los datos obtenidos sugieren que las islas genómicas de *S. ruber* constituirían elementos de tránsito fluido de genes al resto de zonas genómicas ya que la mayoría de eventos de duplicación involucrarían elementos transponibles, mecanismo considerado como uno de los motores de generación diversidad intraespecífica (Thomas y Nielsen., 2005). Aunque frecuentes, sólo una fracción de estas duplicaciones originadas en las HRVs se fijarían en el tiempo. Por otra parte, eventos de duplicación no mediada por transposición afectarían al resto del genoma. Tras la selección, las duplicaciones se irían adaptando al entorno génico, divergiendo en su secuencia y mecanismos reguladores, diferenciando sus perfiles de expresión de manera progresiva.

En este sentido, y con el objetivo de analizar el efecto de los cambios en la secuencia codificante sobre los niveles de expresión, se compararon los datos de expresión de 2537 ortólogos entre las cepas M8 y M31, que constituyen el *core* genoma. Estos genes se clasificaron en función a su valor de ANI (identidad a nivel de secuencia nucleotídica). Las diferencias de expresión se calcularon como el porcentaje que representa la diferencia de expresión absoluta entre ambos genes normalizada por promedio de expresión de cada par ortólogo. Tal como muestra la **figura C1.12**, se aprecia un incremento en la proporción de ortólogos con expresiones similares proporcional al incremento del ANI. Más de la mitad de los genes con un ANI mayor de 99,8 presentaron un cambio porcentual inferior al 20%. Estos resultados indican que, en general, las cambios a nivel de secuencia codificante son un factor que explica parte de las diferencias de expresión observadas entre ortólogos de ambas cepas. Sin embargo, esta fracción contiene genes con un ANI del 100% y diferencias de expresión considerables, incluso estadísticamente significativas como se verá más adelante. Estas diferencias podrían deberse a cambios en los mecanismos de regulación de expresión. Por ello, además se exploró el efecto de los cambios en secuencias reguladoras sobre los niveles de expresión. Entre las regiones genómicas la CR, caracterizada porque sus ortólogos presentan un dN/dS igual a 0, es la que presenta mayor similitud a nivel de secuencia codificante e intergénica, con un ANI promedio de 99,99%. La comparación de los cambios entre ortólogos con un ANI del 100% dentro de la CR y en el resto del genoma permitió explorar el efecto de los cambios acumulados en regiones reguladoras (**figura C1.12**). El conjunto de genes contenidos en la CR fue el más enriquecido

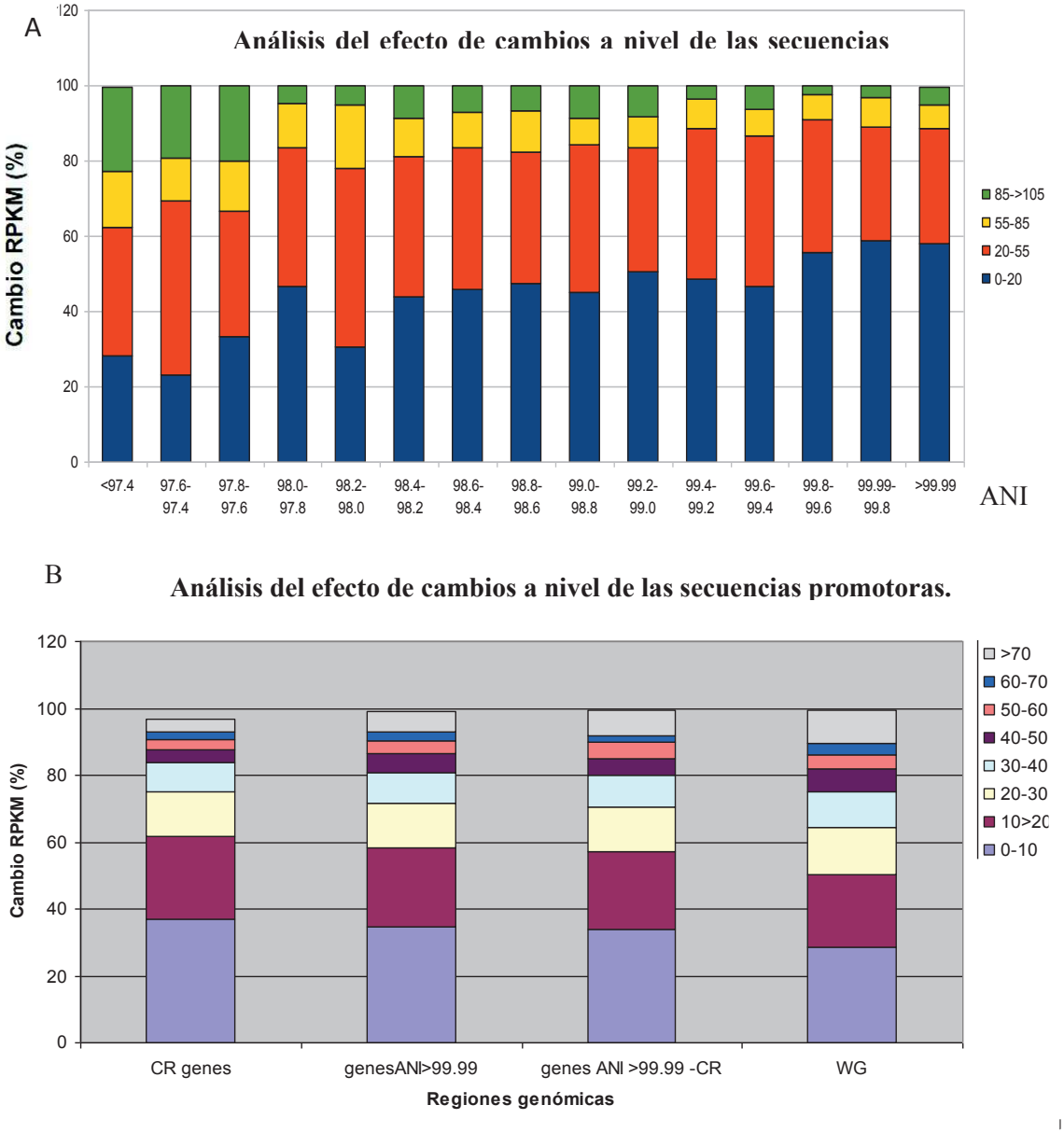


Figura C1.12. **Figura A.** Distribución de la tasa de cambio (%cambio) para los ortólogos de las cepas M8 y M31 de *S. ruber* para cada uno de los intervalos de ANI.
Figura B. Distribución porcentual por niveles de cambio de expresión (RPKM) para los genes contenidos en los distintos grupos considerados: genes pertenecientes a la zona conservada (CR), genes con un ANI de 100 (ANI>99.99), genes con un ANI de 100 pero no contenidos en la zona conservada (genes ANI>99.99-CR) y para los genomas completos (WG, un total de 2552 ortólogos).

para la fracción de cambio inferior al 10%. Estos datos reflejan como no sólo los cambios a nivel de secuencia codificante sino los acontecidos a nivel de secuencia reguladora constituyen factores determinantes a la hora de evaluar la microdiversidad funcional *de S.ruber.* Junto a las diferencias en el genoma accesorio, las diferencias en secuencias codificantes y reguladoras son las responsables de la diversidad en respuestas adaptativas a nichos naturales y evolución de cepas cercanas de una misma especie (Caro-Quintero y Konstantinidis., 2012; Yoder-Himes *et al*., 2009).

Diferencias en la expresión del genoma *core* entre cultivos puros.

Tras el análisis de los transcriptomas individuales en cultivo puro realizamos una comparación de los niveles de expresión de los ortólogos de ambos genomas empleando el software DEseq (Anders *et al*., 2010). La correlación entre ambos transcriptomas presentó valores elevados (r= 0,975), aunque inferiores a los obtenidos para la correlación entre las réplicas en cultivo mixto para cada cepa (r= 0,997 y r= 1 para M8 y M31 respectivamente) (**figura C1.2**). Alrededor de 165 genes compartidos por ambos genomas presentaron diferencias significativas en sus niveles de transcripción ($p<0,05$) (anexo, **tabla S1.11; figura C.1.5**). La proporción de genes diferencialmente transcritos ente M8 y M31 fue de 5 veces mayor en la HRVI y HRVII que en el resto el genoma. Por tanto las HRVs son regiones donde, además de la elevada microdiversidad génica al albergar una proporción elevada de genes específicos de cepa y divergentes, se observa una microdiversidad funcional. La elevada representatividad de genes expresados diferencialmente en las HRVs estaría indicando que las diferencias en los niveles de expresión detectadas se deben no sólo a divergencias en la secuencia codificante sino también a las del entorno génico. Estas últimas afectarían a los patrones de regulación transcripcional, primeros niveles en los cuales comienza a apreciarse divergencia microevolutiva tal como se apunta en estudios previos (Vicente y Mingorance., 2008; Yoder-Himes *et al*., 2009). El agrupamiento de los 165 genes en regiones génicas (**figura C1.5**) es otra evidencia de que ambas cepas presentan diferencias funcionales a nivel de mecanismos de regulación de la transcripción.

Es el caso de los 23 de entre los 165 situados en las HRVI y HRVII, de los cuales 21 muestran una mayor expresión significativa en M31, lo que podría indicar que esta cepa expresa más genes posicionados en estas regiones variables. Algo similar se aprecia en otras regiones génicas como la agrupación de 60 genes con baja expresión involucrados en motilidad, de los cuales 10 presentan diferencias significativas entre cepas y niveles mayores en M8. Estos perfiles reflejan probablemente diferencias microevolutivas funcionales a nivel de regulación de transcripción entre ambas cepas, lo cual tendría sentido al estar implicados una parte importante de estos 165 genes en mecanismos de transducción de señales como se expone a continuación. En algunos análisis de expresión, como los llevados a cabo entre cepas de *A. macleodii* (Kimes *et al*., 2014), se obtuvieron resultados similares en los cuales la fracción de genes con expresión diferencial significativa está algo enriquecida en genes del genoma accesorio, contenidos en islas genómicas y plásmidos.

Tras realizar los test de Fisher (**Tablas S1.12 y S1.13**), comprobamos que el grupo de 165 genes expresados diferencialmente estaba enriquecido en tres categorías COG: N (Motilidad celular) (**figura C1.13**), T (Mecanismos de transducción de señales) y U (trafico intracelular, secreción y transporte vesicular). El 89,5% (34/38) de los de genes de estas 3 categorías presentaron niveles de expresión mayores en M8. Además, el 88% de las diferencias transcripcionales observadas entre ambas cepas parecen estar relacionadas con respuestas ambientales, incluyendo transductores de señales quimiotácticos, sistemas de dos componentes (que como se acaba de mencionar a su vez pueden tener efecto sobre la expresión génica), flagelo o vías de secreción. Contrariamente, entre estos 165 genes encontramos una pobre representación de genes de la categoría COG J (Traducción, estructura ribosómica y biogénesis), acorde con las tasas de crecimiento entre las cepas M8 y M31 al crecer en cultivo puro.

Como se comentó con anterioridad, las principales diferencias de expresión al comparar todos los ortólogos se dieron en aquellos que presentaron un mayor grado de divergencia. En consonancia con estos datos, se observó un enriquecimiento en genes con bajo ANI en la fracción de genes que presentaron expresión diferencial significativa (anexo, **figura S1.2**). Sin embargo, entre ellos también encontramos genes con un ANI elevado, lo cual apunta a diferencias en las vías reguladoras como ya se comentó en los análisis anteriores. Este es el caso

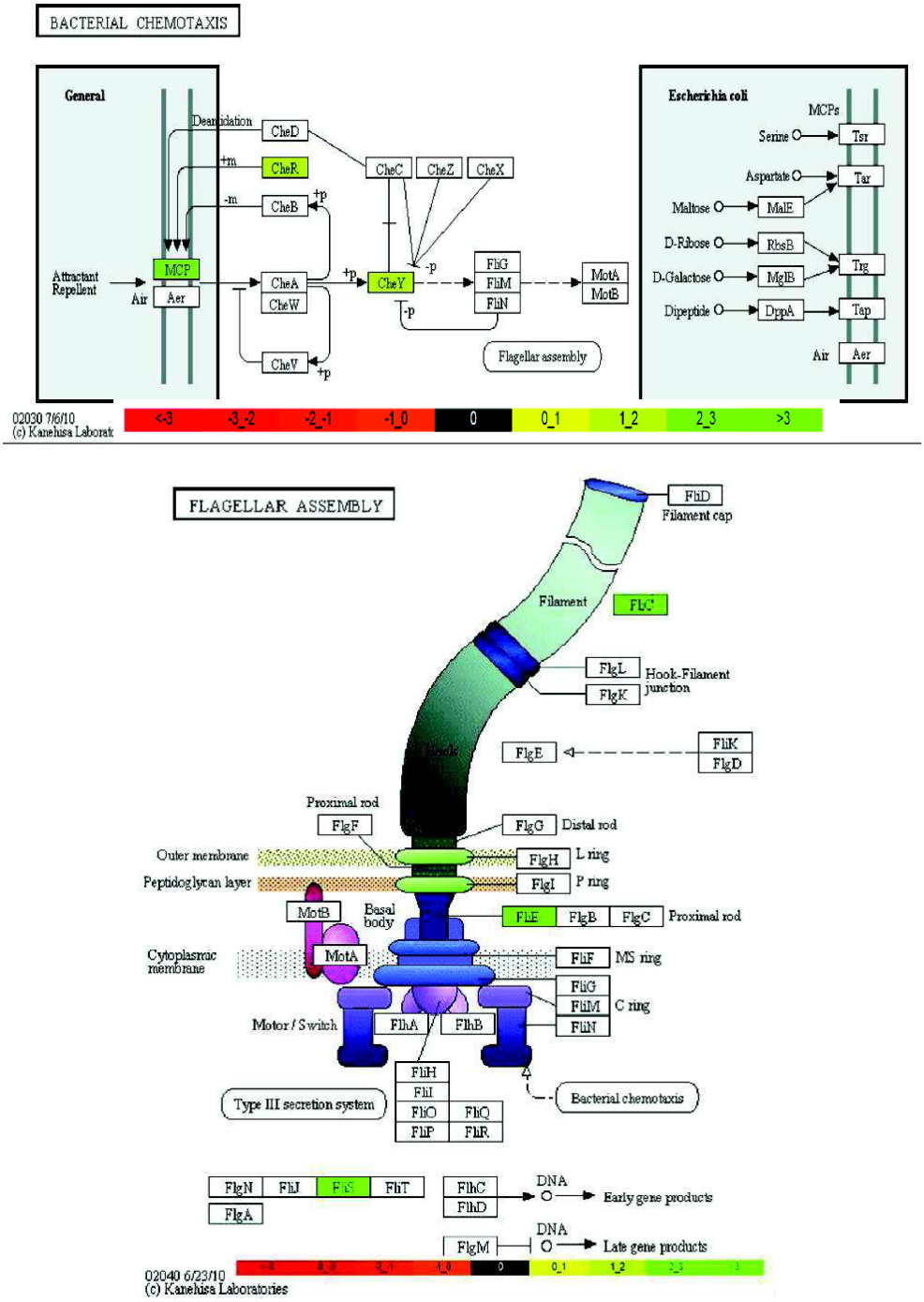


Figura C1.13. Genes con expresión diferencial entre las cepas M8 y M31 de *S.ruber* en cultivo puro pertenecientes a la categoría COG N. Coloreadas en ambas rutas del KEGG, biosíntesis flagelar y quimiotaxis, se muestran los genes con niveles de expresión mayores en la cepa M8.

de 38 genes con un ANI del 100% (**tabla S1.11**) entre los que encontramos muchos transportadores de membrana, genes implicados en transducción de señales y reguladores frente a señales externas. Genes que presentan un ANI elevado no necesariamente deben tener patrones de transcripción parecidos, por lo que análisis de microdiversidad a nivel génico pueden subestimar los niveles reales de la diversidad intraespecífica metabólica y fisiológica entre las cepas analizadas. Análisis comparativos de los transcriptomas de dos cepas cercanas de *B. cenocepacea* llevaron a conclusiones similares, ya que las diferencias transcriptómicas fueron mayores que las observadas a nivel de secuencias codificantes, atribuyéndolas a diferencias en vías reguladoras (Yoder-Himes *et al*., 2009). Además, el hecho de que encontremos una importante representación de genes relacionados con procesos de transcripción expresados diferencialmente entre cepas, y como se verá más adelante entre cultivos puros y mixtos, apoya la importancia y el papel de las diferencias en procesos reguladores globales sobre las diferencias en los transcriptomas de M8 y M31.

En conclusión, aunque las cepas M8 y M31 presentaron patrones generales similares en su arquitectura transcriptómica y a nivel de genes con elevados niveles de expresión, la comparación de sus transcriptomas muestra sutiles pero relevantes diferencias de expresión entre genes compartidos por estas dos cepas tan cercanas y cultivadas en las mismas condiciones. La mayor parte de las diferencias detectadas se apreciaron en genes involucrados en la interacción y detección de señales del ambiente, apuntando a que cada una de ellas podría emplear mecanismos específicos de interacción con el entorno.

5. Metatranscriptoma de cultivo mixtos de M8 y M31

La comparación de los transcriptomas de ambas cepas en cultivo puro y mixto permitió comprobar si la combinación de ambos transcriptomas en cultivo mixto equivale a la adición de los transcriptomas individuales o por el contrario si muestra el efecto de la interacción de ambas cepas. Aunque algunos estudios previos han mostrado variaciones en transcriptomas individuales al enfrentar cepas de especies distintas (Garbeva *et al*., 2011), hasta el momento la aproximación propuesta en este trabajo, el análisis de cultivos mixtos de cepas de la misma especie, no se ha

llevado a cabo, menos aún con cepas muy próximas y con grandes similitudes genómicas como las presentadas por M8 y M31. Un ejemplo es el estudio de la interacción de cepas de especies distintas fue el realizado en *Acidovorax avenae* subesp. *avenae* cultivada individualmente y en presencia de la bacteria *Burkholderia seminalis* (Li *et al*., 2014).

Los estudios de expresión diferencial entre cultivo puros y mixtos combinaron el empleo de dos programas computacionales, Cufflinks (Trapnell *et al*.,2010) y DEseq (Anders *et al*., 2010). Entre ambos programas se detectaron un total de 354 y 446 genes expresados diferencialmente en M8 y M31, respectivamente, al comparar sus transcriptomas en cultivo puro y mixto ($p < 0,05$) (**tablas S1.14 y S1.15**). La mayoría de los genes detectados por DEseq se detectaron también por CUfflinks (**anexo, figura S1.3**), apoyando los resultados obtenidos con cada uno de los *software*. Se excluyeron de los análisis 418 genes, muchos de ellos contenidos en la CR, debido a su elevada similitud, que impidió asignar niveles de expresión de manera inequívoca a una única cepa en cultivo mixto. El conjunto de genes con expresión diferencial mostró niveles de expresión similares entre ambas cepas cuando crecieron en cultivo puro, presentado niveles de correlación elevados ($r=0,989$ para los 354 genes de M8 y $r=0,984$ M31 para los 446 de M31).

Expresión diferencial en genomas *core* y acesorio.

Tal como se muestra en la **tabla C1.2**, la proporción de genes del *core* genoma que presentó expresión diferencial varió entre el 11.8%-28.3% en M8 y 16.6%-33% en M31, dependiendo de si se consideran o no los 418 genes a los que no se puede asignar cambio al pasar de cultivo puro a mixto, mientras que sólo un 7% de los genes del genoma accesorio mostraron cambios de expresión significativos. Los datos obtenidos muestran como, a excepción de la CR como se verá más adelante, la proporción de genes que presentan cambios de expresión diferencial en el *core* genoma es mucho mayor que en genes del accesorio, la mayoría de ellos localizados en plásmidos y HRVs. Tal como se discutió anteriormente, una proporción importante de los genes del genoma accesorio muy probablemente son de reciente adquisición. Este hecho explicaría que la mayoría de ellos no estén adaptados a los mecanismos reguladores

Tabla C1.2. Resumen de los resultados principales tras la comparación de cultivos puros y mixtos de *S.ruber*.

Rasgo analizado	Genoma analizado	
	M8	**M31**
Número de ORFs en el genoma	3303	2898
ORFs en el genoma accesorio	766	361
ORFs en islas genómicas (GI)	280	128
ORFs analizadas en cultivo mixto	3303-418=2885	2898-418=2480
ORFs con expresión diferencial de cultivo puro a mixto	354	446
ORFs con expresión diferencial pertenecientes al core/accesorio	300/54	421/25
ORFs con expresión diferencial en ambas cepas M8 y M31	89	
ORFs con aumento de expresión de puro a mixto	142 (40%)	289 (65%)
ORFs con disminución de expresión de puro a mixto	212 (60%)	157 (35%)
ORFs con expresión diferencial detectados en islas	18	17
ORFs con expresión diferencial detectados en CR	2	3
Principales categorías funcionales representadas y número de ORFs con expresión diferencial		
ORFs involucradas en funciones de trasporte	53	38
ORFs involucradas en sistemas de dos componentes	7	7
Kinasas	4	9
Quimiotaxis y flagelo	6	1
Transcripción: regulación	9	7
Traducción: factores traduccionales/proteínas ribosómicas estructurales/tRNA sintasas/tRNA rRNA metilasas	0/0/3/0	6/20/6/3
Replication y recombinación de DNA	10	9
Proteínas secretadas: Péptidos señal TAT/ Péptidos señal P.	32/3	18/1

del contexto génico y ambiental, lo que afectaría a la capacidad de modulación de sus niveles de expresión. Estudios previos llevados a cabo en *C. difficile* en diferentes condiciones de crecimiento muestran del mismo modo más cambios de expresión en la fracción del *core* genoma que en la del genoma accesorio (Scaria *et al*., 2013). En nuestro estudio, entre los genes que presentaron mayores niveles de cambio entre cultivos puros y mixtos (foldchange > 1; **tablas S1.14 y S1.15**), encontramos una cantidad importante de proteínas hipotéticas a las cuales no se

les ha podido asociar una posible función. En estudios previos con cepas de especies como *B. cenocepacia* o *A. macleodii* se ha detectado también una proporción importante de genes hipotéticos que podrían estar involucrados en respuestas adaptativas al ambiente, (Yoder-Himes *et al.,* 2009; Kimes *et al*., 2014). Este hecho subraya el interés de la caracterización bioquímica de los genes codificantes para HP, que todavía representa una parte considerable de los genomas microbianos. El primero de los estudios refleja cambios adaptativos importantes en clases funcionales como transcripción y traducción de señales, dos de las categorías funcionales con una mayor representación de genes en *S. ruber* entre el grupo de genes que presenta expresión diferencial (**Tabla C1.2**).

Arquitectura genómica y expresión diferencial.

Entre las regiones genómicas de *S.ruber*, la CR contiene los genes que mostraron menor variación en sus perfiles de expresión al comparar ambas condiciones de cultivo (**figura C1.14**). Aunque en esta región están contenidos gran cantidad de los 418 genes con elevada similitud de secuencia, 43 y 35 genes en las CR de M8 y M31 presentaron diferencias suficientes para llevar a cabo el mapeo de manera inequívoca. De entre estos últimos tan sólo 2 genes en M8 y 3 en M31 mostraron expresión diferencial, valores bastante bajos si los comparamos con la distribución observada a lo largo del resto del genoma. De hecho la probabilidad de encontrar valores tan bajos de genes con expresión diferencial en intervalos de 43 y 35 genes a lo largo de los genomas de M8 y M31 es de 0,077 y 0,055 respectivamente. La CR es una región enriquecida en transportadores iónicos (Peña *et al*., 2010), que contiene la isla de halofília, descrita por primera vez en en *S.ruber* M31 (Mogondín *et al.,* 2005), que incluye canales catiónicos y transportadores de aminoácidos catiónicos esenciales para la vida en condiciones halófilas extremas. Estos datos, junto a la estabilidad génica de esta región a la que se apuntó anteriormente y la elevada identidad de secuencia incluso a nivel de las regiones reguladoras sugiere que la CR contiene funciones esenciales bajo una fuerte regulación y control.

Relevancia ecológica de la expresión diferencial: respuestas comunes y específicas de cepa.

Las diferencias en los perfiles de expresión de M8 y M31 fueron significativas desde un punto de vista cuantitativo como cualitativo debido al porcentaje de genes implicados y regiones genómicas (**figura C1.14; tablas S1**) y la relevancia ecológica de los cambios que se detallan a continuación. Cada cepa respondió de manera específica a la presencia de la otra ya que menos de la cuarta parte (89) de los genes con cambios de expresión significativos cambió su expresión en ambas cepas. Por tanto, cada cepa tuvo una reacción específica ante la presencia de la otra. En términos generales una mayor proporción de genes (casi 67%) incrementaron sus niveles de expresión en M31, mientras que en M8 se apreció una mayor presencia de genes que los redujeron **(**anexo, **figura S1.3**). Estos datos reflejan que M31 se mostraría más activa de acuerdo con la proporción de células 10/8 a favor de M31 para este punto en mitad de la fase exponencial.

Sólo 9 de los 89 genes que mostraron expresión diferencial en ambas cepas lo hicieron en sentidos opuestos, es decir con incremento en un caso y disminución en el otro. De estos 9 genes, 8 se sobreexpresaron en M31, y 3 de ellos codificaban para proteínas hipotéticas secretadas. Entre los 89 genes, se identificaron 17 transportadores, lo cual representa un porcentaje considerable, destacando la disminución de expresión de los genes codificantes para xantorrodopsina y bacteriorrodopsina y un incremento en la expresión de transportadores de solutos compatibles como la glicinbetaína. El gen *dpr*A-SMF involucrado en procesos de transformación natural, experimentó represión en ambas cepas. Entre el resto de genes con expresión diferencial sólo en una de las dos cepas, una proporción importante de genes aumentaron la expresión en M31 y la disminuyeron en M8 sin presentar cambios significativos en la otra cepa.

Cada cepa mostro una respuesta adaptativa específica. Como se mencionó anteriormente, en este punto de la fase exponencial observamos una densidad celular mayor en M31. De acuerdo con esto, M31 experimentó una sobreexpresión de varias vías anabólicas, biosíntesis de proteínas ribosómicas (anexo, **figura S1.4**), biosíntesis lipídica, genes que codifican para elementos del replisoma, factores de transcripción, algunas vías de biosíntesis de aminoácidos a

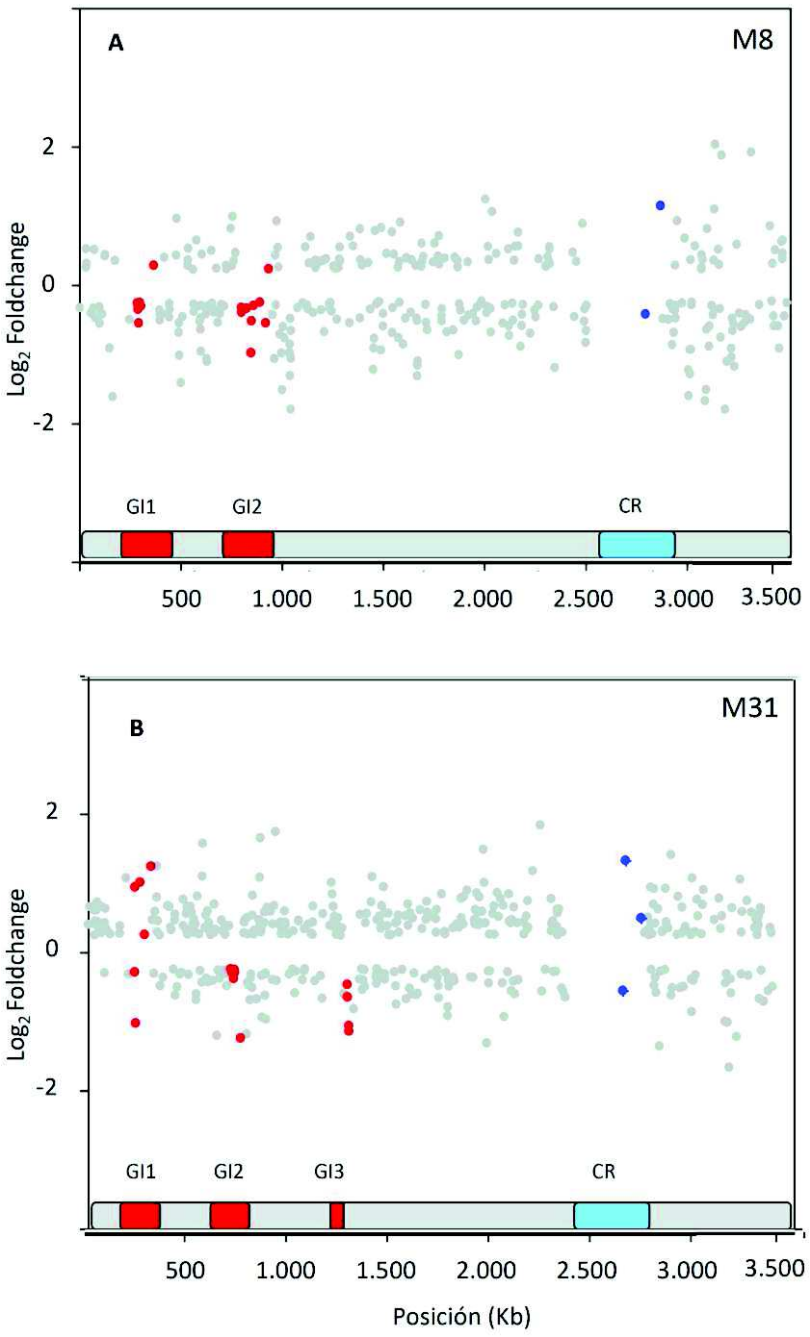


Figura C1.14. Distribución y cambio de los 354 (figura A) y 446 (figura B) genes que presentaron expresión diferencial significativa en M8 y M31respectivamente. Se destaca con puntos rojos aquellos localizados en islas genómicas y con azules los situados en la CR. Los bloques inferiores en rojo y azul delimitan las posiciones de las islas genómicas (GI) y la zona conservada (CR) respectivamente.

través de intermediarios del ciclo del ácido cítrico (CAT) (**figura C1.15, tabla C1.2**) y algunas proteínas específicas de cepa como glicosilasas (**tabla S1.11**). Entre las vías anabólicas sobreexpresadas en M31 se detectó un enriquecimiento significativo en las categorías COG J

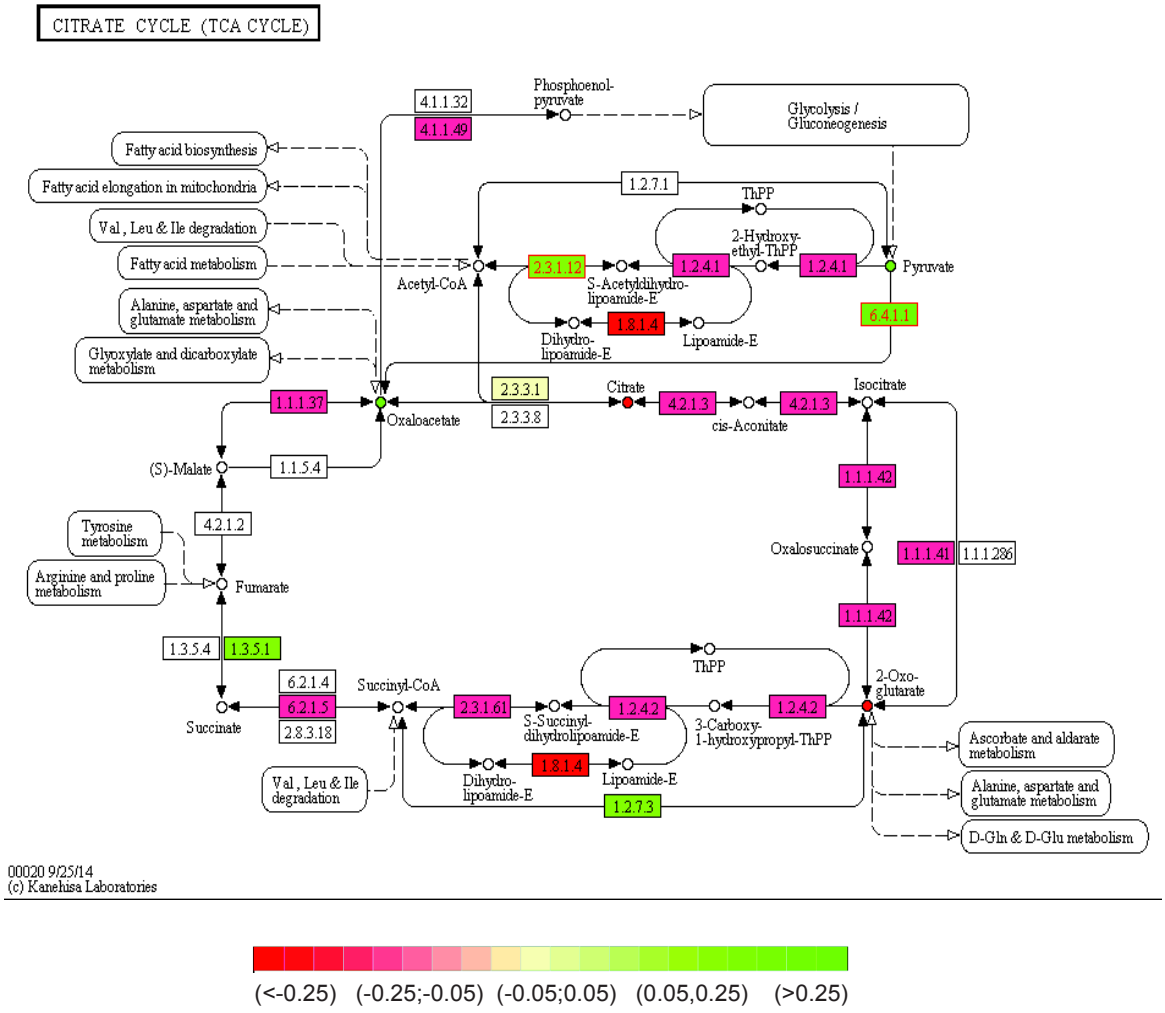


Figura C1.15. Mapa del KEGG en el que se muestran los cambios de expresión diferencial (log_2 fold change) para ciclo de los ácidos cítricos (CAT) (sru:00020) de la cepa M31 mediante un gradiente porcentual. Las cajas verdes representan los genes que incrementan sus niveles de expresión y las rojas las que los disminuyen. En letras rojas se destacan los genes que cambian significativamente sus niveles de expresión (SRU_828 = 6.4.1. 1; SRU_1969 = 2.3.1. 12). Los círculos muestran los incrementos (verde) o decrementos (rojo) de metabolitos intermediarios en base a la expresión de los genes de las proteínas que los sintetizan.

(traducción, estructura ribosómica y biogénesis) y COG O (modificación postraduccional, reciclado proteico y chaperonas) así como en términos GO relacionados con expresión génica, traducción y elementos ribosómicas (**anexo, tabla S1.16 y S1.17**). En conjunto, las respuestas observadas indicarían un incremento notable en las vías del metabolismo central de la cepa M31, acorde con la mayor tasa de división observada. Las vías anapleróticas y la glucolítica suministrarían intermediarios al CAT dada la demanda de rutas anabólicas como las implicadas en biosíntesis de aminoácidos o ácidos grasos. Acorde con el incremento metabólico, se incrementa la demanda de maquinaria de traducción, expresión y plegamiento de proteínas. En particular, la abundancia relativa de transcritos de proteínas ribosómicas se ha sugerido como estimación de las tasas de crecimiento *in situ* (Gifford *et al*., 2012). En este caso encontramos una buena correlación en cultivos puros en donde las diferencias de crecimiento modestas a favor de M31 se reflejan en un promedio de *reads* normalizados para los genes codificantes de proteínas ribosómicas (53,777 y 50,247 RPKM para M31 y M8 respectivamente).

M8 mostró, por el contrario, una represión importante en genes de las categorías COG D (control del ciclo celular), COG N (motilidad celular y secreción) y COG V (mecanismos de defensa) (**tabla C1.2**). La respuesta específica de M8 no se caracterizó por un incremento masivo de expresión de vías anabólicas, sino que experimentó una respuesta general a estrés ambiental. Al contrario que en M31, se detectó la represión de varias glicosilasas específicas de cepa y la mayoría de proteínas implicadas en sistemas de dos componentes que presentaron expresión diferencial (10 de 11). Se detectó sobreexpresión de algunos genes involucrados en respuesta a estrés, tales como los factores sigma E y ECF, y varias subunidades de la NADH deshidrogenasa. Además se apreció un incremento de expresión de canales de resistencia multidroga, resistencia a beta-lactámicos y producción de penicilina y cefalosporinas. Por último, como mecanismo adicional de respuesta a estrés se observó la sobreexpresión del gen de la recombinasa RecR, involucrada en procesos de reparación de DNA tras daño por estrés. El conjunto de respuestas de M8 sugiere que en las condiciones analizadas la presencia de M31 somete a estrés a la cepa M8.

¿Las cepas M8 y M31 detectan la presencia de la otra mediante interacción directa o indirecta?

En conjunto, los datos muestran que ambas cepas estarían reaccionando ante la presencia de la otra adaptando sus perfiles de expresión. Moléculas como los antibióticos se han propuesto como candidatas a la hora de mediar la interacción y señalización entre poblaciones bacterianas (Davies., 2006). En este sentido, los datos obtenidos sugieren que vías como la resistencia a beta-lactámicos y biosíntesis de penicilina y cefalosporinas estarían entre las que presentan los mayores cambios de expresión. Además, entre los genes que presentan expresión diferencial encontramos varios que codifica para canales de resistencia a drogas y transportadores. Las siete vías metabólicas relacionadas con antibióticos en *S. ruber* incluyeron al menos un gen con expresión diferencial en al menos una de las dos cepas (**anexo, tabla 1.11**), aunque el reducido número de proteínas mapeadas en estas vías impide poder analizar si tal enriquecimiento es o no significativo.

En vista de los resultados anteriores parece factible que las cepas de *S.ruber* estén interactuando de manera directa en co-cultivo. Sin embargo entre los genes que presentaron expresión diferencial entre cultivo puro y mixto encontramos gran cantidad de transportadores, algunos de los cuales sólo cambian su expresión en una de las dos cepas. Esta actividad diferencial, entre cultivos puros y mixtos y entre puros de ambas cepas, hace pensar que a medida que avanza el crecimiento bacteriano el medio de cultivo pudiera ir cambiando su composición química. Si así fuese, el efecto o interacción de ambas cepas podría darse no por interacciones bioquímicas derivadas de una actividad específica sino de manera indirecta por la alteración de la composición del medio. Con el objetivo de discernir cuál de las dos hipótesis, interacción directa o indirecta, es la más probable, seleccionamos algunos biolementos (Fe, Co, Ni, Zn, Cu) movilizados por transportadores que presentaron expresión diferencial y que actúan como cofactores de muchas proteínas. Se tomaron medidas de concentración de estos a lo largo de toda la fase exponencial (**anexo, tabla S1.18**), incluyendo el punto correspondiente al análisis transcriptómico, sin que se observaran diferencias significativas ($p<0,05$) tras realizar una comparación de medias entre sus concentraciones en cultivos puro y mixto. Por tanto resulta

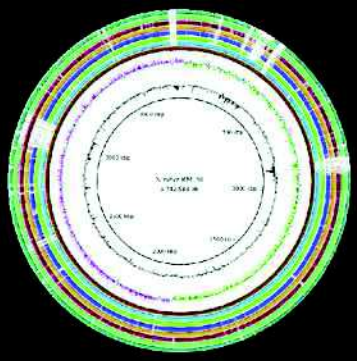
improbable que la interacción entre ambas cepas se deba a una variación en la composición iónica del medio, sosteniendo la hipótesis planteada inicialmente de una interacción molecular específica de cepa. Esta hipótesis se ve sustentada además por los resultados publicados con anterioridad (Antón *et al.*, 2013), en los cuales se muestra que varias cepas coaisladas de *S. ruber* presentan diferencias metabolómicas cuando se crecen en las mismas condiciones de cultivo. Cada una de las 62 cepas analizadas, entre ellas M8 y M31, presentó un perfil metabolómico extracelular diferente. El estudio de M8 y M31, por otra parte, ya había mostrado diferencias metabolómicas en la fracción extracelular (Peña *et al*., 2010). Estos datos sugieren que cada cepa podría modificar la composición del ambiente de manera específica, por lo que su presencia podría ser detectada por microorganismos de la comunidad que cohabiten. Aunque en la presente tesis no se ha determinado la naturaleza de estos compuestos secretados que median la interacción de las cepas M8 y M31, en el citado estudio previo (Antón et al., 2013) se identificaron patrones metabolómicos distintos relacionados con el metabolismo lipídico y la producción de antibióticos entre cepas de *S. ruber* aisladas de una misma muestra ambiental (Antón *et al*., 2013). Sabiendo que los antibióticos son compuestos que tienen funciones reguladoras en microrganismos en concentraciones subhinibitorias en la naturaleza (Davies, 2006; Bernier y Surette 2013), y habiendo detectado expresión diferencial de algunos genes relacionados con la síntesis de antibióticos, parece razonable proponer que, entre otras moléculas, la interacción entre las cepas M8 y M31 podría estar mediada por antibióticos o compuestos relacionados.

Introducción

Objetivos

Materiales y métodos

Resultados y discusión



Capítulo 1

Análisis de las diferencias transcripcionales e interacción de cepas cercanas de *S.ruber* mediante RNAseq.

Capítulo 2

Estudio de los mecanismos y estrategias de diversificación genómica en *S. ruber*

Capítulo 3

Impacto de la recombinación homóloga sobre la evolución de genomas *core* procariotas

Conclusiones

Bibliografía

Anexos

Resumen

En este segundo capítulo se aborda el análisis comparativo de los genomas de 8 cepas de *S. ruber* con el objetivo de cuantificar el grado de diversidad genética existente entre ellos, describirla en detalle y profundizar en los principales mecanismos que mantienen esa microdiversidad y dirigen la evolución de la especie. Esta última cuestión se abordó mediante el análisis de los mecanismos evolutivos que afectan a los genomas *core* y accesorio de la especie. Para ello, en primer lugar se secuenciaron completamente 6 genomas de cepas aisladas de Mallorca y Santa Pola y se diseñó una *pipeline* de ensamblaje. Esta última empleó un único tipo de librería de secuenciación y combinó los resultados de distintos ensambladores, completando los *gaps* mediante mapeo y extensión en una etapa final. Los genomas analizados presentaron tamaños entre 3,5 y 3,8 Mb y contuvieron un número de plásmidos variable. Los cromosomas mostraron una elevada sintenia, interrumpida por las dos zonas hipervariables (HRVs) presentes en todas las cepas, y niveles de identidad de secuencia global (ANI) superior al 97% .

Muchos de los genes específicos de cepa se situaron en las HRVs, plásmidos e inserciones de gran tamaño e *indels,* elementos genómicos que concentran la mayor parte de la microdiversidad. Las HRVs y los plásmidos contuvieron una proporción elevada de genes de las categoría COG M, COG V y codificantes de proteínas hipotéticas, elementos transponibles y de origen vírico. El análisis de la arquitectura de las HRVs reveló patrones de similitud característicos y bloques sinténicos conservados similares a los ya descritos en especies como *Shewanella baltica* (Caro-Quintero *et al.,* 2011*)* y *Alreromonas macleodii* (López-Pérez *et al*., 2014). Algunos de los plásmidos codificaron sistemas de restricción modificación y CRISPR-Cas, los cuales restringen el DNA foráneo incorporado a la célula y los procesos de intercambio génico y transferencia horizontal. Los intercambios genómicos detectados entre HRVs, plásmidos y resto del cromosoma ilustraron la elevada dinámica y movilidad génica de estas regiones. Por otro lado se analizó la evolución del genoma *core* de la especie, que comprendió regiones genómicas extensas de elevada sintenia. Tal como se ha observado en estudios previos con organismos extremófilos tales como *Halorubrum* sp., los mecanismos de recombinación homóloga actuarían homogeneizando el genoma *core* y favoreciendo el intercambio horizontal entre líneas clonales, consituyendo elementos cohesivos dentro de una población.

1. Introducción

Los procesos de variación génica que actúan sobre especies y líneas clonales se han estudiado ampliamente (Thomas y Nielsen 2005; Cohan 2006, Fraser *et al*., 2007, Vos y Didelot 2009; Didelot y Maiden 2010; Polz *et al*., 2013). Cada vez son más las evidencias que muestran variaciones importantes en el contenido génico entre cepas de una misma especie, e incluso coaislados. Estas diferencias en contenido génico a menudo van acompañadas de cambios en la sintenia que son el resultado de grandes reordenamientos genómicos y procesos de recombinación, tanto homóloga como no homóloga, que pueden ocurrir incluso entre cepas pertenecientes a taxones claramente distintos (Mira *et al*., 2010).

Históricamente la mayor parte de estudios genómicos evolutivos se han centrado en bacterias patógenas (Morelli *et al*., 2010, Reeves *et al*., 2011) siendo muy limitados los dedicados a bacterias de relevancia ambiental. Estos estudios analizaban las variaciones de secuencia tras procesos pandémicos de infección (Mutreja *et al*., 2011, Hiller *et al*., 2010), una situación muy particular que, aunque de gran importancia en el ámbito clínico, apenas aportaba datos acerca de los procesos evolutivos en bacterias de vida libre. Esta tendencia ha cambiado en los últimos años, con el aumento del número de publicaciones dedicadas a bacterias de vida libre.

El acceso al pangenoma de una especie (véase apartado 1.1.4 de la introducción) es el primer paso a la hora de comprender la microdiversidad existente en diferentes especies procariotas y los mecanismos evolutivos que mantienen la misma. Actualmente se emplean dos estrategias principales para aproximarse a la microdiversidad de una especie para a continuación analizar los mecanismos que la dirigen: el análisis de genomas completos de aislados o el uso de metagenomas en los que la especie a analizar constituye una población bien representada dentro de la comunidad (Mira *et al*., 2010). La primera implica el estudio del genoma de cepas de una misma especie aisladas de diferentes puntos geográficos y a diferentes tiempos. Aunque inicialmente este tipo de análisis se veía limitado por el número de genomas disponibles, cada vez resultan más completos. Un ejemplo de esta tendencia son los estudios con bacterias ecológicamente relevantes tales como *Shewanella baltica* (Caro-Quintero *et al*., 2011), *Vibrio cycliophicus* (Shapiro *et al*., 2012), *Sulfolobus islandicus* (Cadillo-Quiroz *et al*., 2012), *A.*

maleodii (López- Pérez *et al*., 2014) o *H. walsbyi* (Martín-Cuadrado *et al*., 2015), han permitido analizar la diversidad genómica en la distribución de plásmidos, islas genómicas (GI) e *indels* (i.e. inserciones y deleciones), además de los procesos de recombinación heteróloga e ilegítima que afectan al genoma accesorio de la especie y el papel de la recombinación homóloga en el genoma *core*.

La segunda estrategia de aproximación al pangenoma consiste en el empleo de secuencias metagenómicas procedentes de ambientes en los cuales la especie de interés esté bien representada. Un metagenoma incluye las secuencias de diferentes individuos de una especie dentro de una población, y esta información puede emplearse a la hora de determinar la diversidad genómica de la especie. Para ello debe emplearse al menos un genoma de referencia que permita reclutar o mapear las secuencias de esta especie. Este tipo de aproximación ha permitido explorar la diversidad genómica de especies como *Prochlorococcus marinus* (Coleman *et al*., 2006) o *H. walsbyi* (Cuadros-Orellana *et al*., 2007, Tully *et al*., 2015), identificando las islas genómicas como regiones con baja cobertura que apenas reclutaron secuencias homólogas. Estudios metagenómicos en *Prochlorococcus* sp. (Luo y Konstantinidis., 2011) y poblaciones de *Crenarchaea* marinas (Konstantinidis y DeLong 2008) han revelado la estructura y contenido genómico poblacional de estos organismos. Estos estudios sugieren que la estructura poblacional en *clusters* observada podría derivar del efecto de la recombinación homóloga, que mantendría los niveles de diversidad de secuencia entre genomas de una misma población entre el 1% y el 5% (Konstantinidis *et al*., 2006**;** Konstantinidis y DeLong 2008).

Como aproximación a la descripción de la microdiversidad de la especie *S. ruber* y del pangenoma de la misma, el presente estudio se abordó la secuenciación y ensamblaje de los genomas de 6 cepas de *S. ruber* pertenecientes a dos ambientes similares y próximos geográficamente, dos de ellas aislados de las salinas de Campos de Mallorca (RM158, M1) y 4 de las salinas de Bras del Port de Santa Pola (P13, P18, SP38 y SP73), algunas de ellas coaisladas y otras aisladas a lo largo de 8 años (véase tabla 1M, material y métodos). Como principales objetivos de este segundo capítulo se plantearon:

- Establecer una *pipeline* de ensamblaje y anotación integradas que permitan el ensamblaje de genomas completos a partir de un único tipo de librería de secuenciación.

- Caracterizar la microdiversidad de la especie *S.ruber* mediante el análisis comparativo de 8 cepas secuenciadas, los 6 nuevos ensamblados y los de M8 y M31 descritos ya previamente.
- Analizar los patrones de diversidad y los posibles mecanismos de microdiversificación de la especie y posibles patrones biogeográficos.
- Estudiar el impacto de la recombinación homóloga sobre el genoma *core* de la especie y sus implicaciones en la microdiversidad, evolución de la misma y su estructura poblacional.

2. Evaluación de la *pipeline*: Ensamblaje de genomas, anotación y validación.

Ensamblaje de los 6 cromosomas y 18 plásmidos de las cepas de *S.ruber*.

Se llevó a cabo el ensamblaje de los genomas de 6 cepas de *S. ruber*, dos de ellas aislados de las salinas de Campos de Mallorca (RM158, M1) y 4 de las salinas de Bras del Port de Santa Pola (P13, P18, SP38 y SP73). Durante este proceso se puso a punto una *pipeline,* detallada en detalle en el apartado de material y métodos, que permitiera el ensamblaje completo de estos genomas procariotas empleando un único tipo de librería (figura 3M, material y métodos), en este caso Illumina Miseq *paired ends*, con un tamaño de inserto similar en todas las cepas y una longitud de lectura de secuencia de 100pb (**tabla C2.1**).

En el anexo (**tabla S2.1)** se muestran los resultados obtenidos tras el empleo de las combinaciones de filtrados de calidad y cada uno de los 3 ensambladores analizados: SOAPdenovo (Luo *et al*., 2012), IDBA-UD 1.1.0 (Peng *et al*., 2012), JR-Asembler (Chu *et al*., 2013). Estos ensambladores se seleccionaron a la vista de los resultados presentados por publicaciones recientes con ensayos de ensamblajes de organismos procariotas (Peng *et al*., 2010; Magoc *et al*., 2013., Gurevich *et al*., 2013), en los cuales se mostraba un **N50** y **N90** mayor empleando ensambladores como IDBA-UD o SOAPdenovo. El ensamblador que dio los mejores resultados fue IDBA-UD salvo excepciones como la cepa M1, para la que el mejor ensamblado se obtuvo con SOAPdeNOVO. En ocasiones los mejores ensamblajes se obtuvieron con un mismo ensamblador, pero partiendo de diferentes niveles de filtrado. Tras el ensamblaje, y ya en

una segunda etapa, se reordenaron y fusionaron los *contigs* resultantes combinando los ensamblajes seleccionados con la ayuda de un genoma de referencia (**figura C2.1, anexo figura**

Tabla C2.1. Características de los *reads* secuenciados en el ensamblaje de las cepas de *S. ruber*. La tabla muestra el %GC promedio de las secuencias y su longitud, el *insert size* y los valores de cobertura teóricos si se mapearan los *reads* sin filtrar contra el cromosoma de la cepa M8.

Cepa	Insertsize (bp)	%GC	Reads totales	Tamaño (bp)	Cobertura teórica (X) *
M1-R1	530	65	24657031	101	649,73
M1-R2			24657031	101	649,73
P18-R1	530	65	24123887	101	635,68
P18-R2			24123887	101	635,68
RM158-R1	550	65	23978120	101	631,84
RM158-R2			23978120	101	631,84
SP38-R1	540	65	25433017	101	670,18
SP38-R2			25433017	101	670,18
SP73-R1	530	64	26937140	101	709,81
SP73-R2			26937140	101	709,81

*Definida como el número de bases alineadas por posición del genoma de referencia.

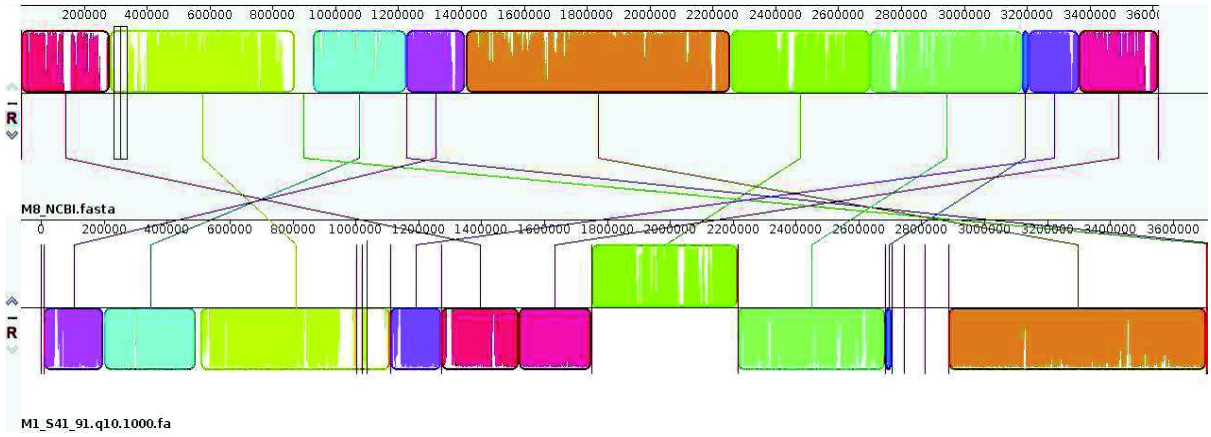


Figura C2.1. Detalle del trabajo de reordenamiento de *contigs* del mejor ensamblaje de la cepa *S. ruber* M1, obtenido con SOAPdeNovo para un intervalo de Kmer 41-91 empleando el programa Mauve. En la figura se muestra el mapeo de los *contigs* de M1(bloques inferiores) contra el cromosoma de referencia de *S. ruber* M8. Aquellos *contigs* representados bajo la línea se invirtieron, y los que no presentaron sintenia se seleccionaron como candidatos a formar parte de plásmidos.

S2.1). Tras la combinación de enamblajes se redujo el número de *contigs* en un 30% aproximadamente. El número de contigs por unir resultante varió según la cepa desde los 7 de M1, la más completa, a los 26 de RM158 y P18 (**tabla C2.2**). En una tercera etapa se completaron los *gaps* restantes no ensamblados mediante mapeo y extensión a partir de los extremos de los *contigs* resultantes de la fusión anterior. En todos los casos se consiguió completar y recircularizar todos los replicones validando la *pipeline* empleada como alternativa a la construcción y secuenciación de varias librerías para obtener genomas completos. Este tipo de aproximación, además de abaratar costes de secuenciación permite obtener genomas ensamblados de alta calidad de secuencia, dado que las Ns o regiones mal ensambladas en los extremos se eliminan durante las etapas 2 y 3. La mayoría de gaps no ensamblados y completados en la etapa 3 fueron de menos de 2 kb, aunque en algunos casos se completaron regiones de hasta 12 kb. La longitud total de la secuencia construida por mapeo y extensión fue de entre 15 y 60 kb por cepa. Tal como apuntan revisiones recientes **(**Scott y Ely., 2014), los ensambladores de nueva generación permiten ensamblar un porcentaje considerable de

Tabla C2.2. Características de los *contigs* resultantes de la etapa de fusión de los mejores ensamblajes. La tabla muestra el número de *contigs* resultante para cada cepa, diferenciando entre los que potencialmente formarán parte del cromosoma y los candidatos a constituir plásmidos en caso de recircularizar y contener origen de replicación propio. La columna Ns muestra el número de bases indeterminadas.

Cepa	*Contigs* potenciales cromosomales	Tamaño del ensamblado (pb)	%GC	N50	N90	Ns	Tamaño del *contig* mayor (pb)
M1	7	3.504.470	66,33	915648	478904	0	1085213
P13	14	3.576.069	66,17	410771	117535	0	659263
P18	26	3.741.260	66,19	225187	59029	0	671772
SP38	27	3.729.980	66,03	319408	84782	0	597512
SP73	11	3.577.792	66,19	477880	137103	0	1027222
RM158	26	3.896.114	65,62	229760	67969	0	497554
Cepa	*Contigs* potenciales extracromosomales	Tamaño del ensamblado (pb)	%GC	N50	N90	Ns	Tamaño del *contig* mayor (pb)
M1	6	65.106	59,29	66181	24026	0	116508
P13	5	584.263	56,04	33485	5484	0	33485
P18	14	1.210.166	62,17	137252	40818	0	137252
SP38	25	411.551	59,38	24284	6521	392	76335
SP73	34	1.124.831	60,81	96087	14262	0	127449
RM158	30	1.210.166	61,75	59337	31746	0	177459

secuencia, más del 98% en todas las cepas secuenciadas en este capítulo, aunque no la completan totalmente, generando la necesidad de estrategias alternativas. En este trabajo confirmamos que las regiones conflictivas se encuentran en zonas con un alto número de secuencias repetidas, ya sean repeticiones en tándem o regiones derivadas de duplicación reciente, que en la mayoría de los casos involucraron elementos transponibles. El cierre de todos los *gaps* y ensamblaje completo de los 19 replicones de las 6 cepas confirma esta *pipeline* como una solución técnica y novedosa al problema planteado, eliminando los gastos de secuenciación derivados de una segunda librería.

Validación del ensamblaje.

Durante la primera etapa de ensamblaje se obtuvieron *contigs* redundantes y sinténicos por diferentes filtrados. En los casos en que se empleó GapCloser (**Li *et al*., 2010**) para reemplazar bases indeterminadas (Ns), se comprobó que el reemplazamiento fue el correcto comparando la región con la del fragmento equivalente de otro ensamblaje sin Ns. Durante la fusión de *contigs* de diferentes ensambladores se unieron aquellos extremos solapantes idénticos, eliminando en ocasiones previamente las últimas posiciones ricas en Ns o que interrumpieron la extensión por un ensamblaje erróneo. Una anotación rápida preliminar de los *contigs* en RAST (del inglés, *the Rapid Annotation Server*) (Aziz *et al*, 2008) confirmó el solapamiento de los extremos al contener la misma secuencia de ORFs y, en el caso de extremos erróneos, su confirmación por la ausencia de estas.

En el caso de la cepa M1 dispusimos adicionalmente de una librería de pirosecuenciación 454 Life Sciencies (Roche) *paired end* de 700 pb con una tamaño de inserto de 3 kb cedida por el Dr. Álex Mira. El mapeo de estas secuencias contra la referencia de M1 al final de la etapa 2 sirvió de control para las etapas 1 y 2 del ensamblaje. Se observó como los *reads* que mapeaban en el extremo de un *contig* lo hacían en el del colindante. Esto confirmó que la mayoría de *gaps* sin ensamblar tenían una longitud inferior a las 2 kb y que la reorientación de *contigs* cromosomales, previa al mapeo y extensión, también era la correcta. Además permitió comprobar que aquellos *contigs* mayores de 10 kb que no mapeaban con el genoma de referencia pertenecían en su mayoría a plásmidos. Por último, y tras el mapeo y extensión, se compararon 2

a 2 todos los extremos extendidos confirmando que el *best hit*, región solapada, se daba con el extremo del *contig* vecino.

La identificación de *contigs* que constituían los plásmidos de las cepas secuenciadas se realizó siguiendo las comprobaciones detalladas en material y métodos (apartado2.2.1). Tras calcular las coberturas de los *reads* empleados en el ensamblaje contra sus replicones ensamblados se apreció como los plásmidos de menor tamaño presentaban coberturas de más de 4 veces superiores a los cromosomas de la propia cepa (anexo, **tabla S2.2**) y los de mayor tamaño mostraron una estequeometría 1:1 o menor. Este último dato podría indicar que incluso dentro de una misma cepa es posible observar diferencias en la estequiometría de los replicones plasmídico a pesar de la presencia de proteínas como *par*A que, como se verá más adelante, está encargada de la regulación del proceso de segregación durante la división celular (Dmowsky y Jagura-Burdzy 2013; Iestwaart 2014). Este fenómeno también podría deberse a la poliploidía del cromosoma en fase exponencial, fenómeno observado en otros organismos halófilos (Breuert *et al*., 2006), aunque los megaplásmidos también podrían ser poliploides. Estas diferencias de cobertura son además una prueba confirmativa de la presencia y el correcto ensamblaje de replicones adicionales a los cromosomas principales, ya que los *contigs* que formaron parte de plásmidos tenían niveles similares entre si y a menudo distintos a los de los *contigs* cromosomales.

Predicción de genes y anotación de los genomas.

El empleo de un sistema de predicción de ORFs común y una anotación lo más completa posible resulta crucial en el desarrollo de estudios de genómica comparativa **(**Kimes *et al*., 2014**).** Con el objetivo de obtener una anotación lo más completa y actualizada posible se anotaron los genomas ya ensamblados en las plataformas Integrate Microbial Genomes (IMG) (*Joint Genome Institute*) (Markowitz *et al*., 2014) y RAST (Aziz *et al*., 2008). La anotación y predicción de ORFs obtenidas mediante la primera plataforma fueron las más completas, por lo que se escogió para compararla con la existente hasta la fecha para las cepas M8 y M31. La predicción de ORFs por la plataforma IMG fue mejor en los dos casos, recuperándose 50 ORFs más en el caso de M8, aunque se excluyeron 118 genes identificados en la anotación original. En el caso de M31 se

localizaron más de 300 ORFs respecto a la anterior anotación, entre estas últimas más de 200 ortólogos con M8. La nueva anotación también fue más completa y permitió reducir el número de proteínas hipotéticas de un 35,32% a un 24,07% (de 1091 a 786 genes) en el caso de M8 y de un 34,46% a un 24,28% (862 a 746 genes) en M31. La mejora en la anotación se puede atribuir en gran parte al incremento notable en el contenido de las bases de datos desde la secuenciación de los genomas de M8 y M31 y, en algunos casos, una mejor predicción de los marcos de lectura. Los mapeos de los *reads* de expresión contra los genomas de M8 y M31 permitieron confirmar que los nuevos genes predichos por la anotación del JGI mapeaban en regiones expresadas significativamente y confirmados por Genemarks (Besemer y Borodovsky 2005). La nueva anotación en KEGG, COG y PFAM para los 8 genomas (anexo, **tabla S2.3**) se completó con 118 genes identificados en la anotación original para la cepa M8, y no incluidos en la nueva, cuyo marco de lectura se confirmó con los datos de expresión y Genemarks tal como se explica en detalle en material y métodos (apartado 2.2.2). Algunos de estos 118 ORFs se identificaron en el resto de genomas mediante comparación de secuencias con el programa Exonerate (Slater y Birney 2005). El genoma de P18 fue el que menos genes de los 118 se localizaron (46) y el de RM158 en el que más (75) (anexo, **tabla S2.4**). Como resultado se añadieron 19 nuevos genes al genoma *core* de la especie. Además la anotación de M8 se completó con la identificaron manual de los codones de inicio y de final de transcripción de 15 genes no identificados por ninguna de las dos predicciones anteriores, pero si con los mapeos de los *reads* de RNAseq.

3. Características generales de los genomas.

Tras el ensamblaje de los 6 nuevos genomas, sumados a los dos ya descritos para la especie, se completaron un total de 26 replicones (8 cromosomas y 18 plásmidos) (**figura C2.3**). Los cromosomas de las cepas secuenciadas hasta la fecha presentaron tamaños que oscilaron de las 3,55 Mb en el caso de M31, cepa tipo, a las 3,76 Mb de SP73. Se anotaron y detectaron un total de 24.516 genes con un tamaño promedio de 1012,5 nucleótidos (la **figura C2.2),** superior al promedio de 924 nucleótidos estimado previamente para un conjunto de 80 genomas procariotas (Xu *et al*., 2006). La ORF de menor tamaño tuvo 117 nucleótidos de longitud y la de mayor 7.770. Tan sólo 48 ORFs en todos los genomas, entre 5 y 8 por genoma, presentaron una

longitud mayor de 5kb, lo que supone el 0,20% del total de ORFs de los genomas anotados, coincidiendo con estudios previos en los que se analizó la presencia de ORFs en 580 genomas procariotas (Reva y Tümmler 2008). El **PI** promedio de los proteomas de las 8 cepas fue de 5,77. Este valor de PI se encuentra dentro del rango calculado para organismos halófilos (Paul *et al*., 2008) e intermedio al calculado para *Archaea* halófilas como *Halobacterium sp.* y *Haloarcula marismortui* (en torno 4,6), y bacterias no halófilas como *Chlorobium tepidum* y *Bacteroides*

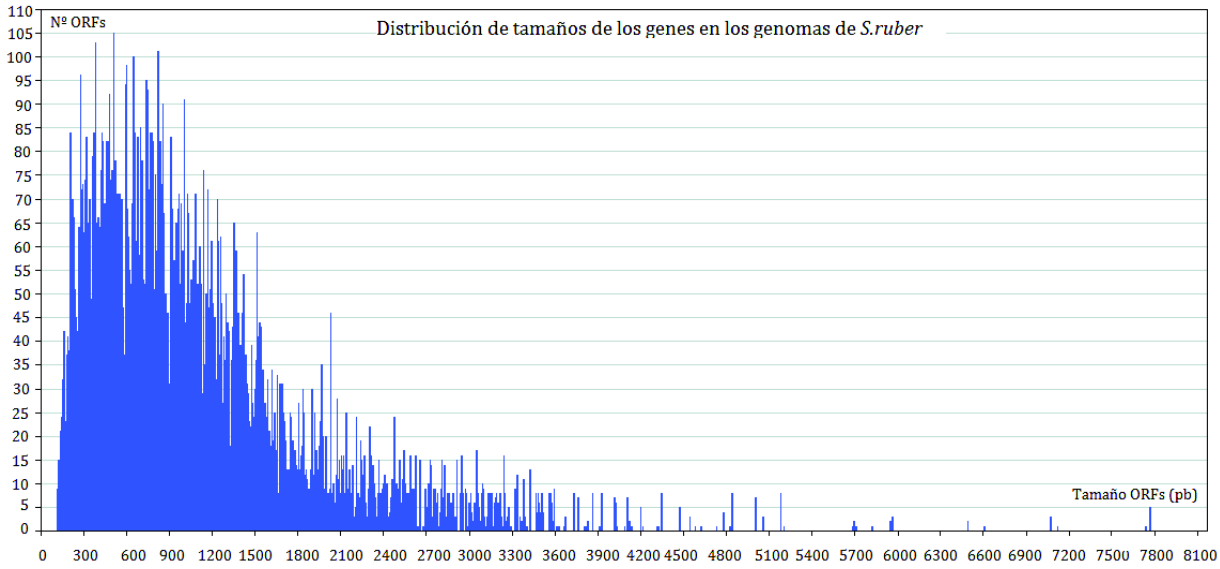


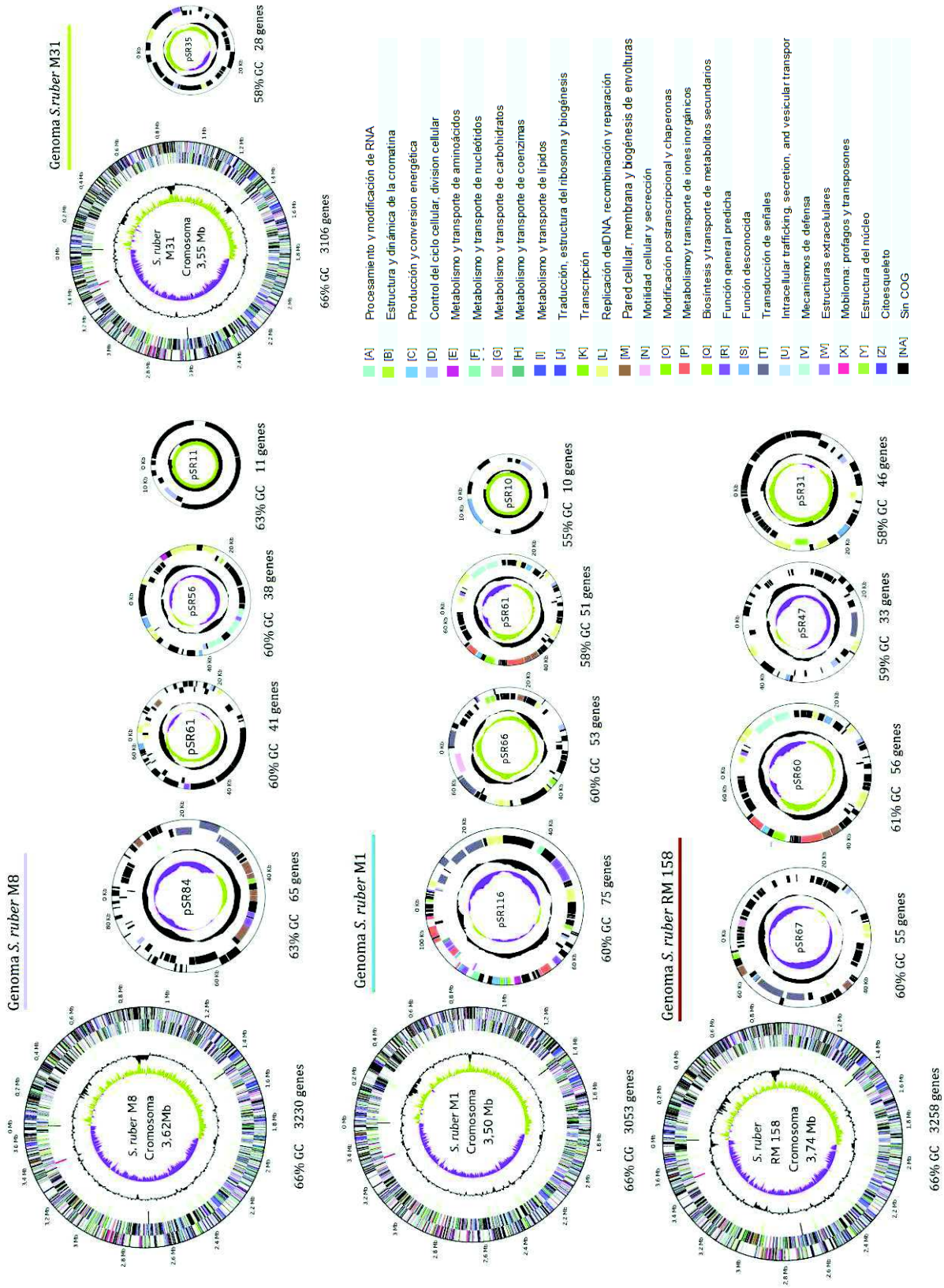
Figura C2.2. Detalle de la distribución de genes por tamaños en los 8 genomas completamente secuenciados de las cepas de *S. ruber*. La mayoría de ORFs mostraron tamaños por debajo de las 2.000 pb.

fragilis (alrededor de 7) (Joo y Kim, 2005; Mongodin *et al*., 2005). Este proteoma ácido se ha observado en *Archaea* halófilas que emplean la estrategia "*salt in*" acumulando K^+ en su citoplasma como mecanismo para equilibrar la presión osmótica y que contienen un porcentaje elevado de residuos ácidos (Asp/Asn, Glu/Gln) y bajo de básicos (Lys) e hidrofóbicos (Fukuchi *et al*., 2003; Oren 2013; Reed e*t al*., 2013). *S. ruber* emplea esta misma estrategia, conteniendo elevadas concentraciones de iones K^+ y Cl^- en su citoplasma (Oren *et al*., 2002), y cumple este perfil proteómico como se muestra en la **tabla C2.3**. En esta tabla aparecen marcados en negrita un subgrupo de diez aminoácidos (Ala, Asp, Glu, Gly, Ile, Leu, Pro, Ser, Thr, and Val) considerados prebióticos y posiblemente la principal fuente de aminoácidos antes de la

emergencia de rutas biosintéticas complejas en el origen de la vida. (Longo *et al*., 2012). Recientemente se han detectado las características mencionadas en proteínas halofílicas (gran porcentaje en residuos ácidos y pobre en hidrofóbicos) al construir proteínas prebióticas con estos aminoácidos (Longo *et al*., 2013). Las proteínas resultantes presentaron características halofílicas y se plegaron correctamente a elevadas concentraciones de sal. Estos hechos sugieren que los ambientes halófilos pudieron jugar un papel importante en los orígenes de la vida. De los diez aminoácidos, ocho se encuentran en mayor proporción en el proteoma de *S. ruber* que en organismos no halófilos como *B. fragilis*, *E. coli* o halófilos moderados como *Halomonas o termófilos como S. islandicus*.

Tabla C2.3. Distribución en la frecuencia y composición de aminoácidos de los proteomas y el promedio para *H. marismortui* ATCC-43049 (Balinga *et al*., 2004), *H. volcanii* DS2 (Hartman *et al*., 2010*)*, *B. fragilis*YCH46 (Kuwahara *et al.,* 2004), *Halomonas elongata* DSM 2581 (Schwibbert *et al*., 2010) y *S. islandicus* LS.2.15 (Reno *et al*., 2009) y el promedio para las cepas de *S.ruber*. La tabla muestra la distribución porcentual de cada aminoácido, destacando en negrita los 10 prebióticos analizados según (Longo *et al*., 2013).

Aminoácido	*H. volcanii*	*H. marismortui*	*S. ruber*	*B. fragilis*	*H. elongata*	*S. islandicus*
Alanina (Ala, A)	11,1%	10,5%	10,7%	6,9%	11,2%	5,6%
Cisteína (Cys, C)	0,7%	0,7%	0,6%	1,3%	0,9%	0,6%
Aspartato (Asp, D)	8,5%	8,4%	7,1%	5,4%	6,1%	4,8%
Glutamato (Glu, E)	8,0%	8,1%	7,1%	6,5%	6,7%	6,9%
Fenilalanina (Phe, F)	3,5%	3,3%	3,4%	4,6%	3,4%	4,3%
Glicina (Gly, G)	8,6%	8,3%	8,3%	6,8%	8,4%	6,5%
Histidina (His, H)	2,0%	2,0%	2,2%	1,9%	2,5%	1,3%
Isoleucina (Ile, I)	3,8%	4,4%	3,7%	7,0%	4,5%	9,7%
Lisina (Lys, K)	2,0%	2,0%	2,1%	6,6%	2,4%	7,7%
Leucina (Leu, L)	9,1%	8,9%	9,8%	9,3%	11,4%	10,3%
Metionina (Met, M)	1,8%	1,8%	1,9%	2,7%	2,5%	2,2%
Asparagina (Asn, N)	2,4%	2,6%	2,5%	5,0%	2,4%	5,1%
Prolina (Pro, P)	4,6%	4,5%	5,5%	3,8%	5,0%	3,9%
Glutamina (Gln, Q)	2,4%	3,1%	3,6%	3,4%	3,6%	2,1%
Arginina (Arg, R)	6,6%	6,0%	7,6%	4,8%	7,6%	4,4%
Serina (Ser, S)	5,8%	5,9%	5,8%	6,2%	5,4%	6,7%
Treonina (Thr, T)	6,2%	6,9%	6,2%	5,6%	5,0%	4,8%
Valina (Val, V)	9,3%	8,7%	8,0%	6,4%	7,2%	7,3%
Triptófano (Trp, W)	1,1%	1,1%	1,3%	1,3%	1,5%	1,0%
Tirisina (Tyr, Y)	2,7%	2,7%	2,7%	4,5%	2,3%	4,8%


Genoma S.ruber M31
S. ruber M31 Cromosoma 3,55 Mb
pSR35
58% GC 28 genes
66% GC 3106 genes
[A] Procesamiento y modificación de RNA
[B] Estructura y dinámica de la cromatina
[C] Producción y conversion energética
[D] Control del ciclo cellular, division cellular
[E] Metabolismo y transporte de aminoácidos
[F] Metabolismo y transporte de nucleótidos
[G] Metabolismo y transporte de carbohidratos
[H] Metabolismo y transporte de coenzimas
[I] Metabolismo y transporte de lípidos
[J] Traducción, estructura del ribosoma y biogénesis
[K] Transcripción
[L] Replicación deDNA, recombinación y reparación
[M] Pared cellular, membrana y biogénesis de envolturas
[N] Motilidad cellular y secrección
[O] Modificación postranscripcional y chaperonas
[P] Metabolismoy transporte de iones inorgánicos
[Q] Biosíntesis y transporte de metabolitos secundarios
[R] Función general predicha
[S] Función desconocida
[T] Transducción de señales
[U] Intracellular trafficking, secretion, and vesicular transpor
[V] Mecanismos de defensa
[W] Estructuras extracelulares
[X] Mobiloma: profagos y transposones
[Y] Estructura del núcleo
[Z] Citoesqueleto
[NA] Sin COG
Genoma S. ruber M8
S. ruber M8 Cromosoma 3,62Mb
pSR84
pSR61
pSR56
pSR11
63% GC 65 genes
60% GC 41 genes
60% GC 38 genes
63% GC 11 genes
66% GC 3230 genes
Genoma S. ruber M1
S. ruber M1 Cromosoma 3,50 Mb
pSR116
pSR66
pSR61
pSR10
60% GC 75 genes
60% GC 53 genes
58% GC 51 genes
55% GC 10 genes
66% CG 3053 genes
Genoma S. ruber RM 158
S. ruber RM 158 Cromosoma 3,74 Mb
pSR67
pSR60
pSR47
pSR31
60% GC 55 genes
61% GC 56 genes
59% GC 33 genes
58% GC 46 genes
66% GC 3258 genes

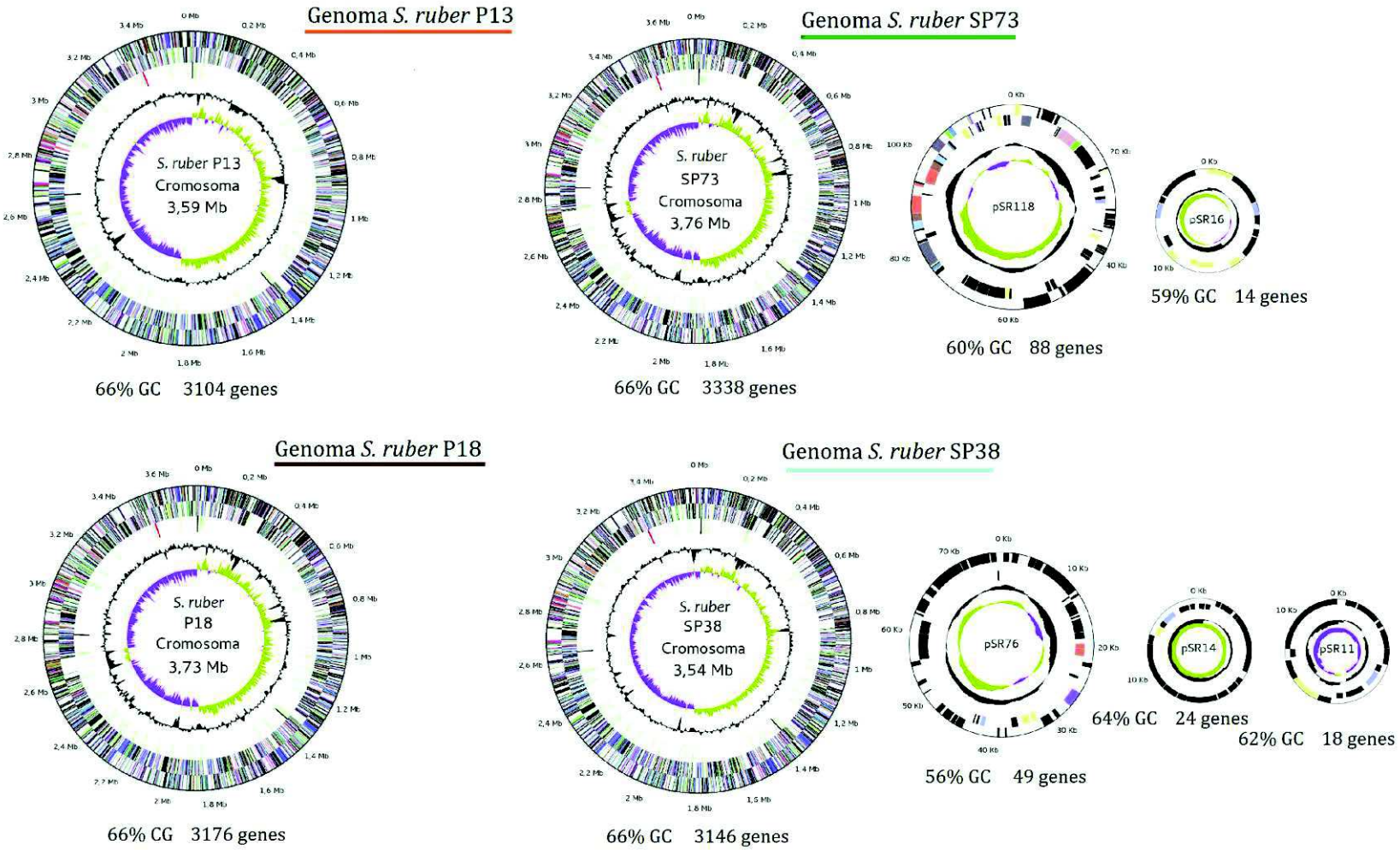


Figura C2.3. Representación circular de los cromosomas de las 6 cepas ensambladas.
De fuera a dentro: Círculos 1-2: anotación según las categorías COG. Círculo 3:tRNAs. Círculo 4: %GC. Círculo 5: GC Skew.
Para cada replicón se indica su tamaño, el genoma al que pertenece, la nomenclatura empleada, número de genes y %GC.

Los datos de expresión del estudio transcriptómico descrito en el capítulo 1 contribuyeron a la identificación de nuevas ORFs y en los casos más claros a la delimitación de operones y 5'UTR de algunos genes. Aunque una librería de expresión específica de hebra o marcada en 5' es la mejor aproximación a la hora de delimitar operones e inicios de transcripción (Croucher y Thompson., 2010), nuestra librería no específica de hebra permitió detectar mediante visualización manual de los mapeos en IGV 623 operones en la cepa *S.ruber* M8, (318 en la hebra directa y 305 en la inversa), 1018 inicios de transcripción y 18 nuevos transcritos. La mayoría de los operones incluyeron entre 2 y 4 genes (**Figura C2.4**) y en total 1800 genes se expresaron en operones en el caso de M8. Un 83,28% (508/610) de los operones predichos *in silico* en la base de datos MicrobesOnline (Dehal *et al.*, 2009) se confirmaron mediante datos experimentales, identificando 115 operones no caracterizados hasta la fecha. Estos datos reflejan el enorme potencial del RNAseq en la comprensión de los mecanismos reguladores, la asignación de genes a los diferentes regulones y la identificación de operones alternativos en

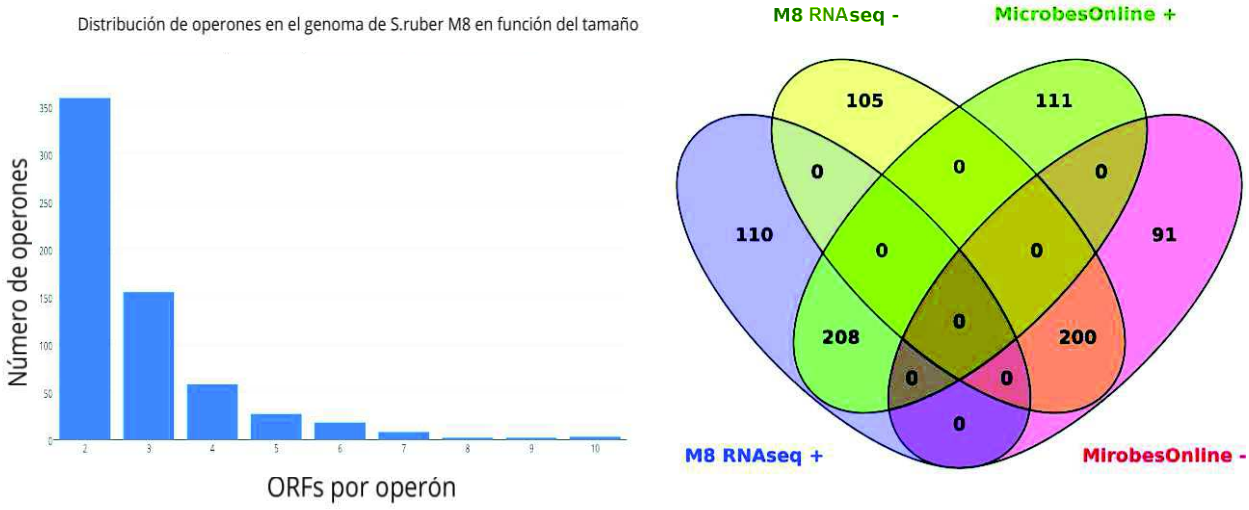


Figura C2.4. A la izquierda, detalle de la distribución de operones en función de los genes contenidos en el genoma de la cepa *S. ruber* M8. El diagrama de Venn muestra la distribución de operones en ambas hebras (+ y -) y aquellos operones confirmados por RNAseq que habían sido predichos *in silico* y se encuentran en la base de datos MicrobesOnline.

diferentes condiciones de crecimiento en organismos procariotas tal como se muestra en trabajos referentes realizados en *Helicobacter pylori* (Sharma *et al*., 2010) y *Mycoplasma pneumoniae* (Guell *et al*., 2009).

Los cromosomas presentaron un %GC y valores de distribución de tetranucleótidos (TETRA) y ANI muy parecidos entre sí, sin seguir un patrón aparente de distribución biogeográfica o época de aislamiento (**anexo, tabla S2.5)**. Estos datos concuerdan con la ausencia de patrones biogeográficos en los análisis previos realizados con patrones de restricción y PFGE (Peña *et al*., 2005) y MLSD (Roselló-Mora *et al*., 2008), en los cuales no se apreció una correlación clara en las matrices de similitud de estos marcadores y el origen de las cepas analizadas. Además, como se discutirá más adelante, estos resultados apoyan el efecto homogenizador de la recombinación homóloga sobre el genoma *core* (ver apartado 5). El alineamiento de los genomas mediante el programa Mauve (**figura C2.5**) mostró que la elevada sintenia detectada en el análisis microevolutivo de las cepas M8 y M31 (Peña *et al*., 2010) constituye una característica extensible a todas las cepas analizadas y no sólo una característica atribuible a la proximidad filogenética de dos cepas cercanas. Este fenómeno se ha descrito en otros estudios microevolutivos en especies con ecología muy diversa como las bacterias de vida libre *Phroclorococcus* sp. (Coleman *et al*., 2006), *Pelagibacter ubique* (Wilhelm *et al*., 2007), *Escherichia coli* (Mira *et al*., 2010), patógenas de relevancia clínica *S. pneumoniae* (Hiller *et al*., 2010), *S. pyogenes* (Mira *et al*., 2010), *C. trachomatis (*Joseph *et al*., 2011), *Acinetobacter baumanii* (Traglia *et al*., 2014). Esto se aprecia también incluso entre distintas especies de un mismo género como es el caso de *Alteromonas* (López-Pérez *et al*., 2014) *Listeria* (den Backer *et al*., 2010) y de géneros cercanos entre sí como *Shigella flexneri* y *E. coli* (Mau *et al*., 2006; Touchon *et al*., 2009). En todos estos casos se observan grandes bloques colineares sinténicos, interrumpidos por *indels* o islas genómicas. En algunos casos se aprecian grandes reordenamientos génicos pero el orden de los genes dentro de los bloques se mantiene. En *S. ruber* además el orden es mayor pues no se observan grandes reordenamientos génicos a excepción de detectado en una de las islas genómicas (GI) de la cepa P13 (GI5-P13). La sintenia observada al alinear todas las cepas revela una arquitectura conservada a nivel de especie, interrumpida en dos regiones situadas en las mismas posiciones relativas dentro del cromosoma y que corresponden a las dos HRVs, descritas con anterioridad en M8 y M31 (Pasic *et al*., 2009;

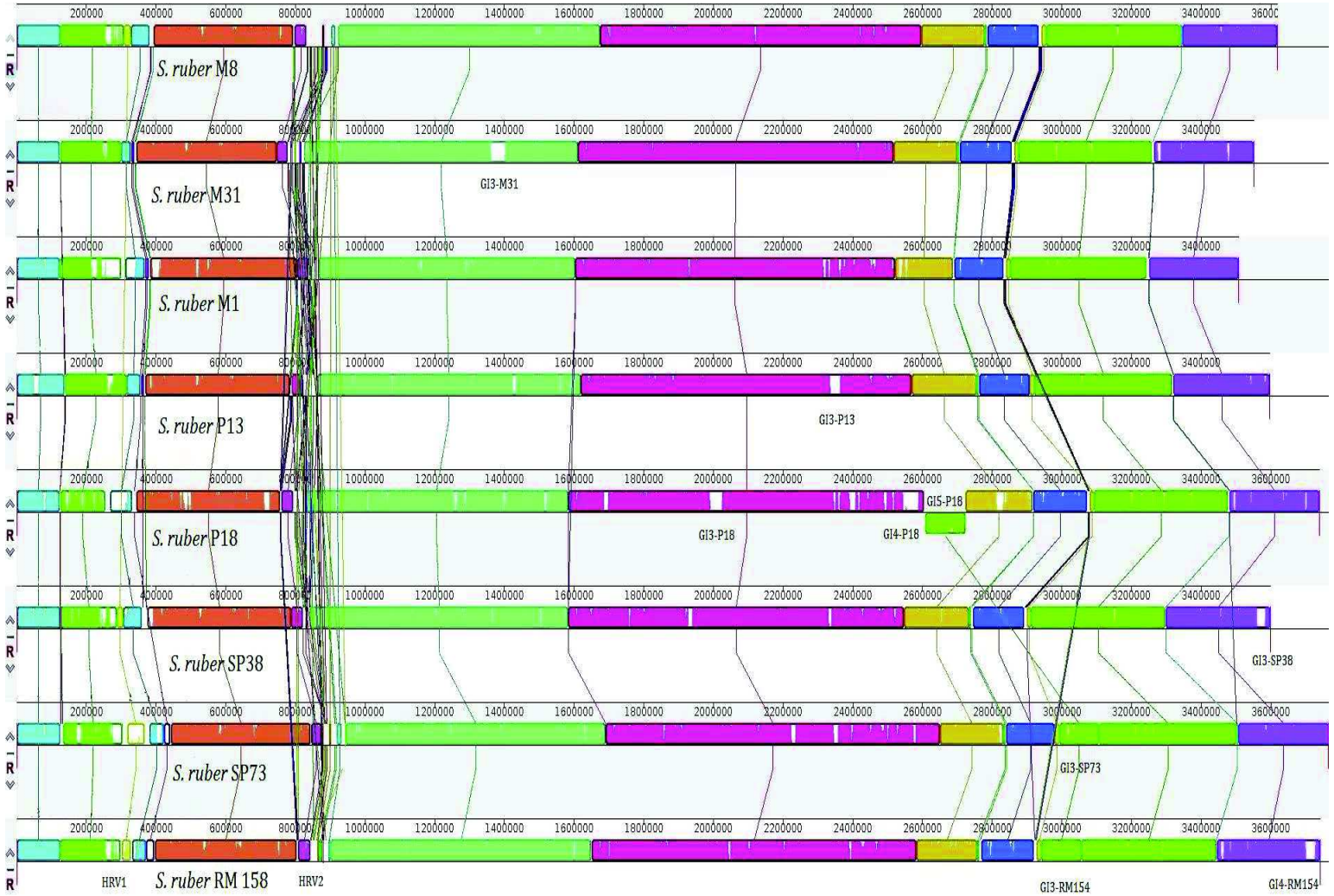


Figura C2.5. Representación del alineamiento de los 8 cromosomas de las cepas secuenciadas de *S.ruber* con el programa *progressive* Mauve. En color se muestran grandes regiones sinténicas o bloques colineares interrumpidas por las zonas himervariables (HRV1 y HRV2), islas genómicas (GI) e *indels*. La altura coloreada de los bloques refleja la elevada identidad de estas regiones, que constituyen gran parte del *core* genoma de la especie. La nomenclatura de las GIs se detalla en la tabla C2.4.

Peña *et al*., 2010). En algunos casos se observó la presencia de una o varias GI adicionales en el cromosoma e *indels* y además se detectó una gran diversidad en elementos plásmidos tal como ya indicaban análisis experimentales previos (Peña, datos no publicados).

La representación de la composición del genoma *core* y el pangenoma con el incremento de genomas secuenciados reveló que el primero de estos alcanza un valor estable con pocos genomas acumulados. El genoma *core* de la especie estaría constituido por un subconjunto de 2434 genes presente en todas las cepas analizadas (entre un 66,6 % y 77,58% del total del genoma de cada cepa) (**figura C2.6**). Gran parte de los genes del genoma *core* de la especie estarían localizados en los bloques sinténicos cromosómicos. Tan sólo una pequeña fracción del mismo se localizaría en las regiones hipervariables (15 y 18 genes en la HRV1 y HRV2 respectivamente). Estas últimas, junto a las islas genómicas y plásmidos, contendrían la mayoría del genoma accesorio tal como sucede en otras especies bacterianas de diversa ecología como *Prochlorococcus sp*. (Coleman *et al*., 2006), *P. ubique* (Wilhelm *et al*., 2007), *S. islandicus* (Cadillo-Quiroz *et al*., 2010), *E. coli* o *S. agalactiae,* incorporándose mediante transferencia horizontal (Mira *et al*., 2010). El promedio que representa el genoma *core* en las cepas de *S. ruber* (73,53%) fue similar al calculado para varias especies bacterianas en un estudio que

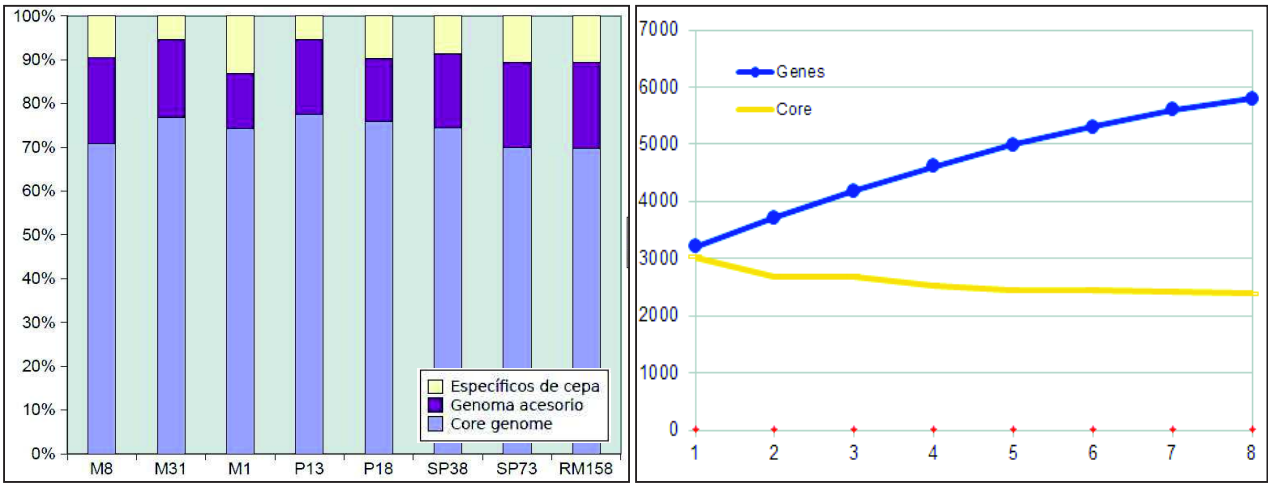


Figura C2.6. A la izquierda, detalle de las fracciones que representan el genoma core (azul), el genoma accesorio (rojo) y los genes específicos de cepa (amarillo) en cada uno de los 8 genomas ensamblaados para las cepas de *S. ruber*. La figura de la derecha representa la evolución del genoma *core* (amarillo) y el pangenoma (azul) con el incremento de cepas secuenciadas. Esta última representa un pangenoma abierto.

incluyó 573 genomas, y que rondó el 72% (Lapierre y Gogarten 2009). Los niveles de diversidad encontrados fueron muy superiores a los de organismos patógenos obligados intracelulares como *Chlamydia trachomatis* con un genoma *core* que incluye más del 90% de los genes, mayores a los de especies de vida libre como *S. islandicus*, 86% con 12 genomas (Castillo-Quiroz *et al*., 2012) y similares a los de otras como *S. báltica*, 70% con 4 genomas analizados (Caro-Quintero *et al*., 2011). En esta última especie el intercambio génico y diversificación juegan un papel importante en la evolución de la especie.

El pangenoma de la especie mostró un incremento constante dibujando una curva que se aproxima a una asíntota y que se aleja de la definida por el genoma *core* (**figura C2.6**). Esta curva asintótica revela la enorme diversidad existente a nivel de plásmidos y GI (incluyendo HRVs), describiendo un pangenoma abierto para *S.ruber.* Esto se ve reflejado en la elevada proporción de genes específicos de cepa ya que cada nuevo genoma secuenciado aporta unos 300 genes específicos de cepa, valor muy similar al observado en especies como *E.coli* con un genoma *core* de alrededor de 2800 genes para el mismo número de genomas secuenciados (Mira *et al*., 2010) o *H. influenzae* con 13 genomas (Hogg *et al*., 2007). Estas especies y otras patógenas han mostrado una enorme dinámica y diversidad de plásmidos e islas genómicas en su estrategia de adaptabilidad patogénica durante el proceso de infección y generación de diversidad policlonal (Hiller *et al.,* 2010; Roberts *et al*., 2014). La diversidad y proporción del pangenoma de *S. ruber* fue similar a la de algunos organismos patógenos como el estudio llevado a cabo con ocho genomas de *S. agalactiae*, según el cual el genoma *core* estaría constituido por alrededor del 80% del total de los genes (Tettelin *et al*., 2005). Este último estudio revela un vasto *pool* de genes accesibles para la especie, requiriendo cientos de aislados para dejar de encontrar genes específicos de cepa y saturar el pangenoma. Por otra parte, en especies con pangenomas cerrados y niveles de clonalidad elevados como *Bacillus antrhacis* se ha descrito el pangenoma completo con tan sólo 4 genomas secuenciados (Mira *et al*., 2010).

En conjunto, la aparente dinámica y diversidad del genoma accesorio en *S. ruber*, incluyendo islas genómicas (GI), entre ellas las regiones hipervariables (HRVs), e *indels*, frente al orden y la elevada identidad de las regiones sinténicas que albergan la mayoría de su *core* genoma, nos llevó a abordar por separado el análisis de la microdiversidad y los mecanismos que

la regulan en cada una de estas dos fracciones genómicas (genoma accesorio en el apartado 4 y genoma *core* en el apartado 5).

4. Descripción de la microdiversidad y mecanismos que actúan sobre el genoma accesorio de *S. ruber*.

La mayoría de análisis genómicos comparativos en los que se ha observado la presencia de un pangenoma abierto compararon aislados de puntos geográficos distintos. *S. ruber* presenta este mismo comportamiento incluyendo coaislados: caso de las cepas M1, M8 y M31 coaisladas en Mallorca (1999); P13 y P18 este mismo año en las Salinas de Bras del Port (Santa Pola), SP38 y SP37, aisladas de esta misma salina en el año 2007 y por último RM158, aislada en Mallorca en 2009 (véase material y métodos, apartado 1.2, **tabla 1M**). La ausencia de clonalidad en el genoma accesorio de la especie, incluso entre coaislados muy próximos filogenéticamente, y su ubicación en plásmidos e islas genómicas muestra la enorme microdiversidad existente y sugiere un papel relevante de los mecanismos de trasferencia horizontal sobre los genomas de *S.ruber*. Estudios recientes han encontrado niveles elevados de diversidad en cepas coaisladas de la misma especie, como es el caso de 9 cepas de *A. macleodii* (Gonzaga *et al*., 2012., López-Pérez *et al*., 2013), 12 genomas secuenciados de *S. islandicus* procedentes de una misma fuente termal situada en la península de Karmchatka (Rusia) (Cadillo-Quiroz *et al*., 2014) o 17 cepas de *S. pneumoniae* aisladas de un mismo paciente tras una infección de carácter policlonal (Hiller *et al*., 2010).

4.1- Zonas hipervariables, islas genómicas (GI) e *indels*.

Los alineamientos de los genomas permitieron detectar aquellas regiones en las que se interrumpía la sintenia, detectando las principales GI e *indels*, elementos que, a parte de los plásmidos, albergan normalmente entre el 60-80% de los genes del genoma accesorio o flexible (Mira *et al.*, 2010., Gonzaga *et al*., 2012., López-Pérez *et al*., 2013) y juegan un papel importante en la adaptación y evolución bacteriana (Bellanguer *et al*., 2013). Las GI detectadas tras la comparación presentaron una longitud entre 10 y 200 kb (**tabla C2.4**), valores típicos para este tipo de elementos génicos, ya que por debajo de las 10kb las inserciones de regiones específicas

Tabla C2.4. Distribución de las GIs en los 8 genomas de las cepas secuenciadas de *S.ruber*. En cada caso se muestra la posición, genes contenidos,% GC y representación del total del genoma (%). SP: Cepas aisladas en las salinas de Santa Pola ; M: Cepas aisladas en las salinas de Campos de Mallorca.

Rasgo genómico analizado		*Genoma analizado*							
		M8 (M1999)	M31 (M1999)	M1 (M1999)	P13 (SP1999)	P18 (SP1999)	SP38 (SP2007)	SP73 (SP2007)	RM158 (M2006)
HRV1 (fGI)	ORFs	127	111	151	139	87	145	205	127
	Tamaño (pb)	151.717	134.119	177.852	140.555	109.177	168.367	196.160	156.467
	Posición inicio	249.507	223.961	235.218	246.865	237.245	230.386	256.113	249.611
	Posición final	401.287	358.080	413.070	387.420	346.422	398.753	452.273	406.078
	% GC	61,3	62,1	60,5	61,9	62,7	62,5	60,4	60,7
	% del genoma	3,96	3,74	4,72	3,90	2,92	4,55	5,03	4,03
HRV2 (fGI)	ORFs	134	82	44	77	50	42	103	93
	Tamaño (pb)	129.923	88.135	51.245	74.900	56.785	50.096	104.730	94.722
	Posición inicio	829.018	774.703	841.588	819.096	793.772	817.511	872.601	840.001
	Posición final	958.941	862.838	892.833	893.996	850.557	867.607	977.331	934.723
	% GC	58,2	59,1	62,2	60,8	62,4	62,5	58,6	58,7
	%del genoma	3,39	2,46	1,36	2,08	1,59	1,35	2,69	2,44
HGT-GIs			**GI3-M31**		**GI3-P13**	**GI3-P18**	**GI3-SP38**	**GI3-SP73**	**GI3-RM158**
	ORFs								
	Tamaño (pb)		42.752		27.300	33.800	23.800	113.670	113.682
	Posición inicio		1.360.489		2.335.400	1.990.500	3.560.400	2.993.000	2.943.003
	Posición final		1.404.241		2.362.700	2.024.300	3.584.200	3.106.670	3.056.685
	% GC		55,8		50,6	58,7	51,0	61,4	61,5
	%del genoma		1,19		0,76	0,90	0,64	2,91	2,93
						GI4-P18			**GI4-RM158**
	ORFs								
	Tamaño (pb)					43.500			27.640
	Posición inicio					2.544.500			3.700.270
	Posición final					2.588.000			3.727.910
	% GC					66,0			57,6
	%del genoma					1,16			0,72
						GI5-P18			
	ORFs								
	Ta maño (pb)					114.610			
	Posición inicio					2.608.780			
	Posición final					2.723.390			
	% GC					61,4			
	%del genoma					3,06			

o accesorias se consideran *indels* (Juhas *et al.*, 2008; Bellanguer *et al*., 2013). El número y tamaño de las mismas fue variable en los diferentes genomas, con tan sólo dos en M8 y hasta cinco en el caso de P18, valores habituales según el rango definido en un estudio realizado con 675 genomas procariotas (Fernández-Gómez *et al*., 2012). La proporción del genoma representada por las GI osciló entre 6,1% y un 10,6%, una proporción muy elevada en comparación a otros *Bacteroidetes* y clases de bacterias, en las que suele encontrarse entre el 2% y el 5%. Tan sólo algunos casos puntuales como en *E. coli* se han detectado valores mayores en torno al 17% (Ochman *et al*., 2000). La variabilidad en el tamaño de las GI parece ser una característica a las distintas especies de la clase *Bacteroidetes* (Fernández-Gómez *et al*., 2012). Esta característica parece extenderse a *S. ruber* ya que, pese a haber comparado coiaslados, entre las 8 cepas de *S. ruber* analizadas existe una gran variabilidad génica a nivel de GI como se verá más adelante.

Las GI detectadas en este trabajo mostraron características típicas de este tipo de elementos génicos: un contenido en GC y un CAI para los genes albergados inferior al del resto del genoma tal como se había observado previamente en las GI descritas en los genomas de M8 y M31(Pasic *et al.*, 2009; Peña *et al.*, 2010). Además estás regiones, tal como se observó ya en la cepa tipo M31 (Pasic *et al*., 2009), estuvieron infrarrepresentadas en metagenomas de ambientes halófilos extremos como los de las salinas de San Diego (alta salinidad) **(**Rodríguez-Brito *et al*., 2010) o Santa pola (cristalizador CR30) (Ghai *et al*., 2012) (**figura C2.21)**. El menor reclutamiento de secuencias en estas regiones indica que su representación es inferior a las regiones sinténicas que forman parte del genoma *core*, y que contienen una proporción importante de genes específicos de cepa. Estos reclutamientos, realizados de manera análoga en otras especies de halófilos como *H. walsbyi* (Cuadro-Orellana *et al*., 2007, Tully *et al*., 2015) confirman la presencia de este tipo de GI.

Basándonos en su estructura y dinámica, es posible clasificar las GI detectadas en *S.ruber* en dos grupos. El primero incluye las denominadas HRVs, presentes en las 8 cepas de *S. ruber.* Estas GI cumplieron algunas características comunes: 1) se localizaron en regiones similares en los 8 genomas analizados (**figura C2.5, figura C2.7, tabla C2.4**), 2) se situaron entre los mismos genes del *core* genoma y 3) contuvieron genes comunes o con funciones similares. Este tipo de isla genómica, muy común en *Bacteroidetes*, no suele estar flanqueada además por

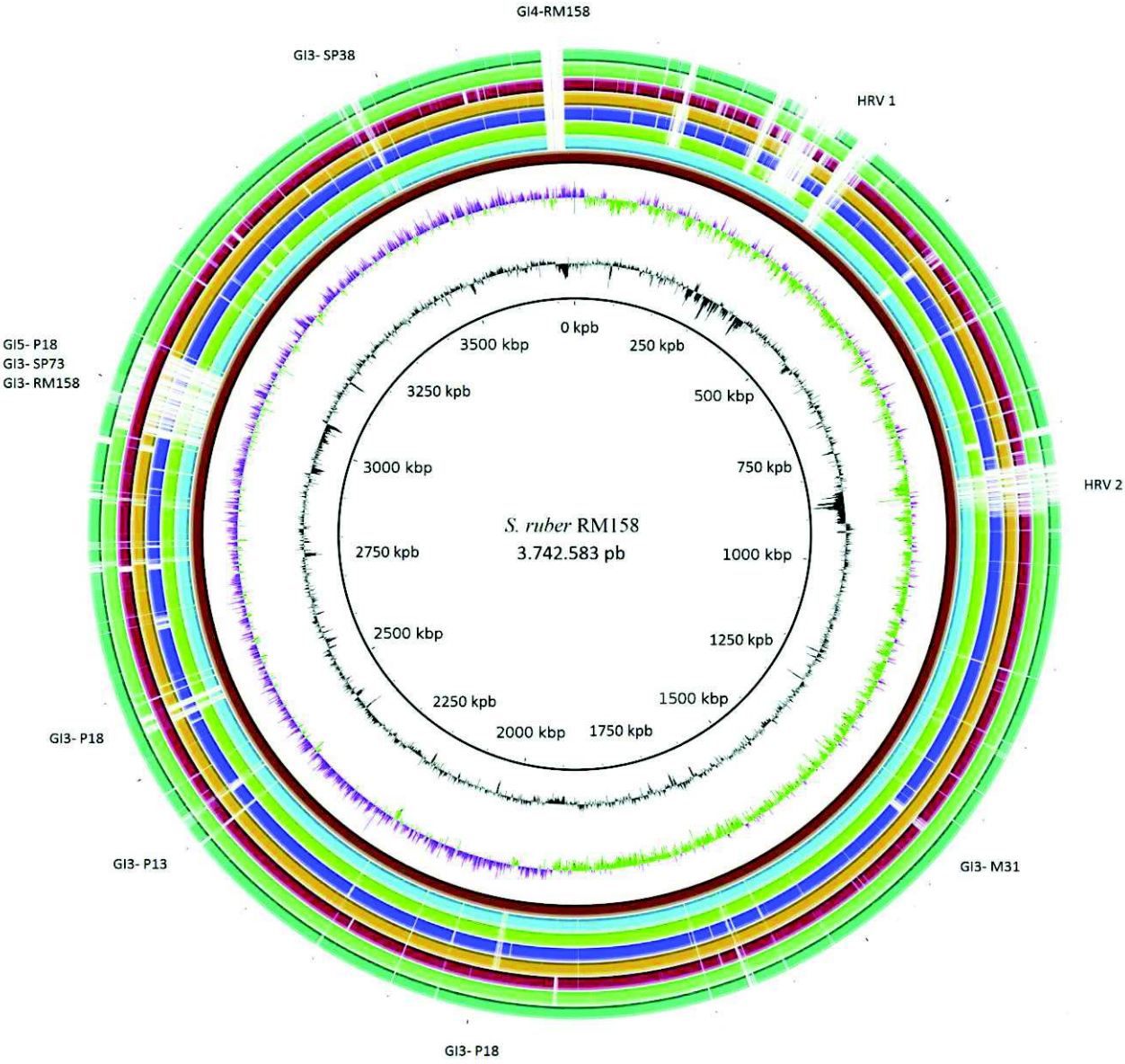


Figura C2.7. Alineamiento de los genomas de las 8 cepas de S. ruber de fuera a dentro: SP73 (verde azulado), SP38 (verde claro), P18 (ocre), P13 (naranja), M1 (azul oscuro) M31 (verde claro) y RM158 como referencia (marrón). Los anillos centrales muestran el GC *skew* y el contenido en GC. Las regiones sin colorear dentro de cada anillo reflejan las zonas no sinténicas correspondientes a las GI detectadas en las diferentes cepas y que se indican en la figura siguiendo la nomenclatura de la tabla C2.4.

tRNAs ni secuencias repetitivas en tándem, y se conocen como islas genómicas flexibles (fGI) o (HR-GIs) (Fernández-Gómez *et al*., 2012; Gonzaga *et al*., 2012; López -Pérez *et al*., 2014). Dentro de las fGI, podemos encontrar dos tipos a su vez: (1) Las fGIs aditivas, en las cuales se encuentran *clusters* conservados de genes flanqueados de transposones, transposasas e integrasas de virus y por lo que estarían asociadas a elementos móviles como virus y (2) fGI de reemplazamiento, en las cuales encontramos *clusters* de genes totalmente diferentes entre cepas comparadas aunque con funciones similares y normalmente implicados en estructuras celulares esenciales como envueltas. El segundo grupo, conocido como HGT-GI, incluye GI presentes sólo en algunas cepas de la especie o específicas de cepa, normalmente flanqueadas por regiones repetitivas o tRNAs y con un elevado contenido en proteínas virales y transposones. Este segundo grupo incluiría el resto de GI detectadas en *S. ruber* (**tabla C2.4**)

Islas genómicas flexibles (fGIs). Las HRVs de *S.ruber* se asemejan por sus características a 3 de las 5 islas detectadas en *Prochlorococcus sp*. (Coleman *et al*., 2006), y a las descritas en *H. walsbyi* (Cuadros-Orellana *et al.*, 2007) y *A. macleodii* (López-Pérez *et al*., 2013, 2014). La HRV1 mostró un tamaño mayor a la HRV2 (en promedio 154,3 Kb y 136.5 ORFs para la HRV1; 81,3 Kb y 78,7 ORFs para la HRV2) (**tabla C2.4**). Ambas albergan una elevada microdiversidad, un porcentaje elevado de genes del genoma accesorio (**figura C2.7, tabla C2.4**) y aportan respectivamente 13 y 15 genes al genoma *core* de la especie, situados en *clusters* sinténicos como se comentará más adelante.

Funcionalmente, una característica remarcable de estas zonas fue el enriquecimiento significativo de genes pertenecientes a las categorías COG M (pared celular, membrana y biogénesis de envolturas), especialmente en la HRV2, y de la categoría COG L (replicación, recombinación y reparación) ($p<0,05$, $FDR<0,05$) (**figura C2.8, tabla S2.6**), confirmando que las características funcionales de estas GI observadas para las cepas M8 y M31 (Pasic *et al*., 2009; Peña *et al*., 2010) son extensibles a la especie. Dentro de la categoría COG L, tal como se observó en estudios anteriores en GI de varias bacterias marinas (Fernández-Gómez *et al.*, 2012), se detectaron gran cantidad de elementos móviles, en su mayoría transposasas que no presentan homólogo en otras cepas. Estos mismos estudios identificaron una proporción elevada de genes relacionados con componentes de membrana y envoluras celulares. Estas regiones, tal como sucede en elementos móviles como plásmidos, contuvieron una proporción elevada (58%) de

genes sin función COG asignada, valores próximos al promedio detectado para la mayoría de GI de procariotas patógenos (53%) (Hsiao *et al*., 2005) y de marinos de vida libre (55-60%) (Fernández-Gómez *et al.*, 2012). Dentro de estos últimos, en el 70% de los casos se aprecia un enriquecimiento significativo de HP en GIs, aunque no en especies de *Bacterioidetes* donde la proporción de estos genes también es elevada en el resto del genoma tal como sucede en *S.ruber*. La elevada proporción de HP presentes en GI se corresponde con la presencia de nuevos genes adquiridos mediante transferencia horizontal en procesos de adaptación microbiana como puede ser la resistencia antibióticos, cambios en la virulencia durante procesos de infección o resistencia a metales pesados (Dobrind *et al*., 2004, Hsiao *et al*., 2005, Roberts y Kreth 2014). En el caso de *S. ruber*, gran parte de los genes de la categoría COG M codificaron para sulfotransferasas y glicosiltransferasas, la mayoría de ellas ubicadas dentro del *cluster* de biosíntesis del antígeno-O, una de las tres partes que conforman la estructura del lipopolisacárido bacteriano, principal componente de la membrana externa de las bacterias gram negativas, junto con el lípido A y el núcleo o *core* (Caroff y Karibian 2003; Wang y Quinn 2010). Se trata de una cadena de polisacárido altamente variable compuesta por entre 1 y 8 unidades de azúcar de gran importancia en las envueltas de las bacterias gram negativas y está involucrado, entre otras cosas, en los procesos de reconocimiento virus-hospedador e inmunoespecificidad bacteriana (Samuel y Reeves 2003; Wang y Quinn 2010). Este *cluster*, encontrado en fGIs de reemplazamiento de microorganismos acuáticos como *A.macleodii* (Gonzaga *et al*., 2012, López-Pérez et al., 2013) y *H. walsbyi (*Martín-Cuadrado *et al*., 2015)*,* contiene genes que suelen agruparse en 3 categorías: implicados en la síntesis de nucleótidos glicosilados, glicosiltransferasas y genes que procesan el antígeno-O (Samuel y Reeves 2003). La elevada proporción de genes relacionados con envolturas celulares en *S. ruber* y especies como *H. walsbyi* (Cuadros- Orellana *et al*., 2007) y su gran diversidad intraespecífica, entre linajes clonales, podrían responder a la presión adaptativa ejercida por los virus del ambiente o a eventos de desecación, ya que las glicosiltransferasas estarían involucradas en ambos procesos (Rodríguez-Valera *et al*., 2009; Peña *et al*., 2010; Mira *et al*., 2010., López -Pérez *et al*., 2014). Precisamente las principales diferencias en los metabolomas de las cepas de la especie se encontraron a nivel de glicosilasas, sulfotransferasas y lípidos de membrana (Roselló-Mora *et al*., 2008; Peña *et al*., 2010., Antón *et al*., 2013). Dada la importante presión y papel de los virus

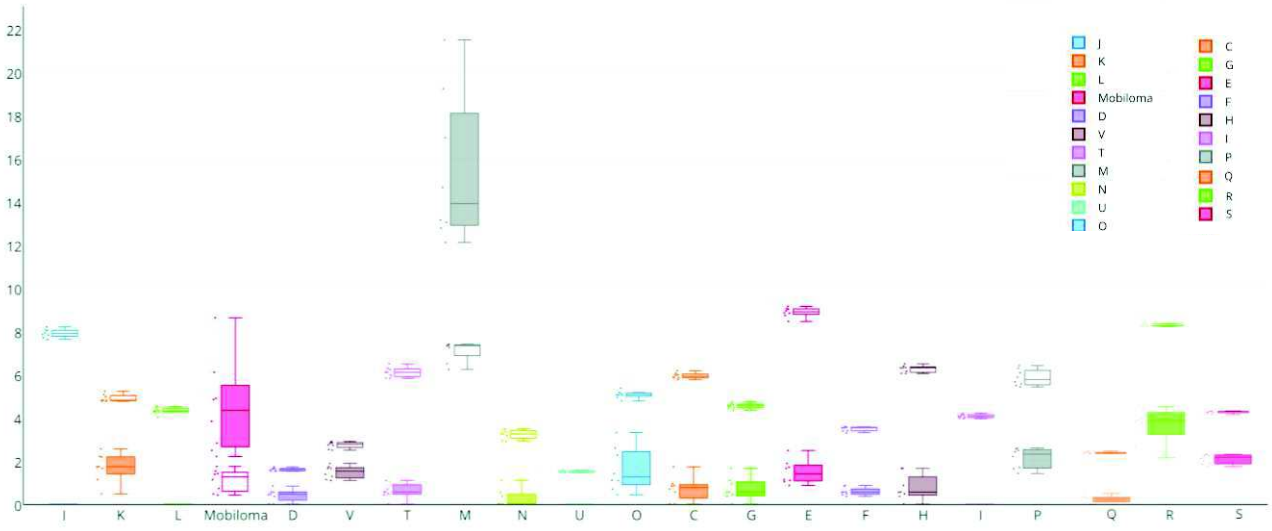


Figura C2.8. Diagrama de caja que muestra la distribución porcentual (eje de ordenadas) para los genes contenidos en el genoma completo (caja transparente) y en las HRVs (caja rellena) de las ocho cepas de *S.ruber* secuenciadas en función de su anotación COG. Se observa un incremento en la proporción de genes del mobiloma y de la categoría COG M en las HRVs.

en ambientes halófilos extremos (Santos *et al*., 2012), estas diferencias podrían afectar al rango de hospedador de los fagos ambientales y a la susceptibilidad por infección, influyendo notablemente en la evolución de las especies microbianas ya sea por impacto de la lisis diferencial o de los procesos de transducción (Rodríguez-Valera *et al*., 2009). Su presencia en fGI puede contribuir a la generación de diferentes dianas de reconocimiento de fagos, diluyendo la presión ejercida por estos y generando diferencias en la susceptibilidad por infección entre cepas cercanas, como las diferencias observadas entre M8 y M31 (Peña *et al*., 2010). Además encontramos genes que codifican para transportadores TonB y proteínas de origen vírico (recombinasas y terminasas de fago en su mayoría) y tirosín recombinasas específicas de sitio como XerD, estas últimas localizadas habitualmente en fGIs (Fernández-Gómez *et al*., 2012; Bellanger *et al*., 2013).

Un análisis estructural comparativo de estas regiones permitió detectar una serie de bloques sinténicos o *clusters* (**figura C2.9**) similares a los descritos en especies como *A. macleodii* (López -Pérez *et al*., 2013) o *S. baltica* (Fernández-Gómez *et al*., 2012), pero también entre fGIs aditivas de diferentes géneros de *Bacteroidetes* marinos (Fernández-Gómez *et al*.,

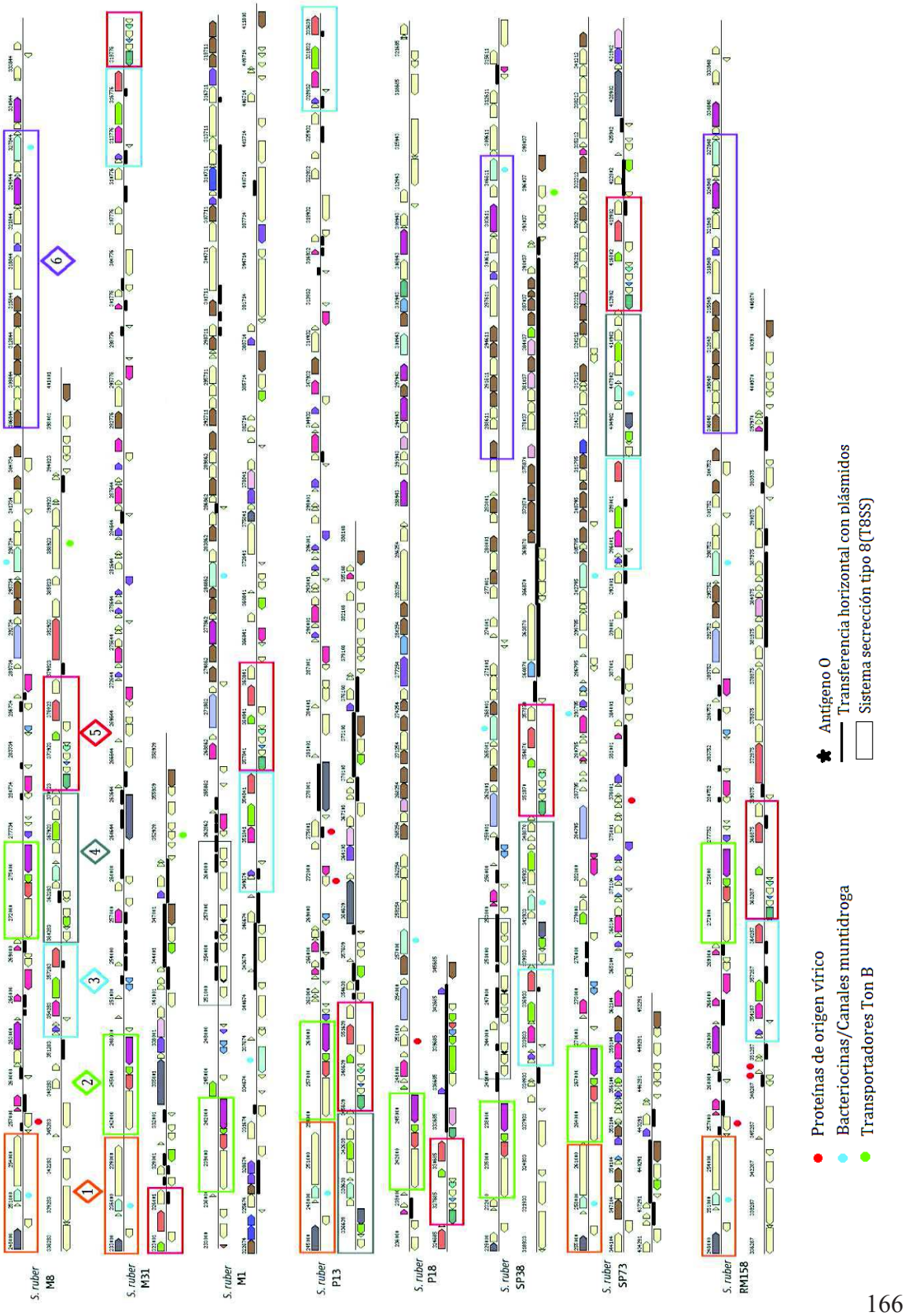

S. ruber M8
S. ruber M31
S. ruber M1
S. ruber P13
S. ruber P18
S. ruber SP38
S. ruber SP73
S. ruber RM158
1
2
3
4
5
6
Proteínas de origen vírico
Bacteriocinas/Canales muntidroga
Transportadores Ton B
Antígeno O
Transferencia horizontal con plásmidos
Sistema secreción tipo 8(T8SS)

Figura C2.9. Representación del contenido génico de las HRV1 (página anterior) y HRV2 (página actual) de las 8 cepas secuenciadas de *S.ruber*. Los genes se muestran coloreados según su categoría funcional COG (misma leyenda de colores de la figura C2.4). Enmarcados aparecen los *clusters* homólogos, numerados del mismo modo que en la figura C2.9). Destacados, se indican los genes de origen vírico, transportadores TonB y genes que codifican para el antígeno O. El sistema TSS8 aparece enmarcado en negro.

2012). Estos bloques sinténicos, más frecuentes en la HRV1, encuadrados en la **figura C2.9**, en ocasiones están presentes en todas las cepas, conteniendo genes del *core* genoma y entre los que se intercalan elementos transponibles y genes de origen viral. Además de estos bloques conservados, en la zona central de la HRV2 encontramos otros funcionalmente similares posicionados en las mismas zonas dentro de esta GI, que responderían con *clusters* de genes típicos detectados en fGIs de reemplazamiento de especies como *A. macleodii*. Estos últimos *clusters* compartirían una proteína común, en el caso de un gen codificante para bacteriocinas y el codificante para la biosíntesis polisacárido de la cadena O (**figura C2.9**), del que se habló anteriormente en este mismo apartado.

Otro aspecto analizado de las fGI fue su dinámica y evolución considerando los *clusters* sinténicos y las regiones con contenido génico variable situadas dentro de ellas. Los mecanismos y estrategias de variabilidad de estas regiones suscitan especial interés debido a su relación funcional con el ambiente y su papel en los procesos de adaptación (Fernández-Gómez *et al*., 2012). Estas regiones variables contienen transposasas específicas de cepa y proteínas de origen vírico, incluyendo integrasas y recombinasas, que estarían sometidas a una movilidad constante, tal como apuntaba el análisis de los eventos de duplicación recientes mostrado en el capítulo 1 (apartado 4). Esta distribución de elementos móviles y *clusters* se ha observado en algunas GI de microorganismos acuáticos como *A. macleodii* (Gonzaga et al., 2012, López-Pérez *et al*., 2013) y *H. walsbyi* (Cuadros-Orellana *et al*., 2007; Martin-Cuadrado *et al*., 2015), y refleja el carácter dinámico de las mismas. De esta manera las HRVs podrían incorporar material de eventos conjugativos-integrativos, como se mencionará más adelante, o DNA libre mediante recombinación ilegítima y transposición. Estos mecanismos de variabilidad serían distintos a los que predominan sobre el *core* genoma y los *clusters* sinténicos. A menudo los genes virales están rodeados de genes de tamaño reducido pobremente anotados, que podrían tener también origen viral o ser los restos de un profago.

Los *clusters* homólogos mencionados anteriormente y situados dentro de las HRVs mantienen un elevado grado de sintenia, lo que nos llevó a plantearnos si, como el resto del genoma *core*, podrían estar sometidos a recombinación homóloga. Para comprobarlo se realizó un análisis de recombinación con los alineamientos de 7 estos *clusters* (4 de la HRV1 y 3 de la HRV2) empleando los programas RDP4 (Martin *et al*., 2010) y Dual Brothers **(**Suchard *et al*.,

2005). Los árboles y los gráficos resultantes mostraron en todos los casos la existencia de puntos de ruptura de recombinación, y una genealogía clonal diferente para cada uno de ellos (**figura C2.10**). Estos datos reflejan el importante efecto de la recombinación homóloga incluso en regiones sinténicas cortas (entre 5 y 12 kb) situadas en fGI, fenómeno similar al observado en la HR1-GI de *S. baltica* (Fernández-Gómez *et al*., 2012) acorde con el impacto de este mecanismo en la evolución general de los genomas de esta especie (Caro-Quintero *et al*., 2011). Estudios

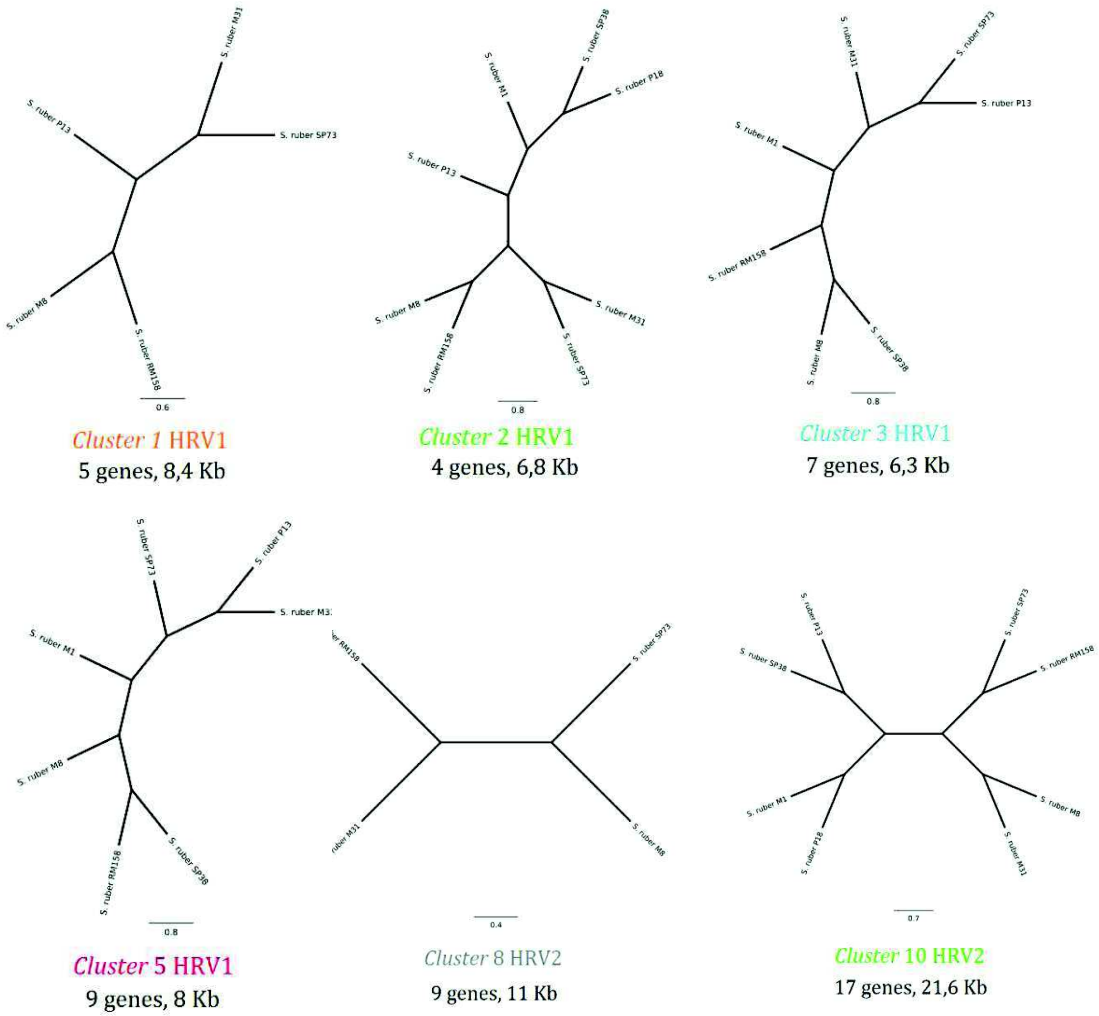


Figura C2.10. Arboles filogenéticos generados durante el análisis de recombinación homóloga con el programa Dual Brothers para seis de los 10 *clusters* identificados y mostrados en la figura 2.8. La incongruencia de los árboles releja el efecto de la recombinación homóloga sobre las líneas clonales de *S. ruber*.

previos realizados en procariotas acuáticos de vida libre indican que la recombinación homóloga es uno de los mecanismo que contribuye a la diversidad en fGIs con genes implicados en envueltas celulares (Fernández-Gómez *et al*., 2012). Al comparar la estructura de las HRVs de las 8 cepas destacan dos fenómenos interesantes. El primero es la similitud entre las HRV1 de las cepas M8 y RM158 ambas aisladas de las salinas de Campos (Mallorca) en 1999 y 2009 respectivamente. Los análisis de recombinación para los 4 *clusters* estudiados a lo largo de la misma los sitúan como las cepas más cercanas, lo cual sugiere que la isla completa, aprovechando la sintenia de las regiones flanqueantes a la misma, pudo recombinar entre ambas cepas o un clon cercano a estas en un evento relativamente reciente. La HRV1 de estas dos cepas es idéntica a excepción de dos inserciones, una de ellas de 7kb en el extremo 3' de la isla de RM158. Intentando averiguar el posible origen de esta región, observamos que presentaba homología con uno de sus plásmidos, RM158-pSR67, lo cual nos lleva a pensar que los plásmidos podrían estar contribuyendo a la diversidad observada en las HRVs. En segundo lugar, las cepas M31 y P13, aisladas en Mallorca y Santa Pola respectivamente en 1999 y 2009, presentaron una composición génica y estructura idénticas salvo en las secuencias situadas en la región 5'de sus HRV1. Tres de los cuatro *clusters* situaron a ambas cepas como las más próximas filogenéticamente, a excepción del situado en la región 5' de la HRV1, dato que apoya que estemos observando el mismo fenómeno de transferencia horizontal que entre M8 y RM158.

Islas genómicas HGT-GI. El resto de GI detectadas fueron de tipo HGT-GI. A excepción de las GI5-P18, GI3-SP73 y GI3-RM158, homólogas y colineares con una longitud de alrededor de 114Kb, el resto de GI fueron específicas de cepa y su tamaño fue menor al observado en las HRVs (entre 22 y 45 kb) (**figura C2.11, tabla C2.4**). Todas estuvieron flanqueadas por tRNAs o regiones repetitivas, lo que podría facilitar su integración y excisión como eventos de transferencia horizontal (revisado en Haecker y Kaper 2000). Las cinco de menor tamaño, GI3-M31 (42Kb), GI3-P13 (27 kb), GI-P18 (33 kb), GI3-SP38 (23 kb) y GI-4 RM158 (27 kb) presentaron un enriquecimiento en HP y una elevada proporción de ORFs de corta longitud, en las tres primeras flanqueando aguas arriba y abajo una recombinasa de fago o una recombinasa XerD. XerD es una recombinasa específica de sitio que participa, mediante recombinación homóloga, en la resolución de dímeros a monómeros intra e intermoleculares durante la segregación (Blakely y Sherratt 1996). Los elementos integrativos e integrativos conjugativos

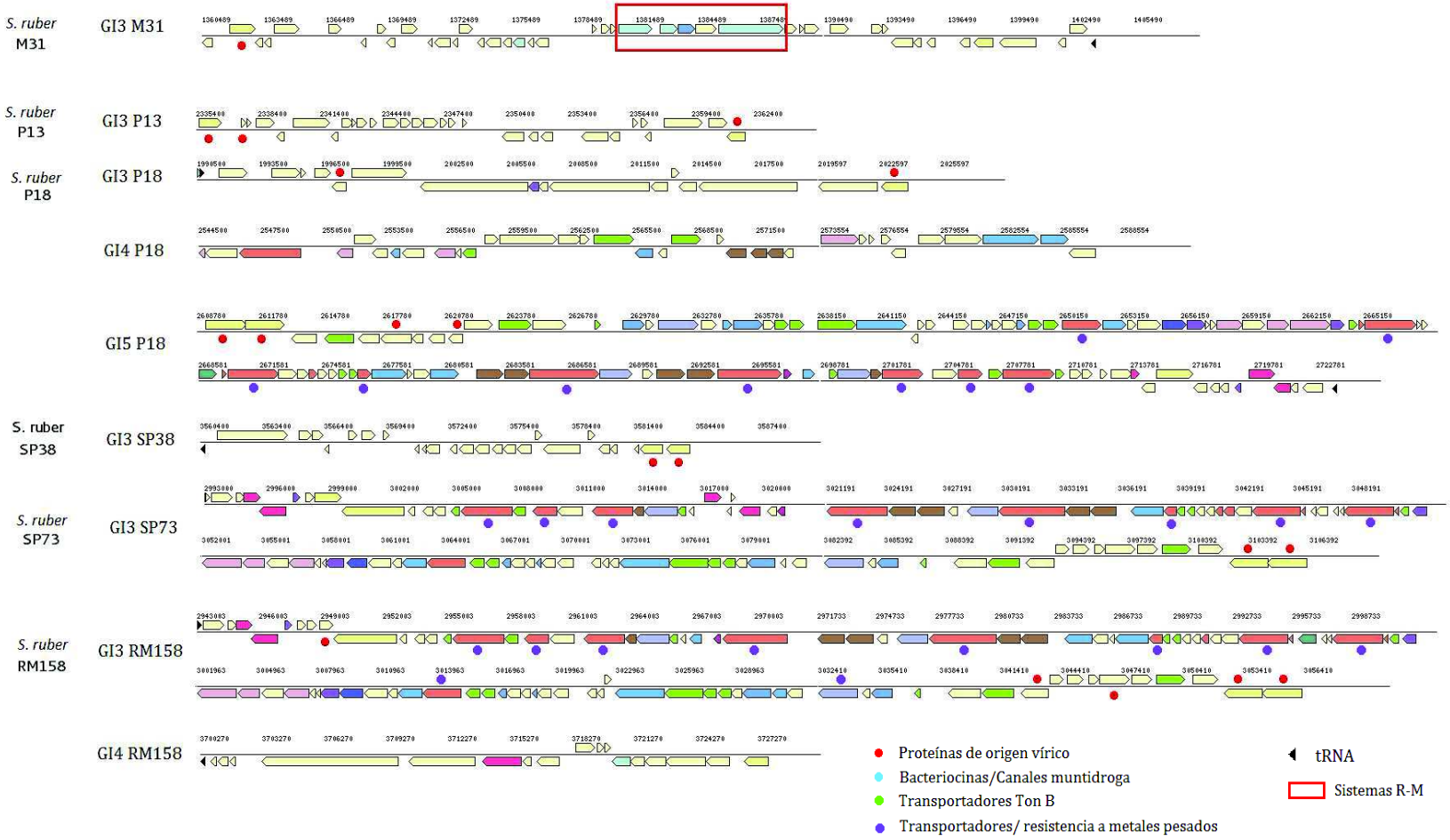


Figura C2.11. Representación del contenido génico de las HGT-GI de las cepas M31, P13, P18, SP73 y RM158 de *S.ruber*. Los genes se muestran coloreados según su categoría funcional COG (misma leyenda de colores de la figura C2.4). Destacados, se indican los genes de origen vírico, transportadores TonB e implicados en la resistencia o transporte de metales pesados. En el caso de GI3-M31 aparece enmarcado su RM-I).

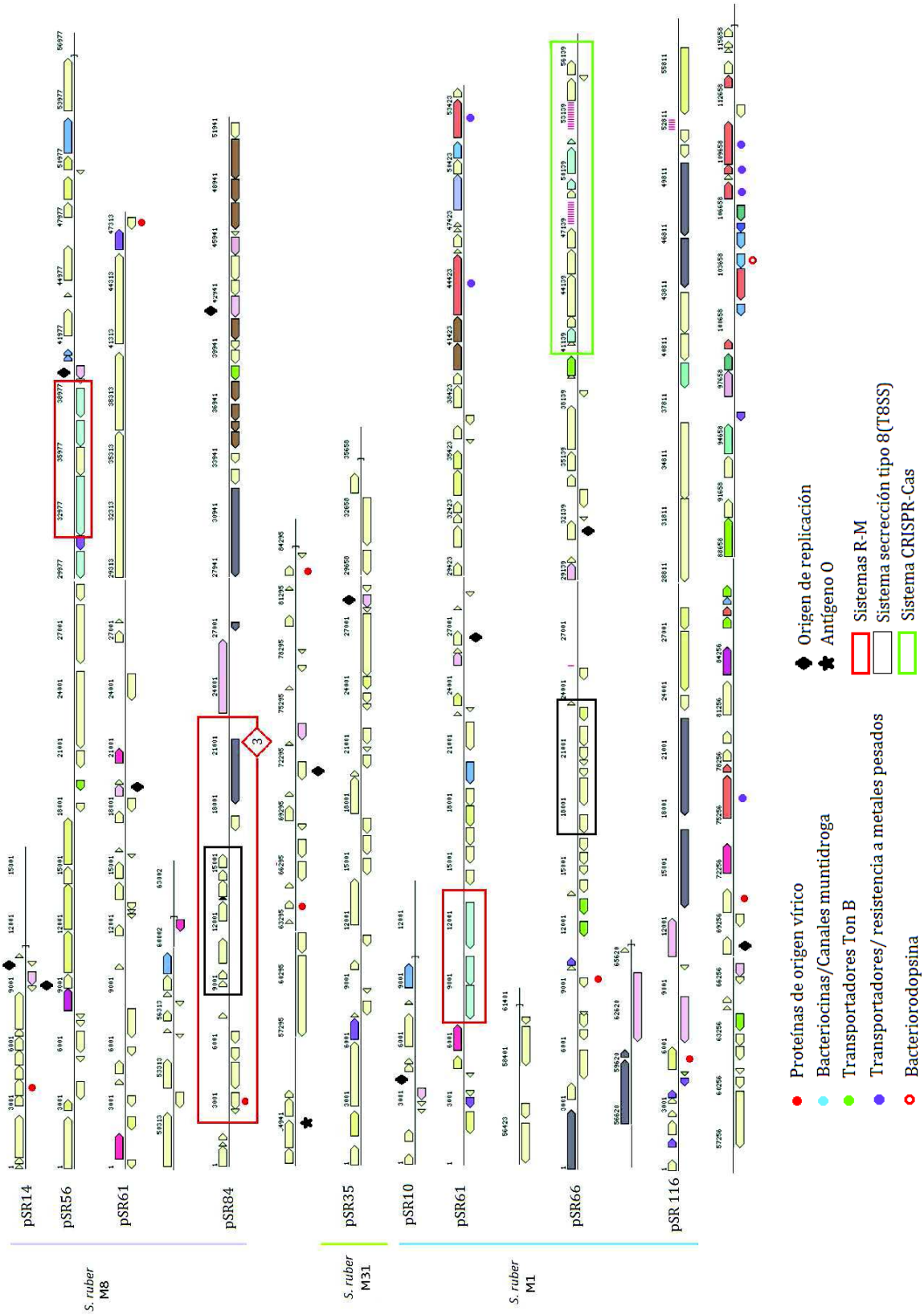
(ICEs) emplean este tipo de recombinasa para su integración o escisión (Das *et al*., 2013; Bellanger *et al*., 2013). Su presencia en GIs y zonas flanqueantes a las mismas, en plásmidos y en regiones recombinadas, apoya un papel relevante en los procesos de recombinación homóloga e intercambios entre plásmidos y GIs como se ha observado en *Neisseria gonorrhoae* (Dominguez *et al*., 2010) o *N. memingitidis* (Woodhams *et al*., 2012) en procesos de escisión y ganancia de GIs respectivamente. Como se acaba de comentar, en el caso de *S. ruber* algunas de las GI están flanqueadas por regiones repetitivas o tRNAs, secuencias que las recombinasas XerD selecionan preferentente para la integración de secuencias (Bellanger *et al*., 2013). Esto podría indicar que en origen se trataran de islas generadas por la integración de un fago lisogénico inactivado o de plásmidos. Funcionalmente destaca la presencia de un sistema de restricción-modificación tipo I (RM-I) en la GI3-M31, homólogo al ubicado en la cepa M8 en el plásmido pSR56. De mayor tamaño que las anteriores, la GI-4P18 (43 kb) contiene gran cantidad de componentes de membrana (COG M), algunos implicados en la resistencia a disolventes orgánicos, y factores de transcripción (COG K) entre ellos sigma 70.

Las tres GI de mayor tamaño, GI5-P18 (114 kb), GI3-SP73 (113 kb) y GI3-RM158 (113 kb) presentaron una composición idéntica, salvo por una inserción de alrededor de 3 kb mediada por transposones en la GI3-SP73. Constituyen 3 regiones sinténicas entre estas tres cepas, situadas en SP73 y RM158 tras una zona variable de menos de 10 kb presente en todos los cromosomas secuenciados. En P18, esta GI se sitúa unas 250 kb antes en sentido 3' e invertida. Esta GI está flanqueada en su extremo 5' por una tRNA-ser y contiene en sus extremos cuatro integrasas de fago rodeadas de nuevo de ORFs cortas. Destaca la gran cantidad de genes codificantes para transportartadores iónicos (COG P) y de carbohidratos (COG G) y elementos estructurales de membrana (COG M). Muchos de estos transportadores codifican para proteínas que participan en procesos implicados en resistencia a metales pesados (Cu, Mg, Au, Co Zn, Cd) y transporte de metales cofactores de proteínas (Zn, Fe, Co), por lo que esta GI podría considerarse como una isla de metalorresistencia, por analogía a la de halofila descrita para M31 previamente (Mongodín *et al*., 2005). La presencia de elementos de resistencia a metales pesados es común en elementos móviles y GI y se asocian a estrategias de adaptación de diversos procariotas (Hsiao *et al*., 2005; Bellanger *et al*., 2013), entre ellos organismos acuáticos como *A. macleodi*i (Ivars-Martinez *et al*., 2008; López-pérez *et al*., 2013) o *H. walsbyi* (Martín-Cuadrado

et al., 2015), patógenos como *Legionella pneumophila* (D'Auria *et al*., 2012) o saprófitos de vida libre como *Pseudomonas putida* (Sharma *et al*., 2014).

4.2- Plásmidos.

El número de los plásmidos detectados fue variable en las diferentes cepas, oscilando entre los cuatro encontrados en RM158, M1 y M8, y la ausencia de los mismos en P13 y P18 (**figura C2.4; figura C2.12**). Además el rango de tamaños varió entre las 10 kb y 116 kb. Se trata de elementos que contribuyen casi en su totalidad al genoma accesorio de la especie, ya que contienen una elevada proporción de genes específicos de cepa. El contenido en GC de los mismos fue menor que el del cromosoma variando entre el 64% y el 55%. Todos contuvieron un origen de replicación típico de plásmidos de bajo número de copias con una proteína anexa implicada en la replicación. En algunos casos el gen implicado codifica para la proteína ParA (*plasmid partition protein*), involucrada en la división del plásmido (Schumacher 2006; Mierzejewska y Jagura-Burdzy 2012; Iestwaart *et al*., 2014) y en otros para la proteína iniciadora de la replicación RepB (*replication initiation protein*), esta última implicada en la replicación y segregación plasmídica en bacterias gram negativas (Cevallos *et al*., 2008; Pinto *et al*., 2012). Como se indicó en el apartado 1, los plásmidos con tamaños menores de 40 kb presentaron una cobertura al mapear los *reads* usados en su ensamblaje incluso 4 veces superior a la de los cromosomas, lo cual indica que, aunque bajo, existe un número de copias adicionales que es variable. Funcionalmente, la mayoría de plásmidos presentaron una proporción de genes de HP muy superior a la de los cromosomas, que ronda el 24%. Sólo los plásmidos M8-pSR56 y M8-pSR84 mostraron un contenido similar al de los cromosomas, mientras que 13 de los 18 plásmidos ensamblados contuvieron más de un 35% de HP, en ocasiones superando el 60%. Aunque no existe una correlación clara, parece que los plásmidos de mayor tamaño tuvieron en general un %GC mayor y un contenido menor de HP. Los plásmidos de menor tamaño (M8-pSR14; M31-pSR35; M1-pSR10; SP73-pSR16; SP38-pSR11; SP38pSR14; RM158-pSR31) fueron los que además de contener una proporción mayor de HP presentaron agrupaciones de ORFs cortas siempre cerca de una recombinasa de virus, una recombinasa específica de sitio, usualmente XerD, o una terminasa (**figura C2.12**). Destaca el caso del plásmido RM158-pSR31


pSR14
pSR56
pSR61
pSR84
pSR35
pSR10
pSR61
pSR66
pSR 116
S. ruber M8
S. ruber M31
S. ruber M1
Proteínas de origen vírico
Bacteriocinas/Canales muntidroga
Transportadores Ton B
Transportadores/ resistencia a metales pesados
Bacteriorodopsina
Origen de replicación
Antígeno O
Sistemas R-M
Sistema secreción tipo 8(T8SS)
Sistema CRISPR-Cas

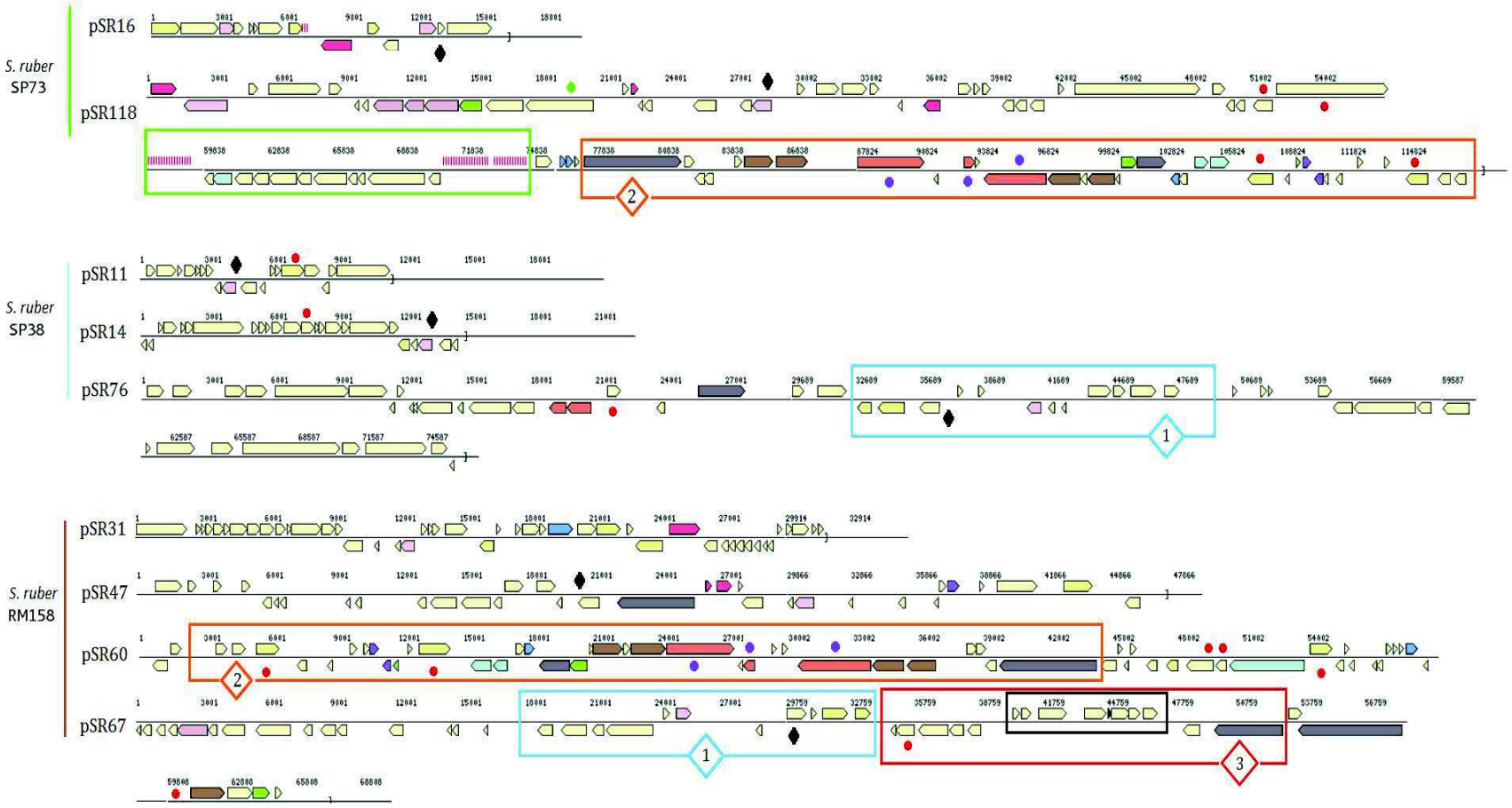


Figura C2.12. Representación del contenido génico de los plásmidos de las cepas M8, M31, M1 (página anterior) y SP38, SP73 y RM158 (página actual) de *S.ruber*. Los genes se muestran coloreados según su categoría funcional COG (misma leyenda de colores de la figura C2.4). Destacados con puntos, se indican los genes de origen vírico, transportadores TonB e implicados en la resistencia o transporte de metales pesados. Encuadrados según la leyenda, de identifican los sistemas CRSPR-Cas, RM-I y T8SS. Numerados y encuadradros, las tres regiones sinténicas compartidas entre tres parejas de plásmidos mencionadas en el texto.

que contiene 45 ORFs en sólo 31 Kb. Los plásmidos de mayor tamaño se caracterizaron por contener elementos pertenecientes a sistemas barrera para la captación y estabilización de DNA del medio, sistemas de secreción y elementos de membrana. En los plásmidos M8-pSR56 y M1-pSR61 se ubican dos sistemas de modificación-restricción de tipo 1, de los que se hablará a continuación. Como mecanismo barrera, destaca la presencia de sistemas CRISPR-Cas en los plásmidos M1-pSR66 y SP73-pSR118, que se describirán más adelante. Los plásmidos M8-pSR61, M1-pSR66, M8-pSR84 y RM158-pSR67 albergan agrupaciones de genes que codificaron para sistemas de secreción. El primero de ellos contiene un sistema de secreción pobremente caracterizado en el que están involucradas las 3 ORFs de mayor longitud del mismo. El resto de plásmidos contuvieron un sistema de secreción de tipo VIII (T8SS, *curlin like*) implicado en la síntesis de fimbrias en bacterias gram negativas. Este tipo de sistemas de secreción tipo se encontró en las HRV1 de las cepas M1 y SP38 (**figura C2.9)**, en regiones situadas entre los *clusters* de recombinación homóloga que presentaron homología con secuencias plasmídicas. La presencia de sistemas de secreción en GI y plásmidos es común en bacterias con estrategias ecológicas diversas. Se ha detectado la presencia de sistemas de secreción asociados a virulencia de tipo I, III, IV y VI en *Bacteria* en GI y plásmidos de bacterias marinas de vida libre (Persson *et al*., 2009; Fernández-Gómez *et al*., 2012) y de tipo II, III y IV en plásmidos y GIs de patógenas (revisado en Kado 2009). El sistema tipo II, con similitudes estructurales con el tipo IV, está implicado en la síntesis de pili. Los dos últimos sistemas, II y IV, están involucrados en la secreción de factores de virulencia vitales durante la patogénesis, procesos de adaptación a la interacción virus-hospedador y formación de biofilms y, en el caso del sistema de tipo IV, en la transferencia de DNA mediante conjugación, locomoción y adherencia (revisado en Melville y Craig 2013; Imam *et al*., 2011). En concreto los dos últimos plásmidos mencionados, M8-pSR84 y RM158-pSR67, compartieron una región sinténica de 17 ORFs, alrededor de 10,1kb con una identidad de secuencia del 99,6% (bloque 3, **figura C2.12**) que incluyó este sistema de secreción de tipo 8. Un análisis con BlastN comparando la secuencia de todos los plásmidos entre si reveló otras dos regiones homólogas extensas, a las que denominamos bloques 1 (15,7 kb entre SP38-pSR76 y RM158-pSR67) y 2, de 41,24 kb (entre SP73-pSR118 y RM158-pSR60) (**figura C2.12**).

Finalmente, con una elevada proporción de genes relacionados con elementos estructurales de membrana (COG M) y transportadores iónicos (COG P), encontramos los plásmidos M8-pSR84, RM158-pSR60 y M1-pSR116. El primero contiene un *cluster* de genes (COG M situado tras el gen codificante para el polisacárido de la cadena O (antígeno O en M8) como el descrito en la HRV2 de la mayoría de cepas. Entre los genes de la categoría COG P encontramos algunos relacionados con la resistencia a metales pesados y el transporte de hierro. La composición y estructura génica de este *cluster* no coincide con la de ninguno de las HRV2, sin embargo el hecho de encontrar este *cluster* no sinténico en un plásmido sugiere que éstos podrían actuar como vehículo transmisor incorporando secuencias a las GI. La comparación de este mismo plásmido, M8-pSR84, con las demás HRV1 permitió identificar en la HRV1 de SP38 una región de 30 kb (posiciones 360.498-390.549) con una elevada identidad y sintenia (93%, e-value=0). Estos resultados sugieren que los plásmidos actuarían como un vehículo muy dinámico de acceso al *pool* ambiental mientras que las GI serían la puerta de entrada al cromosoma y la zona de intercambio de secuencias como la del *cluster* descrito en pSR84. Este fenómeno se ha descrito entre plásmidos y GIs de especies como *S. baltica* (Fernández-Gómez *et al*., 2012). La integración de plásmidos en cromosomas es un fenómeno conocido en bacterias, donde se ha descrito en especies como *E.coli* o *B. subtilis* (Casei *et al*., 1991). La expresión de este *cluster* en las cepas M8 y M31 se da a niveles muy elevados como se vio en el capítulo 1, por lo que el entorno genómico de la HRV podría ser más inductor de la expresión que el de un plásmido, en donde registramos niveles de expresión bajos para los genes de estas cepas mediante RNAseq. Esto podría deberse a que las GI además de incorporar nuevos genes juegan un papel determinante en la modulación de su transcripción, regulación y transducción (Coleman *et al., 2006*).

En base a estas últimas evidencias, y con el objetivo de profundizar en la dinámica y evolución de estos elementos móviles, se realizó una comparación con BlastN de cada plásmido contra los cromosomas del resto de las cepas para identificar regiones plasmídicas que pudieran haberse integrado. La mayoría de identidades se dieron a nivel de las HRVs lo que indica que existe una mayor dinámica e integración preferencial de genes entre los plásmidos y GIs dentro los cromosomas. Aunque las GIs abarcan entre un 6,1% a un 10,6% de la secuencia de los genomas (**tabla C2.4**), la proporción de secuencias plasmídicas reclutadas contra estas regiones

fue muy superior a este porcentaje en los 8 genomas estudiados (entre un 14,81%, en el caso de P18 para y un 48,14% para M31) como se muestra en la **figura C2.13**. La mayoría de *hits* se concentraron en las regiones específicas de la HRV1 y en muchos casos se dieron con plásmidos de otras cepas. En esta misma figura muestra con una barra acumulada el porcentaje de secuencia compartido de cada plásmido con los cromosomas de las 8 cepas. Este último dato podría indicar que existe un proceso frecuente de pérdida y adquisición de plásmidos y explicaría que cromosomas de cepas distintas compartan en su genoma accesorio secuencias presentes en plásmidos de otras cepas. Análisis comparativos similares en *A. macleodii* muestran intercambios similares de más de 20 kb entre plásmidos presentes en diferentes cepas de localizaciones

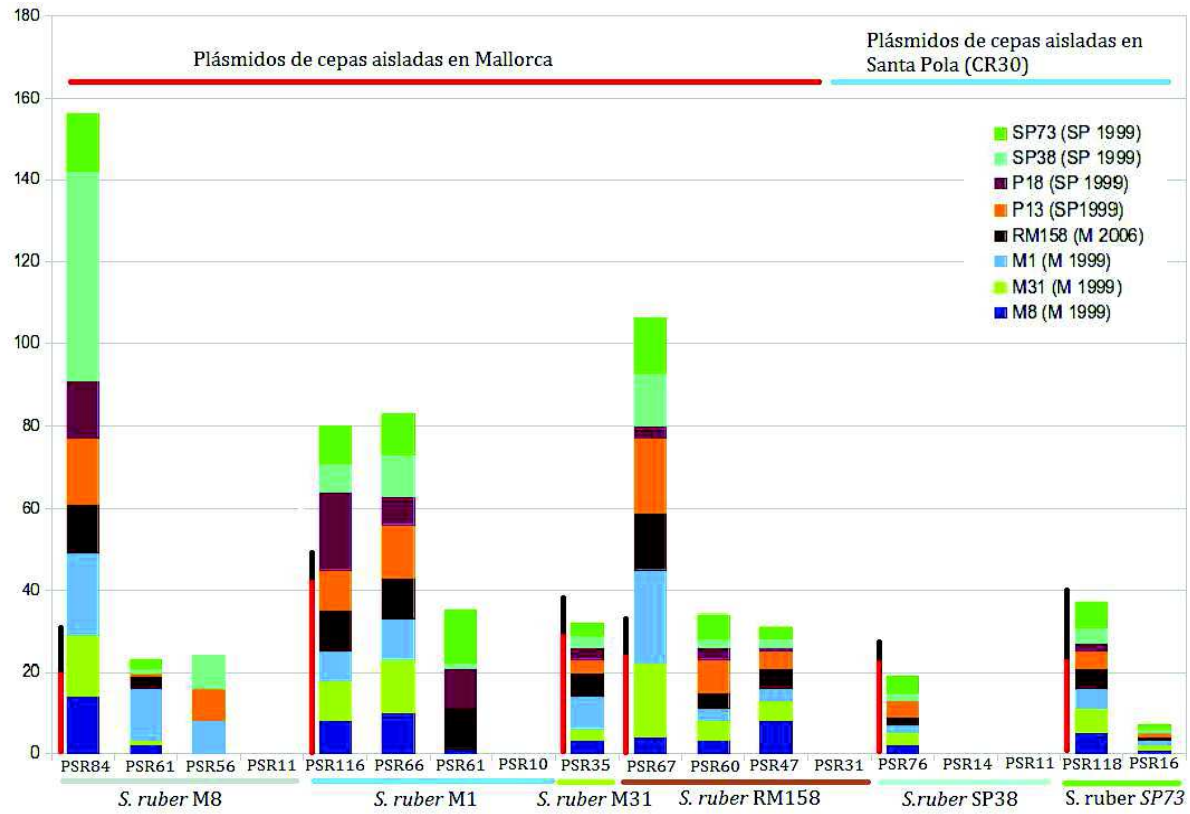


Figura C2.13. Porcentaje acumulado de secuencia que comparte cada uno de los plásmidos de las 6 cepas que los albergan a en cada cromosoma de las 8 cepas secuenciadas de *S.ruber* mostradas en la leyenda. Para cada cepa se muestra una barra adicional con el porcentaje de secuencias del total recibido de plásmidos en la HRV1 (rojo) y en el resto de GI (negro). En los casos de las cepas P13 y P18 los valores (no mostrados) fueron 37,11% y 4,6% para la HRV1 y 42,26% y 14,81 respectivamente en todas las GI.

distantes geográficamente y en ocasiones entre plásmidos y GI detectados en otras líneas clonales (CF) (López-Pérez *et al*., 2013), tal como sucede en *S. ruber*. Pese a ello no encontramos algunos de los elementos necesarios en los plásmidos de *S. ruber* para poder considerarlos ICE y que están presentes en plásmidos móviles (Smillie *et al*., 2010., Bellanger *et al*., 2013). Aunque contienen genes codificantes para su integración y escisión, tirosin recombinasas específicas de sitio (XerD) e integrasas, carecen de los elementos básicos para su movilización autónoma, entre los que se incluyen una relajasa (MOB) y los genes codificantes para los complejos MPF y CP que constituyen el sistema de secreción de tipo IV (T4SS) tal como sucede en la mitad de los plásmidos de especies del dominio bacteria (Smillie *et al*., 2010; Guglielmini *et al*., 2011). Otra alternativa es de que pudieran ser elementos integrativos mobilizables (IME) (Burrus *et al*., 2002; Bellanger *et al*., 2013), pues además de presentar los elementos más comunes en la integración y escisión de IME podrían emplear desde otro replicón, en *trans,* los elementos MPF. Sin embargo, aunque los cromososomas de todas las cepas contienen los genes necesarios para constituir los sistemas de secreción tipo II y tipo III, que comparten elementos comunes con el de tipo IV y el aparato de transformación (Imam *et al*., 2011, Korotkov *et al*., 2012., Melville y Craig 2013), no hay indicios de la presencia de este último. La presencia de los genes *com*EC y *com*F en todas las cepas de *S. ruber*, elementos esenciales en el sistema de transformación bacteriano, sugiere este mecanismo junto con la tranducción como alternativas a la transferencia de plásmidos tal como se ha observado entre géneros de Borrelia (Qiu *et al*., 2004).

Los plásmidos de cepas aisladas de las salinas de Mallorca compartieron muchas más secuencias con los cromosomas que los de las aisladas de Santa Pola, estableciendo una diferenciación biogeográfica respecto al acceso al *pool* ambiental. Los reclutamientos de las secuencias de los metagenomas de las salinas de Campos Mallorca (datos no publicados) y Santa Pola, cristalizador CR30 (Ghai *et al*, 2012) contra los plásmidos de las cepas de *S. ruber* (**figura C2.14**) sustentan esta observación pues se detectaron elevados niveles de reclutamiento a identidades superiores al 99% en los plásmidos que compartieron más fragmentos con los cromosomas secuenciados: RM158-pSR60, M1-pSR66, RM158-pSR47, SP73-pSR118, M1-pSR61, M8-pSR84, M1-pSR116 y SP38-pSR76. Estos niveles difícilmente son atribuibles exclusivamente a secuencias de origen plasmídico más si cabe cuando se aproximan a los valores

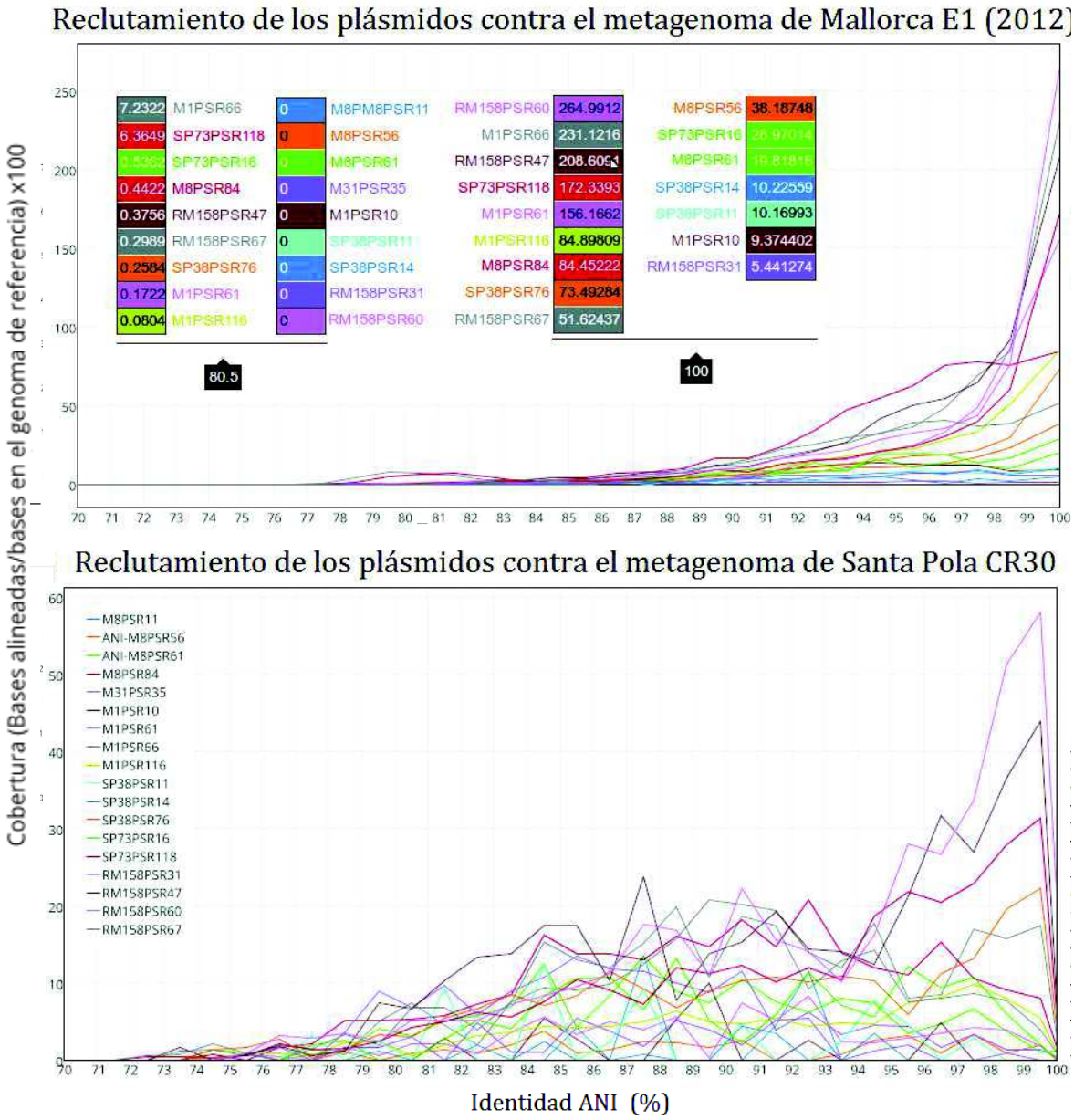


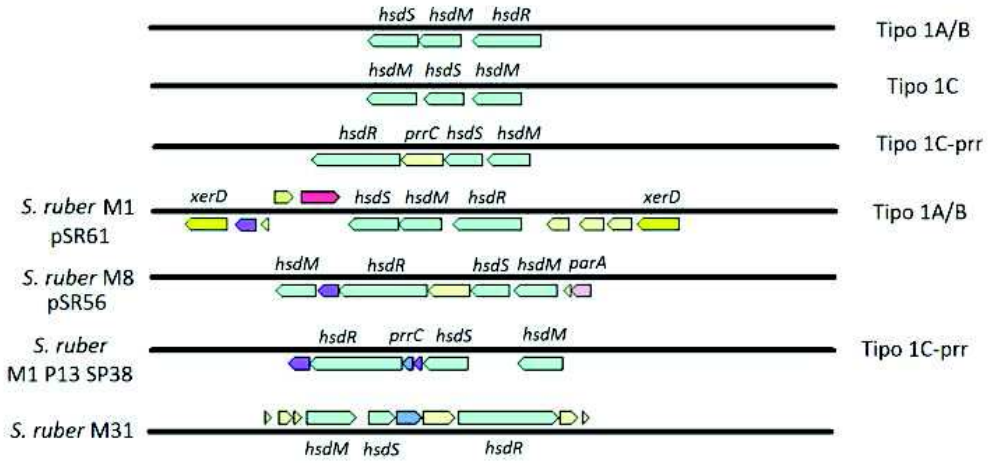
Figura C2.14. Reclutamiento de los metagenomas del cristalizador CR30 de Santa Pola (Ghai *et al*., 2012) y las Salinas de Campos de Mallorca contra los plásmidos secuenciados de las cepas de *S. ruber*. Los reclutamientos muestran cómo se reclutan *hits* con identidades menores 94%. En detalle se muestran las coberturas a identidades del 100% y del 80,5%.

de abundancia detectados para los cromosomas de las 8 cepas en estos mismos ambientes (**figura C2.21**) Para el resto de plásmidos, hay un mayor reclutamiento con secuencias con valores de identidad entre el 86-94%. Estos resultados podrían indicar que: (1) originalmente proceden de cepas de especies distintas divergiendo sus secuencias una vez se estabilizaron en las cepas de *S. ruber*, (2) que las propias cepas de *S. ruber* los contengan y diverja su secuencia desde su incorporporación o (3) que estas secuencias se encuentren de manera ubicua en diversas especies que han intercambiado en algún momento secuencias desde otros replicones, tal como se ha observado con los plásmidos mencionados anteriormente en *S.ruber* y en cepas de *Alteromonas*(López-Pérez *et al*., 2013). En este último caso la divergencia de secuencia por procesos de mutación acumulados entre el donador y el plásmido explicaría que se den mayores reclutamientos a identidades inferiores. El conjunto de datos refleja un intercambio frecuente de plásmidos o de fragmentos de los mismos entre cepas o CF de *S. ruber*, el cual podría darse incluso entre aislados de diferente localización geográfica como sucede en *Alteromonas* (López-Pérez *et al*., 2013) y por lo tanto con una distribución e impacto global.

Elementos barrera codificados en plásmidos: Sistemas restricción-modificación y CRISPR-Cas.

Entre las cepas de *S.ruber* encontramos dos importantes elementos barrera que dificultan la incorporación de secuencias mediante transferencia horizontal: sistemas MR y sistemas CRISPR-Cas. Ambos se ubicaron en regiones del genoma accesorio (GI, *indels* y plásmidos), reflejo de la compleja interacción entre elementos barrera y los procesos de intercambio génico tal como se comentó en la introducción (apartado 1.2.4). Existen antedecentes de la presencia de ambos sistemas en plásmidos y GI de diversas especies procariotas (Corvaglia *et al*., 2010; Fernández-Gómez *et al*., 2012, revisado en Bellanger *et al*., 2013). Se había detectado la presencia sistemas MR en las dos cepas de *S.ruber* secuenciadas (Mogondin *et al*., 2005; Peña *et al*., 2010), ubicado en la M31-GI3 y en M8 en el plásmido M8-pSR56, pero no la de elementos CRISPR-Cas. Basándonos en la dinámica e intercambio detectados entre plásmidos y GI, y en la presencia de una proteína de antirrestricción, común en algunos ICE que contienen sistemas MR como en *S. agalactiae* (Brochet *et al*., 2008) o *Mycoplasma agalactiae* (Marenda *et al*., 2006), es

probable que el sistema de M31 se incorporara al cromosoma desde un plásmido. Los sistemas MR encontrados en M8 y M31 fueron de tipo I, que están constituido por una proteína compuesta de tres subunidades, codificadas por los genes *hsdS, hsdM y hsdR* (**figura C2.15**). Se trata de una enzima multifuncional que cataliza la restricción- modificación y además tiene funciones topoisomerasa y ATPasa (Bickle y Krüger, 1993). Además de estas dos cepas, se detectó la presencia de este sistema en los cromosomas de M1, P13, y SP38 en la misma posición relativa, flanqueados por dos XerD e interrumpiendo la sintenia con el resto de cepas (**figura C2.16**). Estas evidencia sugiere que el sistema pudo incorporarse en alguna de ellas cepas mediante transferencia horizontal y en el resto posteriormente recombinación homóloga de



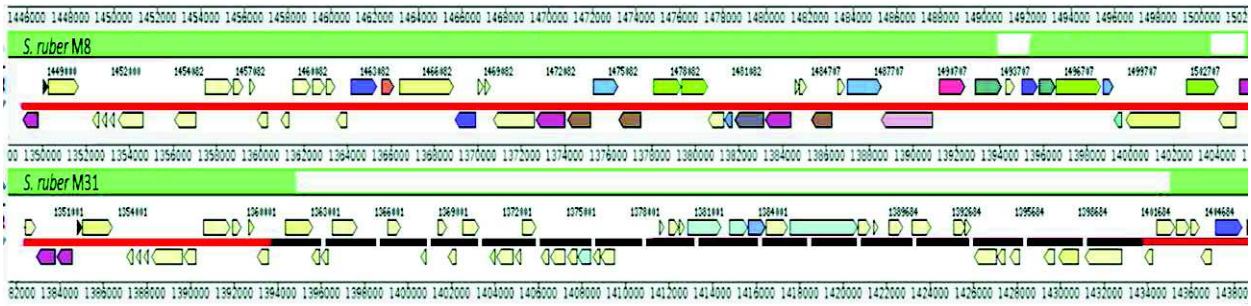


Figura C2.15. Figura superior: Composición y tipo de sistemas MR detectados en las 8 cepas de *S. ruber*. La figura inferior: Disposición de la M31-GI3 y el sistema MR albergado en la misma. Se muestra la inserción generada por la GI mediante el alineamiento local de las cepas M8 y M31, señalando en rojo la parte sinténica y en negro discontinuo la GI.

extremos en la misma región genómica. El sistema MR detectado en las tres cepas (M1, P13, y SP38) fue de tipo IC-prr y el orden de los genes idéntico (*hsdR,* HP, nucleotidiltransferasa *(prrC) hsdS y hsdM*) (**figura C2.15**). En el caso de M1 además se identificó este sistema en el plásmido M1-pSR61, presentando esta cepa una copia extra y distinta del sistema con una distribución tipo

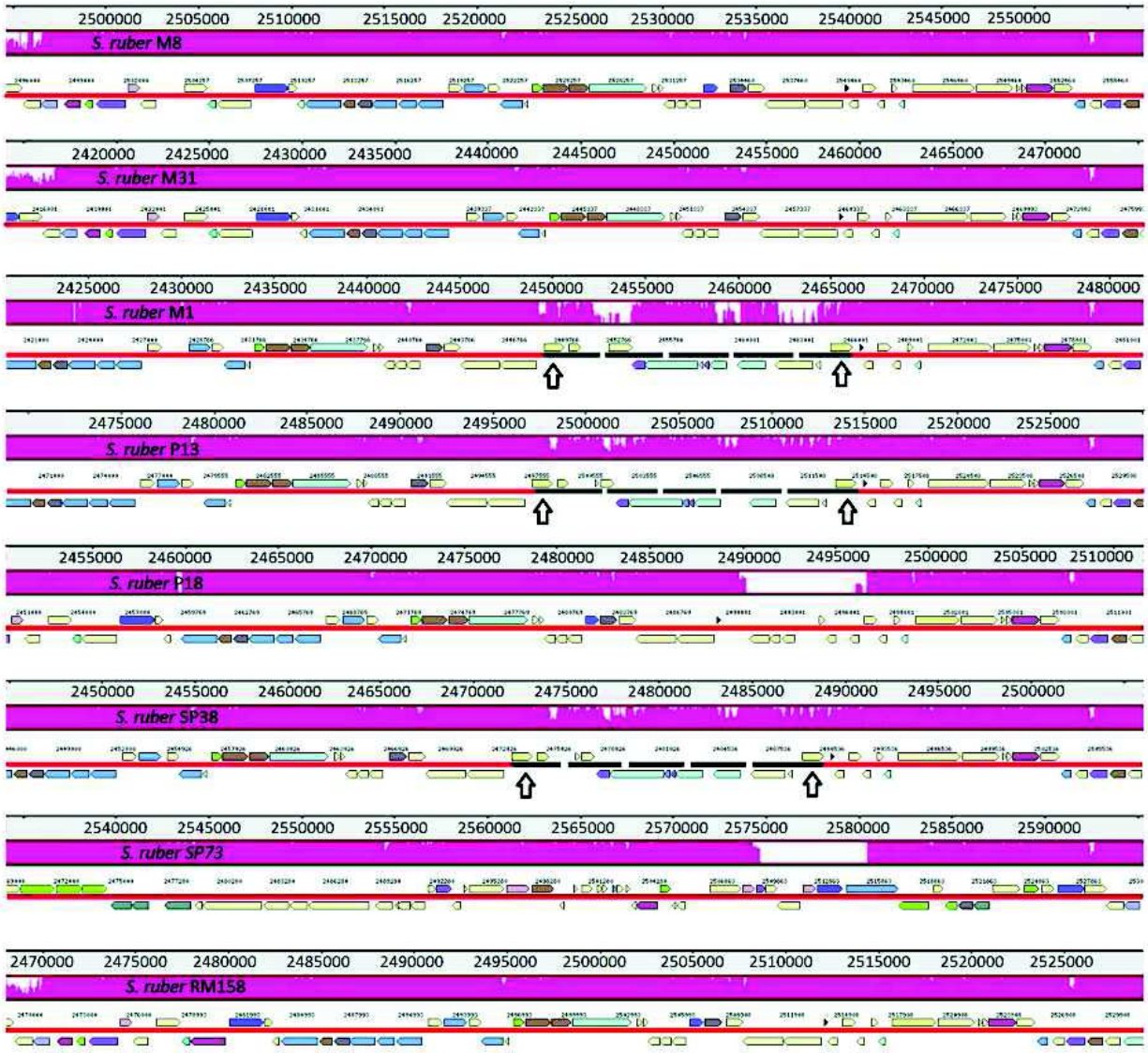


Figura C2.16. Alineamiento local de las 8 cepas secuenciadas que incluye el *indel* donde se localiza el sistema MR de las cepas M1, SP38 y P18. La línea roja señala la región sinténica compartida por las 8 cepas y la negra discontinua el *indel*. Las flechas señalan la posición de las tirosín recombinasas XerD.

1A/B. La presencia de diferentes sistemas MR en plásmidos y cromosomas de una misma cepa se ha descrito en especies con varias cepas secuenciadas, entre ellas *H. pylori* (Corvaglia *et al*., 2010) o *E. coli*, y condiciona el *pool* ambiental real disponible para cada cepa pues limita las secuencias heterólogas favoreciendo la incorporación de otras homólogas por recombinación. Esto, junto al efecto sobre los procesos de transfección, llevaría a un acceso e incorporación diferencial de secuencias al genoma accesorio en diferentes cepas y en ocasiones, favorecería fenómenos de especiación incipiente (Kommireddy y Valakunja 2013), como sucede entre cepas coaisladas de *S. islandicus* (Cadillo-Quiroz *et al*., 2012) o *H. pylori* (Corvaglia *et al*., 2010).

Por otra parte se detectó la presencia de sistemas CRISPR-Cas completos conteniendo agrupaciones CRISPR con 38 y 105 espaciadores en los plásmidos M1-pSR66 y SP73-pSR118 respectivamente y genes *cas*, así como agrupaciones CRISPR adicionales de mucho menor tamaño en los plásmidos M1-pSR116 (5 espaciadores) y SP73-pSR16 (3 espaciadores). Analizando la composición de genes *cas* se caracterizó el tipo de sistema, que resultó ser distinto en cada cepa. En el caso de SP73 fue el subtipo I-E de *E. coli*, identificado tras encontrar los genes *cas2, cas1, (cas6/cse3/casE), cas5e, case4, (casB/Case2), (casA/cse1), y cas3* según la clasificación actual (Haft *et al*., 2005; Makarova *et al*., 2011) flanquados por dos agrupaciones de 33 y 62 espaciadores (**figura C2.17, anexo tabla S2.7**). La primera de estas agrupaciones albergó la repetición degenerada en su extemo 3' y el segundo en 5', lo que permitió orientarlas. Además 30kb aguas arriba del gen *cas1* se encuentra un gen codificante para proteína RAMP (del inglés, CRISPR *associated/Repeat associated mysterious protein*). El *best hit* de los genes *cas* fue con los ortólogos de *Rhodothermus marinus* R-10 (DSM4254), bacteria gram negativa marina aislada por primera vez de una fuente termal marina costera en Islandia (Nolan *et al*., 2009) y la especie más cercana filogenéticamante a *S.ruber* en el omento de su aislamiento. Esta bacteria del filo Bacterioidetes presenta sistemas CRISPR-Cas en sus dos replicones, cromosoma (NC_013501) y plásmido (NC_013502) con 3 agrupaciones (154 espaciadores) y 6 agrupaciones (83) respectivamente. El cromosoma de esta especie contuvo la agrupación de genes *cas* en orden idéntico a la encontrada en SP73-pSR118. En el caso del plásmido M1-pSR66, el subtipo identificado fue el I-B de halófilos, debido a la presencia de c*as5e, cas3 y cas7-1C*. En este caso el *cluster* de genes *cas* estuvo flanqueado en 3' por un agrupamiento CRISPR e interrumpido por otro. Se ha detectado la presencia de sistemas CRISPR-Cas en el 40% y el 90% de los genomas

de bacteria y archaea secuenciados (Grissa *et al*., 2007). Su presencia en elementos del genoma accesorio como plásmidos podría indicar que juegan un papel adaptativo importante en el ambiente hipersalino, en que ya que entre las funciones primarias de este sistema está la de conferir inmunidad frente a infección por virus (Mojica *et al*., 2000; Barrangou *et al*., 2007;

Figura C2.17. Descripción de los sistemas CRISPR-Cas de las cepas M1 y SP73 de *S.ruber*. Figura a: Alineamiento de los sistemas CRISPR-Cas de *S.ruber* SP73 y *Rhodothermus marinus* DSM 4252. Figura b: Distribución de los array CRISPR en el cromosoma y plásmido de *Rhodothermus marinus* DSM 4252. Figura c: Composición génica y secuencia protoespaciadora de los sistemas CRISPR-Cas de las cepas M1 y SP73 de *S.ruber*. En rojo se muestran los *arrays* CRISPR.

Sorek *et al*., 2008) y en este ambiente se han observado las mayores densidades de poblaciones víricas detectadas en ambientes acuáticos (Guixa-Boixareu *et al.,* 1996; Santos *et al*., 2010, 2012). La presencia de espaciadores en diferentes plásmidos de las cepas de *S. ruber*, SP73 y M1, sugiere que estos elementos podrían actuar en *trans*, empleando los genes cas codificados en otro plásmido. La funcionalidad de estas pequeñas agrupaciones favorecería la selección y permanencia de ambos plásmidos en la célula, por lo que los procesos adaptativos en los que estuviese involucrado este sistema podrían darse con la incorporación de agrupaciones si la cepa receptora contiene los genes *cas*.

Se evaluó la funcionalidad y ventaja adaptativa que ofrecen estos sistemas en *S.ruber* buscando las secuencias protoespaciadoras que pudieron originar los espaciadores presentes en las agrupaciones de *S. ruber*. Para ello realizamos una búsqueda exhaustiva en las bases de datos especializadas como CRISPRdb (Grissa *et al*., 2007) y mediante BlastN con los metagenomas de las salinas de San Diego (Rodríguez-Brito *et al*., 2010),SantaPola CR30 (Ghai *et al*., 2012), Mallorca (datos no publicados) y metaviromas de ambientes hipersalinos como las salinas de San Diego (Rodríguez-Brito *et al*., 2010), Santa Pola CR30 (Santos *et al*.,2010), Lago Tyrrell (Emerson *et al*., 2012) y Lago Tuz (Boujelben *et al*., 2012) además de metaviromas dirigidos y virus secuenciados de de *S. ruber* (Villamor *et al*.,, datos no publicados) y *H. walsbyi* (Martín-Cuadrado *et al*., 2013) (**tabla C2.5).** Para SP73 un 24% (26/108) de los espaciadores presentes en SP73 tuvieron buenos *hits* con los metagenomas analizados, la mayoría de ellos tuvieron menos de 5 *hits* y con el metagenoma de Mallorca y en segundo lugar el Lago Tyrrell, y un 15.7% (17/108) con los metaviromas, la mayoría estanques de alta salinidad de San Diego. En M1, un 51% (22/43) de los espaciadores tuvieron *hits* con alguno de los metagenomas analizados, la mayoría con Mallorca y en número menor de 4 (**figura C2.18 a**). Los metaviromas en los que se encontró un mayor número de protoespaciadores potenciales fueron, dentro de San Diego sólo los de los estanques de mayor salinidad y, entre los metaviromas dirigidos, el de *S. ruber*.

Para M1 un 18,6% (8/43) de los espaciadores dieron *hit* con un protoespaciador viral. En esta cepa, algunos de los espaciadores situados cerca de la repetición degenerada proporcionaron un mayor número de *hits* (**figura C2.18 a**), en contra de las observaciones en estudios previos donde los espaciadores adquiridos más recientemente son los que reclutan más secuencias

Tabla C2.5 Características de los metaviromas y metagenomas empleados durante el estudio microevolutivo de las 8 cepas de *S.ruber*, incluyendo localización, año de muestreo y características del ambiente estudiado. En cada caso se detalla el tipo de análisis en el que se empleó a lo largo de este capítulo.

Metagenoma	Características principales	Análisis realizado	Referencia
San Diego (EEUU)	Salinas de San Diego, estanques de salinidad baja, alta y media	Reclutamiento contra cromosoma y plásmidos Búsqueda de protoespaciadores	Rodríguez–Brito *et al*., 2010
Santa Pola (España)	Metagenoma del cristalizador CR30.	Reclutamiento contra cromosoma y plásmidos Búsqueda de protoespaciadores.	Ghai *et al.,* 2012
Mallorca (España)	Estanque de alta salinidad, cristalizador.	Búsqueda de protoespaciadores	2012 (Roselló-Mora *et al.,* datos no publicados)

Metaviroma	Características principales	Análsisis realizado	Referencia
Santa Pola (España)	Metaviroma del cristalizador CR30.	Reclutamiento contra virus aislados de *S. ruber* Búsqueda de protoespaciadores	Santos *et al*., 2010
Lago Tuz (Turquía)	Muestreos estacionales de un ambiente hipersalno	Reclutamiento contra virus aislados de *S. ruber* Búsqueda de protoespaciadores	Boujelben *et al*., 2012
Lago Tyrrell (Australia)	Muestreos estacionales de dos localizaciones en dos años distintos	Reclutamiento contra virus aislados de *S. ruber* Búsqueda de protoespaciadores	Emerson *et al*., 2012
Metavitomas dirigidos	Metaviromas dirigidos para *S. ruber* y *H. walsbyi*	Búsqueda de protoespaciadores	Villamor *et al*., datos no publicados

debido a los procesos coevolutivos y la adquisición de resistencias por parte de los virus (Anderson y Banfield 2008). Ninguna de las dos cepas compartió espaciadores, lo que indica la gran microdiversidad existente dentro de una misma especie. Se han observado ejemplos de CRISPR divergentes entre cepas coaisladas de *S. islandicus* (Held *et al.*, 2010).

La comparación de los espaciadores por BlastN con las bases de datos sólo permitió averiguar la identidad de unos pocos protoespaciadores, la mayoría fueron genomas procariotas (**figura C2.18a; anexo tabla S2.7**), tal como se ha observado anteriormente (Emerson *et al*., 2013). Esto podría reflejar tanto la mutación o selección en genomas víricos para evadir la función de los protoespaciadores (Anderson y Banfield 2008), como la enorme diversidad de virus en este ambiente y su infrarrepresentación en las bases de datos, ya que sí se detectaron *hits* con metaviromas. Algunos de los protoespaciadores identificados para la CRISPR de SP73 corresponden con bacterias acuáticas o marinas, acorde con la homología de sus sistema con *Rhodothermus marinus* y su incorporación mediante HGT. Dos espaciadores de la CRISPR de M1 y otro de la de SP73 obtuvieron como *hit* el *contig* 170 del metaviroma del cristalizador CR30 de Santa Pola (Santos *et al*., 2010). La adquisición independiente de espaciadores para un mismo elemento móvil sugiere que las poblaciones víricas de los ambientes de Mallorca y Santa Pola comparten un *pool* de virus. Además, la adquisición sucesiva y rápida de resistencias de manera independiente en distintas líneas clonales con diferente fondo genómico prevendría el barrido clonal en episodios periódicos por las resistentes y evitaría la pérdida de diversidad (Held *et al*., 2010., Lythgoe y Chao 2003).

Destaca el elevado número de *hits* (75) encontrado en los metaviromas del lago Tyrrell en 2007, cuando para estos mismos estanques el número de *hits* fue menor en el año 2010 (5) (**figura C2.18 c**). Estas diferencias no son atribuibles a una mayor profundidad de secuenciación y apoyan la persistencia de espaciadores para virus que, dentro de un mismo estanque, pueden presentar fluctaciones periódicas en su abundancia que podrían ajustarse a episodios de dinámica virus-hospedador descritos por el modelo *kill the winner* (Thingstad y Lignell 1997**).**

Por otra parte decidimos comparar las secuencias de virus aislados de *S. ruber* con las de los espaciadores de las dos cepas, M8 y SP73. En el caso de M1encontramos un espaciador, situado cerca de la repetición degenerada, para el que identificamos como protoespaciadores dos virus aislados de la cepa M8 (Villamor *et al*., datos no publicados) y otro en el metaviroma

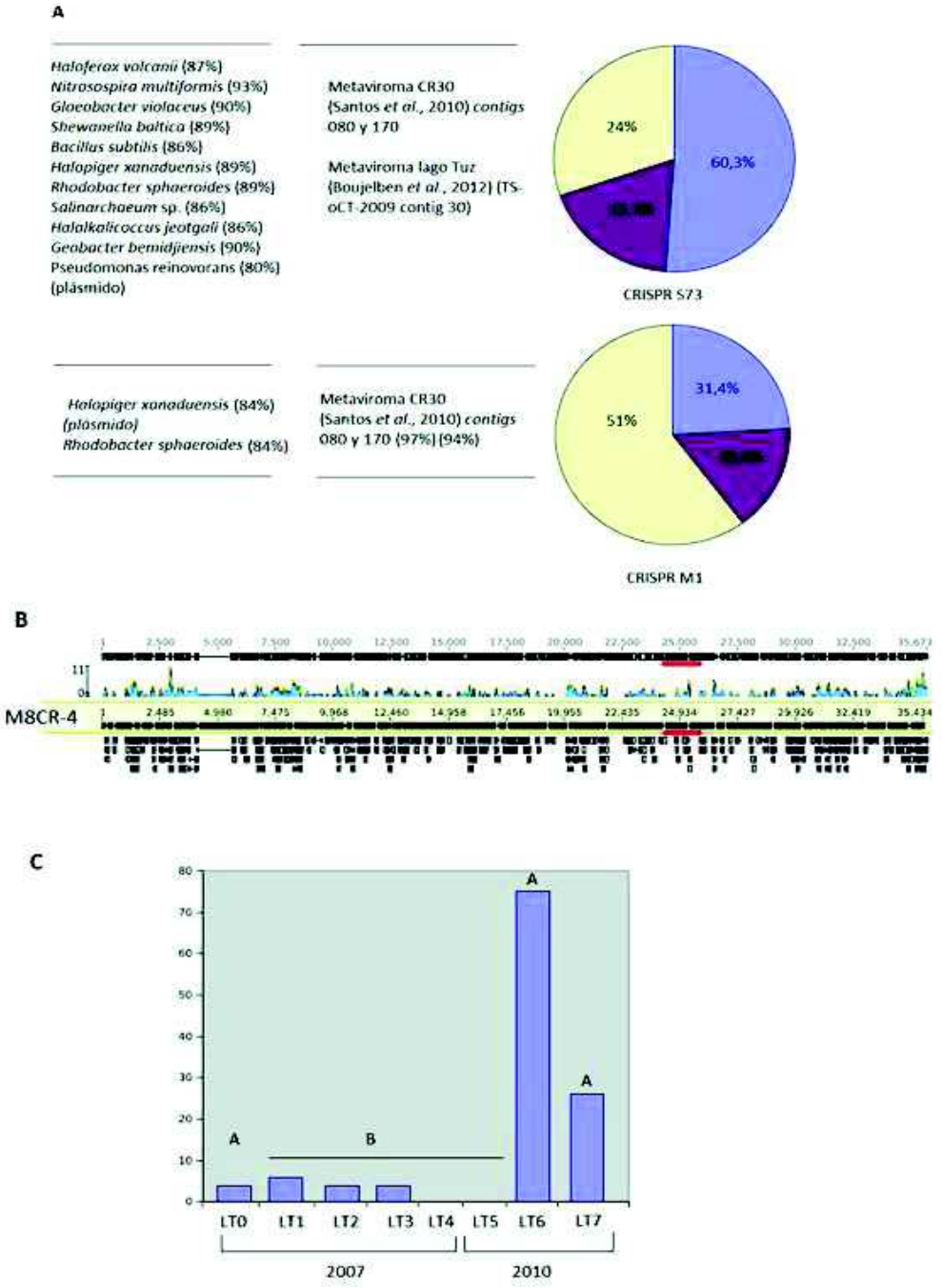


Figura C2.18. Descripción de los sistemas CRISPR-Cas de las cepas M1 y SP73 de *S.ruber*. Figura a: Protoespaciadores identificados desde las bases de datos públicas y proporción de espaciadores con hits con metagenomas y metaviromas ambientales. Figura b: Reclutamiento del virus M8CR-4 contra el metaviroma de San Diego. En rojo se indica el protoespacador. Figura c: Hits de los espaciadores de SP73 contra los metaviromas del lago Tyrrell mostrando diferencias según el año de muestreo.

dirigido de *S.ruber*. M8, M31 y M1 son tres cepas muy próximas filogenéticamente y coaisladas de las Salinas de Campos (Mallorca) en 1999. De las tres, M1 fue la que mostró una mayor resistencia a infección por virus, de hecho no es infectada por ninguno de los dos virus aislados en los que se identificó el protoespaciador ni tampoco por virus de la cepa M31 (Villamor *et al*., datos no publicados), lo que apoyaría la función inmunitaria del sistema CRISPR-Cas en M1 contra infección por virus. Este mismo espaciador también tuvo *hits* con los metaviromas de alta salinidad de San Diego. El reclutamiento de las secuencias de este metaviroma contra el genoma del virus M8CR-4 permitió reconstruir el 85% de su secuencia (**figura C2.18 b**). La región protoespaciadora fue una de las que más fragmentos reclutó lo que también podría indicar que formase parte de un genoma *core* vírico y explicaría la funcionalidad del espaciador, uno de los adquiridos hace más tiempo.

En conjunto los datos anteriores sugieren por una parte la persistencia de algunas especies de virus en el ambiente, así como la especificidad de hospedador, pese a las estrategias de evasión por ambas partes mediante la modificación de la secuencia diana en el caso del virus y sus receptores de superficie por parte de la bacteria. El hecho de encontrar protoespaciadores en metaviromas en localizaciones geográficas distintas sustenta la existencia de especies víricas o genes comunes aunque también de una fracción variable. En el caso de las cepas M1y SP73 de *S. ruber*, el hecho de no encontrar ningún espaciador común sugiere, aunque no necesariamente, la existencia de diferencias biogeográficas en la comunidad vírica de las salinas de Mallorca y Santa Pola y la adaptación de los sistemas CRISPR a las mismas, de acuerdo con el mayor número de *hits* de los espaciadores de las CRISPR de SP73 contra los metagenomas y metaviromas del cristalizador CR30 (**anexo tabla S2.7).**

5. Efecto de la recombinación homóloga sobre el *core* genoma de *S.ruber*.

El alineamiento de los genomas de las 8 cepas aisladas y secuenciadas de *S. ruber* mostró una alta sintenia, que contrasta con las diferencias y reordenamientos observadas dentro de las GI. Estas regiones sinténicas están constituidas 130 bloques colineares (**figura C2.7**) que contuvieron la mayoría de los 2434 genes del genoma *core*. La arquitectura genómica es uno de los aspectos más estudiados en bacterias, ya que localización génica y orden afecta a los niveles

de regulación y expresión génica y por tanto a la fisiología de la especie (Ochman *et al*., 2000, Coleman *et al*., 2006., Mira *et al*., 2010). En este sentido resulta especialmente interesante estudiar los mecanismos evolutivos que actúan sobre el genoma *core* y que mantienen este grado de sintenia ya que se trata además de un fenómeno observado en otros miroorganismos acuáticos de vida libre como *S. baltica*, *A. macleodii*, *S. islandicus* o *H. walsbyi*, en los cuales la recombinación homóloga se postula como uno de los mecanismos que controlan los niveles de diversidad en la fracción del genoma *core*. Estos antecedentes, que incluyen incluso especies con las que *S.ruber* comparte ambiente, junto a los datos aportados por estudios anteriores basados en análisis MLSA (Roselló-Mora *et al*., 2008), sugieren que la recombinación homóloga podría tener un gran impacto sobre los genomas de *S.ruber*.

Con el objetivo de observar la diferencias existentes en el genoma *core* entre las distintas cepas de la especie, el impacto de la recombinación homóloga sobre el mismo y posibles patrones espacio-temporales, se llevó a cabo un estudio de genealogía clonal que permitió clasificar los 8 genomas en 5 marcos clonales (CFs, del inglés *Clonal frame*) empleando el programa ClonalFrame v1.2 (Didelot y Falush 2007). Este programa permitió establecer las relaciones clonales entre las cepas aisladas y estimar la contribución relativa de la recombinación y mutación mediante el cálculo de dos parámetros: la tasa de recombinación/mutación (r/m) y la relación rho/theta (σ/θ) (Vos y Didelot 2009). El término CF se emplea para designar un linaje bacteriano con aparente clonalidad y procedencia de un ancestro común pero sometido a reemplazamientos de fragmentos genómicos, mutaciones y selección con una historia común (Milkman y Bridges 1990., Vos y Didelot 2009). En el caso de *S.ruber* algunas de los 5 CFs incluyeron dos cepas (**figura C2.19**), cuya distribución resultó congruente con los datos de ANI globales y la distribución de tetranucleótidos mostradas anteriormente y en las que no se apreció ninguna distribución biogeográfica aparente. Aquellas cepas incluidas en un mismo CF tuvieron valores de ANI elevados y tetranucleótidos similares y los valores de dN/dS menores que los observados entre cepas de diferentes CFs. Estos resultados sugieren que la recombinación homóloga es más frecuente entre cepas cercanas entre las cuales muchos de los reemplazamientos sinónimos se atribuyen a eventos de recombinación homóloga más que a mutaciones puntuales como se ha observado en especies como *Clostridium difficile* y *Staphylococcus aureus* (Castillo-Ramírez *et al*., 2011) o en *S. islandicus* (Cadillo-Quiroz 2012).

En el caso de *S. islandicus* se detectaron valores de dN/dS mayores entre poblaciones divergentes que entre cepas cercanas dentro de cada una de ellas. De este modo la recombinación homóloga actuaría como mecanismo evolutivo predominante que, en función de la selección purificadora, eliminaría variantes no sinónimas en la población tal como sucede en organismos de vida libre y población efectiva elevada como *E.coli* con valores de dN/dS bajos (0,081) (Jordan *et al*., 2002), en donde además la recombinación homóloga actúa como mecanismo evolutivo relevante (Mau *et al*., 2006). En un estudio evolutivo realizado por Larsson y colaboradores en 2009 con 13 genomas del género *Francisella* se muestra como en la especie *F. tularenis*, patógeno intracelular facultativo con baja población efectiva, apenas se encuentran signos de recombinación homóloga, mientras que en la especie de vida libre *F. novicida* se detectó recombinación homóloga en casi el 20% de su genoma. El dN/dS para *F. novicida* fue de 0,087 mientras que para *F. tularensis* el valor fue próximo a 0,5 reflejando el efecto homogenizador significativo de la recombinación homóloga. Los valores de dN/dS en *S. ruber* serían similares a los encontrados en bacterias patógenas como *H. pylori* (0,158) o *N. meningitidis* (0,188) y menores a los de patógenas obligadas como *Chlamidophyla pneumoniae* (0,568) (Jordan *et al*., 2002), esta última con tamaños de población efectiva bajos y cuyo genoma carece de sistemas de reparación o recombinación (García-González *et al*., 2013). Las relaciones (r/m) y (σ/θ) para *S. ruber* fueron 1,52 y 0,29 respectivamente, mostrando que una mayor proporción de SNPs detectados se deben a eventos de recombinación homóloga más que a mutación. Aunque los eventos de mutación resulten más frecuentes que los de recombinación, los valores de r/m obtenidos atribuyen un 50% más de SNPs a los procesos de recombinación homóloga. La relación r/m obtenida fue elevada si se compara con la observada en otras especies de microorganismos con distintos estilos de vida (Didelot y Vos 2009), y próxima la detectada en organismos extremófilos como *S. islandicus* (Whitaker *et al.,* 2005) y *Halorubrum* sp (Papke *et al*., 2004, 2007) con valores de r/m 1,2 y 2,1 respectivamente, especies en las que la recombinación homóloga es el principal mecanismo de evolución del genoma *core*. Con el fin de confirmar este aparente alto impacto de la recombinación homóloga sobre el genoma *core*, se construyeron árboles de máxima verosimilitud (ML, del inglés *maximum-likelihood*) utilizando regiones sinténicas seleccionadas al azar a lo largo del mismo, incluyendo también las regiones alineables situadas en las fGIs. Los resultados obtenidos (**figuras C2.10 y C2.19**) muestran

como la topología de los árboles varió en función de las regiones seleccionadas. Estas incongruencias filogenéticas confirman que la recombinación homóloga rompe de manera frecuente la estructura clonal de la población como se ha descrito en especies como *A. macleodii* (López-Pérez *et al*., 2013) con niveles de recombinación incluso inferiores. En *S. ruber*. los niveles de r/m fueron similares a los detectados en *Halorubrum* sp. (Papke *et al*., 2004), con la que comparte ambiente y que también presenta una gran población efectiva. Estos niveles se sitúan aproximadamente en la mitad del rango propuesto por Fraser y colaboradores (2007).

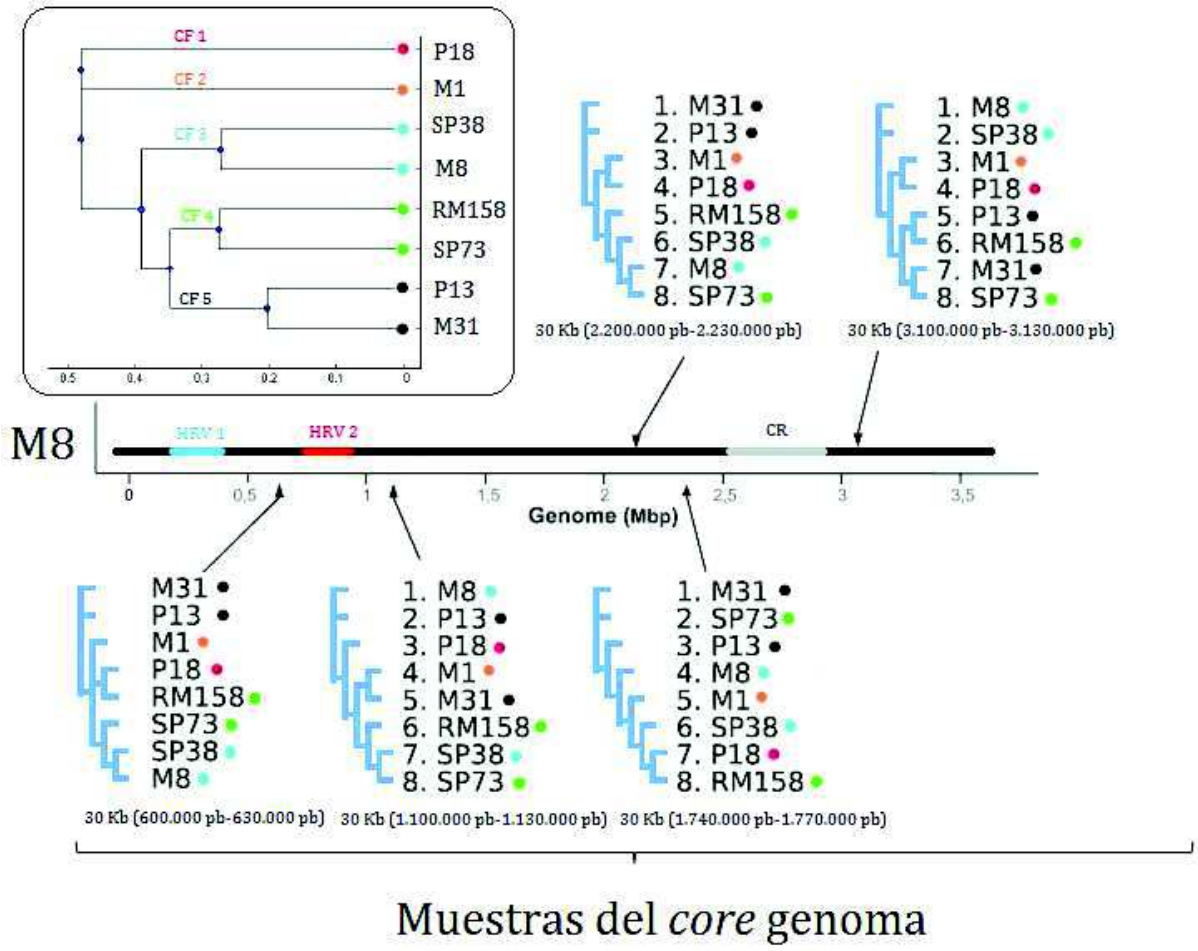


Figura C2.19. Árboles filogenéticos ML para las 8 cepas secuenciadas de *S.ruber*. Enmarcado se muestra el árbol generado a partir del alineamiento del genoma *core* completo y que muestra la distribución delas 5 líneas clonales, cada una identificada con un círculo de color. El resto de árboles filogenéticos se realizaron con regiones de 30 kb seleccionadas al azar dentro del genoma *core*. La incongruencia de estos últimos muestra el efecto de la recombinación homóloga a lo largo del genoma *core*.

Según estos autores, valores de r/m entre 0,25 y 4 indicarían la existencia de poblaciones diferenciables y estables en las cuales se produciría una homogenización o mezcla de líneas clonales fruto de la recombinación homóloga que actuaría fuerza cohesiva entre *clusters* de cepas cercanas. En *Halorubrum sp*. esta homogeneidad estructural se observó en aislados de diferentes puntos geográficos, como Santa Pola y Argelia (Papke *et al*., 2007). Como en el caso de *Halorubrum,* los aislados de *S. ruber* presentan diferente linajes cercanos atendiendo al *cluster* de genes o genes analizados.

Estudios recientes con genomas y metagenomas de microorganismos marinos como *Prochlorococcus* (Luo y Konstantinidis, 2011) y poblaciones de *Crenarchaea* marinas (Konstantinidis *et al*., 2011) muestran una estructura poblacional en *clusters*, que podría responder tanto al efecto de la recombinación homóloga (Fraser *et al*., 2007) como al de la presión selectiva descrita en la teoría de ecotipos (Cohan 2006). Estas poblaciones, que podrían aproximarse a especies, estarían constituidas por un conjunto de genomas de cepas ecológicamente homogéneas con una identidad de secuencia elevada. En estos estudios el reclutamiento de metagenomas contra los genomas de referencia del mismo ambiente muestra una discontinuidad entre el 80-95% de ANI, determinando las diferencias entre genomas de esta población entre menos del 1% y el 5% (Caro-Quintero *et al*., 2011). En el caso de *S. ruber* comprobamos si se observaba este tipo de estructura poblacional fruto de los elevados niveles de recombinación homóloga detectados mediante el reclutamiento de secuencias de los metagenomas de las salinas de Bras del port (Santa Pola) (**figura 2.20**), Campos (Mallorca), lago Tyrrell (Australia) y San Diego (EEUU) contra los cromosomas de las 8 cepas secuenciadas. Como se muestra en la **figura C2.21**, los cuatro ambientes analizados mostraron una distribución normal con máximos de reclutamiento en torno al 98% de identidad y superior al 95% en la mayoría de secuencias. Esta distribución homogénea responde al efecto de la recombinación homóloga sobre el genoma *core* de la especie. En el caso de los metagenomas de los ambientes donde se aislaron las cepas secuenciadas, el porcentaje de secuencias reclutadas por debajo del 95% fue menor indicando cierto grado de divergencia para la especie en puntos geográficamente alejados. Por otra parte, cuando se estimó el reclutamiento de los metagenomas contra los plásmidos se obtuvo un porcentaje considerable de fragmentos con identidades inferiores al 95% (**figura C2.14**). Como elementos móviles del genoma accesorio, los plásmidos contienen una

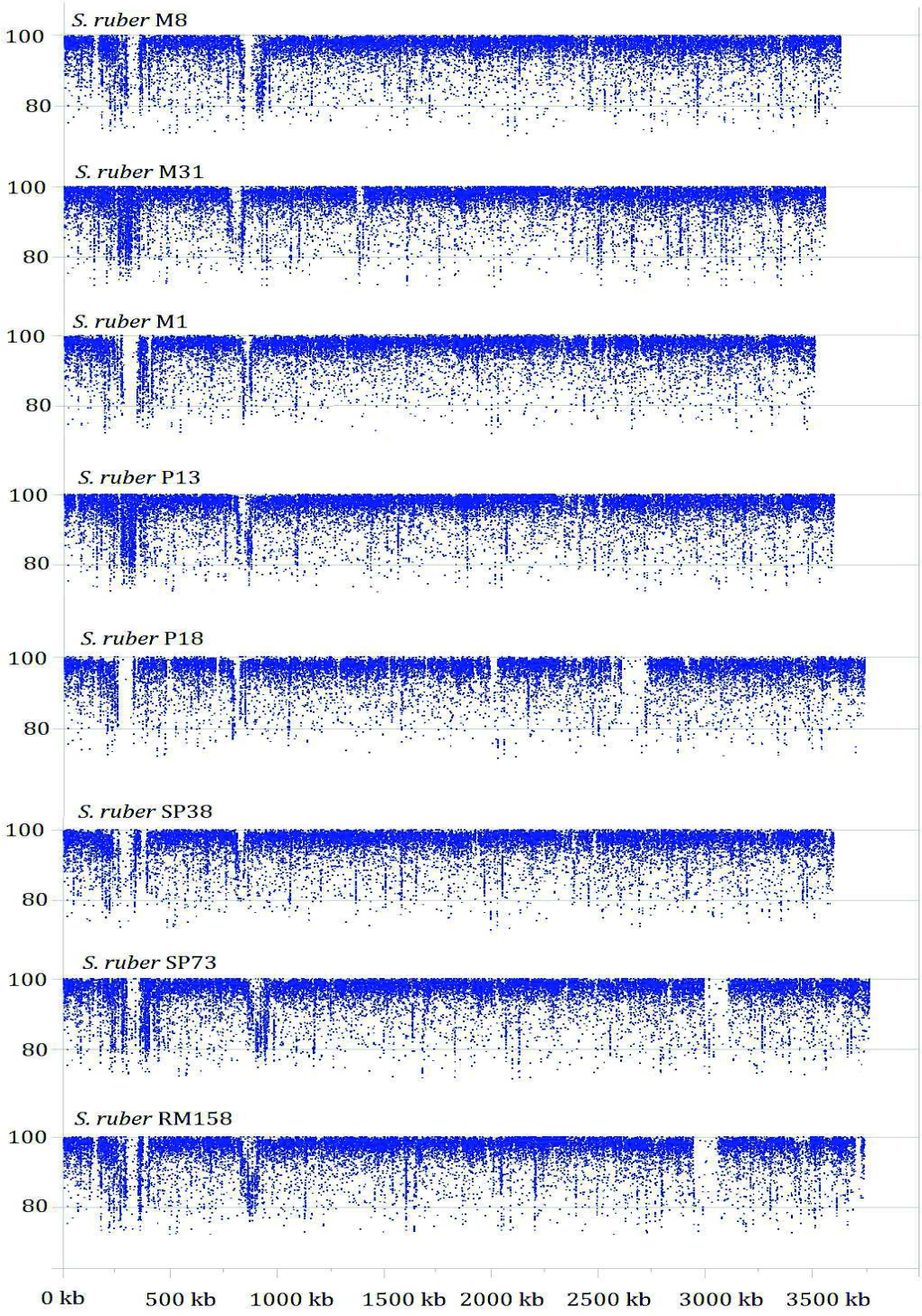


Figura C2.20. Reclutamiento del metagenoma de CR30 de Santa Pola (Ghai *et al*., 2012). La mayoría de *hits* reclutados contra secuencias del genoma *core* lo hicieron con identidades superiores al 95% mientras que los que lo hicieron contra las fGI mostraron identidades menores que este nivel en muchos casos.

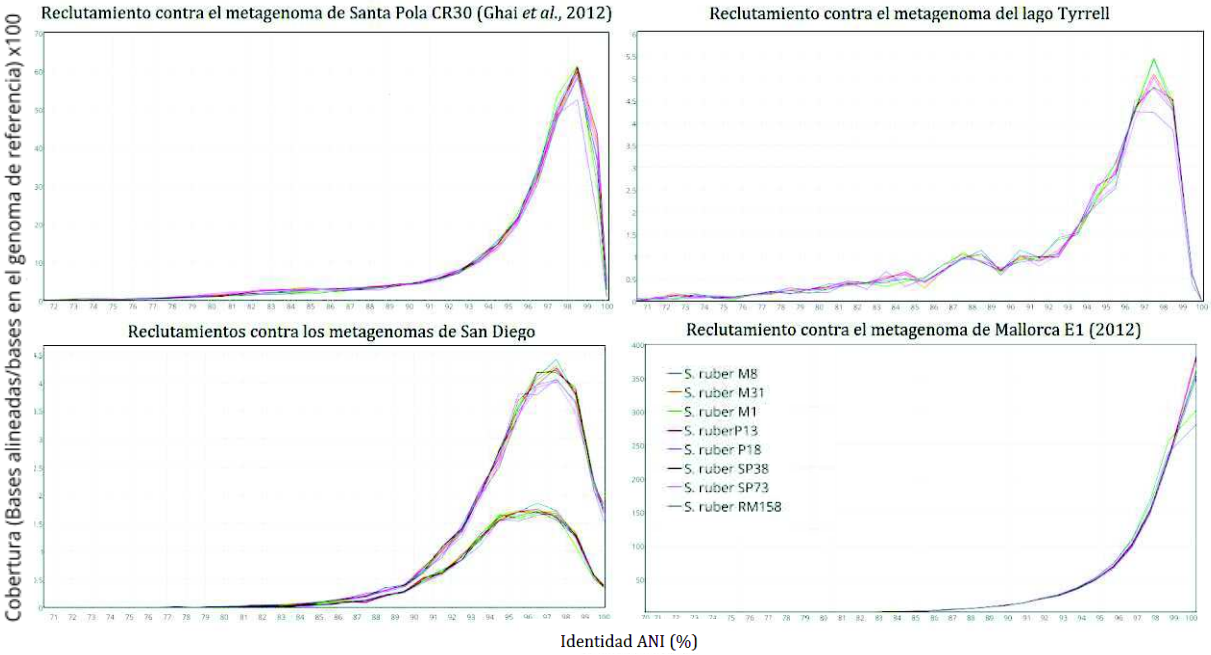


Figura C2.21. Reclutamiento de los metagenomas del cristalizador CR30 de Santa Pola, lago Tyrrell , salinas de San Diego (estanques de alta salinidad) y Salinas de Campos de Mallorca. Los reclutamientos muestran como la mayoría de *hits* se reclutan con identidades superiores al 94%. Esta distribución, similar a la observada en poblaciones discretas de *Prochlorococcus* (Konstantinidis y De Long., 2008) identifica una única población discreta y de *S. ruber,* con una divergencia de secuencia intrapoblacional de entre el 1-6%.

Los niveles de recombinación homóloga en *S.ruber* explicarían las incongruencias filogenéticas observadas en estudios de MLSA llevados a cabo con genes *housekeeeping* (Roselló-Mora et al. 2008), así como la ausencia de patrones biogeográficos claros cuando se compararon los patrones de PFGE (Peña *et al*., 2005). Sin embargo, estos patrones biogeográficos sí se detectaron al analizar diferencias metabolómicas (Roselló-Mora et al., 2008, Antón et al., 2013) y distribución de eventos HGT (Peña *et al*., 2014), probablemente debido a la composición del genoma accesorio recientemente adquirido desde el *pool* ambiental. En el caso de *Halorubrum*, el árbol filogenético del gen del rRNA 16S se mostró totalmente incongruente con el derivado de MLSA lo que sostiene que este *locus* podría ser uno de los más afectados por recombinación homóloga como resultado de su conservación a nivel de secuencia. Aunque tradicionalmente este *locus* y otros implicados en transcripción y traducción de DNA se consideraban resistentes a la HGT (hipótesis dela complejidad) (Jain *et al*., 1999), cada vez son

proporción elevada de genes específicos de cepa lo que, junto a la posibilidad de ser transferidos a diferentes especies, explicaría la mayor divergencia de secuencia y un menor efecto de la recombinación homóloga sobre las mismas más los estudios que muestran la transferencia de rRNA y proteínas ribosómicas (Mau *et al*., 2006., Williams *et al*., 2012., Zhaxybayeva *et al*., 2009). Incluso fracciones de un gen de rRNA de pocas cientos de bases puede ser transferido de manera independiente al operón (Boucher *et al*., 2004), generando ruido al emplear este gen como marcador filogenético. Multiples eventos de conversión génica sobre este gen conducirían al final a la homogenización del mismo, de tal manera que las copias divergentes iniciales serían indetectables. Este mismo fenómeno podría estar afectando al mismo *locus* en *S. ruber*, lo que explicaría la enorme homogeneidad en su secuencia e indicaría que la frecuencia a la que la recombinación homóloga se produce podría estar determinada por el gen particular implicado más que por la divergencia de la cepa, pudiendo producirse múltiples eventos de recombinación entre diferentes líneas clonales simultáneamente (Papke *et al*., 2007).

Existen diversos factores ambientales y genéticos que ayudan a entender el impacto de la recombinación homóloga en la evolución de *S. ruber* y otros organismos extremófilos como *Halorubrum* y *S. islandicus*. Los ambientes extremos están sometidos a condiciones ambientales que pueden afectar a la estabilidad del DNA, condiciones tales como las elevadas temperaturas, concentraciones de sales y metales pesados y la elevada incidencia de radiación UV. Por este motivo, los microrganismos hipertermófilos y halófilos extremos han desarrollado estrategias que favorecen la reparación y estabilidad del DNA, entre los que destacan la maquinaria de reparación y recombinación (García-González *et al*., 2013; revisado en van Wolferen *et al*., 2013). Muchos de estos organismos, que además son competentes, se cree que han adquirido evolutivamente esta capacidad al favorecer la reparación del DNA y el acceso a una fuente importantísima de carbono en medios que contienen elevadas concentraciones de DNA debido a las condiciones fisicoquímicas mencionadas sumada a la lisis por infección viral (Thomas y Nielsen. 2005; Rodríguez-Valera *et al*., 2009; Cadillo-Quiroz *et al*., 2012). *S. ruber* contiene en su genoma *core* el gen *comEA* (*rec*2), uno de los mejor caracterizados en bacterias gram negativas competentes y que se sabe conforma el canal que internaliza el DNA en la célula controlando la tasa de transferencia (Takeno *et al*., 2011). Además en el genoma *core* también encontramos la presencia del sistema de reparación y recombinación *Rec*FOR, análogo al

*Rec*BCD, y el SOS. El sistema *Rec*FOR tiene funciones similares a las de las restrictasas, diferenciando entre el DNA propio y el ajeno por medio de un elemento en *cis*, la secuencia Chi, ausente del DNA viral pero presente en el genómico. Esta secuencia es específica de grupos de bacterias y actúa como un código de barras contribuyendo al mantenimiento de los genomas *core* (Halpern *et al*., 2007). El sistema actúa como una restrictasa con el DNA ajeno, y como sistema de reparación con el DNA homólogo reparando roturas de DNA de doble cadena mediante recombinación homóloga.

Por otra parte, en el caso de *S. ruber* encontramos una presencia pobre de sistemas MR y CRISPR-Cas. Atendiendo a su tamaño genómico, las cepas de *S.ruber* deberían contener al menos entre tres y cuatro sistemas MR por genoma tal como muestran estudios previos (Vasu y Nagaraja 2013). Organismos competentes como *H. pylori, H.influenzae o S. pneumoniae* presentan abundancia y diversidad de sistemas MR. Sin embargo en *S. ruber* su distribución es similar a la encontrada en endosimbiontes y patógenos intracelulares como *Buchnera, Chlamydia, Chlamydophila, Coxiella*, los cuales apenas se enfrentan a infecciones víricas y tienen tasas de HGT bajas. Se sabe que microorganismos que carecen de sistema *Rec*BC tienen una distribución más elevada de sistemas MR, por lo que las recombinasas podrían suplir el papel del sistema *Rec*BC (Vasu y Nagaraja 2013). Del mismo modo un sistema *Rec*FOR adaptado y optimizado junto a un sistema de envueltas celulares variable y complejo, definido por la distribución de genes de envoltura presentes en las fGI, podrían actuar como sistemas de defensa principales en *S. ruber* frente a la infección vírica. De hecho se ha demostrado que el sistema *Rec*FOR es capaz de suplir la pérdida de sistemas MR (Handa *et al*., 2009). La ausencia de sistemas CRISPR-Cas en algunas cepas, pese a la presión selectiva de virus del ambiente, se justifica en publicaciones previas por la mayor complejidad y diversidad de proteínas de envuelta y glicosilasas codificadas en sus fGIs (Martín-Cuadrado *et al*., 2015). Se ha demostrado además en organismos extremófilos como *Halobacerium* sp que en ambientes hipersalinos o termófilos la maquinaria de recombinación, incluyendo el sistema SOS, (Papke et al., 2007, McCready *et al*.,2005) y sistemas pili tipo IV (Ajon *et al*., 2011) pueden inducirse por alta radiación UV favoreciendo la recombinación homóloga. En el caso de *Archaea* como *H. volcanii* y *H. salinarum* se ha observado una marcada poliploidía en fase exponencial (Breuert *et al*., 2006), mecanismo adicional que podría favorecer la integridad cromosomal mediente recombinación de

replicones homólogos. En conjunto, la expresión de sistemas MR y restrictasas, junto a la especificidad de los sistemas SOS y *Rec*FOR en la recombinación de secuencias homólogas, contribuiría al mantenimiento de la identidad de especie, actuando como mecanismos barrera frente a la entrada de DNA de otras especies y su fijación (Murray, 2002, Jeltsch 2003). De este modo estos mecanismos estarían favoreciendo la recombinación homóloga por selección negativa de fragmentos heterólogos. Además, en el caso de entrada de DNA de especies cercanas, la generación de fragmentos de restricción cortos favorecería su recombinación homóloga generando variabilidad génica (Price y Bickle 1986). Las cepas de *S. ruber* que se agruparon en una misma CF tuvieron el mismo contenido en sistemas MR, y niveles de recombinación homóloga parecidos. Estos niveles fueron mayores en aquellas cepas que contuvieron un sistema MR (M31 y P13 en el CF5, y SP38 y M8 en el CF3) (**figura C2.19**).

Por último, y para valorar el impacto directo de la recombinación homóloga sobre el genoma *core* y los *clusters* de genes individuales, se realizó un análisis de fragmentos recombinados entre los genomas de *S. ruber* secuenciados empleando el programa RDP4 (Martin *et al*., 2011). Se detectaron en total 958 eventos de recombinación que en promedio abarcaron un 31% del cromosoma de las cepas analizadas. El 75% (725/958) de los eventos detectados comprendieron zonas de al menos 10 kb, 211 abarcaron regiones entre 10-50 kb, 22 regiones superaron las 50kb y 6 superaron las 100 kb. Los eventos de recombinación detectados en general fueron superiores en tamaño y número a los detectados en análisis similares con cepas de *E. coli* (Mau *et al*., 2006) o *A.macleodii* (López-Pérez *et al*., 2013), lo que destacaría la relevancia de este mecanismo en *S. ruber*. En la naturaleza se han detectado eventos de transferencia de grandes regiones mediante recombinación homóloga en organismos con diferentes estilos de vida como *E. coli* (Mau *et al*., 2006), *S. aureus* (Castilllo-Ramírez *et al*., 2011) o *Archaea* halófilas del género *Haloferax* (Naor *et al*., 2010).

La distribución de estos eventos a lo largo del cromosoma y en las diferentes cepas no fue homogénea. En algunas cepas como P13, M8 y SP38 la recombinación homóloga abarcó más del 40% del cromosoma, mientras que en otras como P18 implicó solo al 10,4%. Curiosamente esta última cepa constituye su propia CF, tiene las HRVs más cortas de entre las cepas secuenciadas, carece de plásmidos y es la que compartió menos secuencias con plásmidos de otras cepas. Todo esto nos lleva a pensar que quizá codifique en su secuencia algún sistema barrera adicional que

no se haya caracterizado.

La detección de eventos de recombinación fue congruente con la topología del árbol de CF (**figura C2.19**) y los valores de ANI y TETRA. Aunque el programa empleado pudo resolver más señales de recombinación desde líneas CF más divergentes, CF1y CF2, hacia el resto, muy probablemente la frecuencia sea mucho mayor entre cepas de una misma CF o entre CF cercanas. Esto justifica que zonas con dN/dS próximas a 0 como es el caso de la CR entre M8 y M31 con 385 Kb de longitud (Peña *et al*., 2010) no se hayan resuelto por este tipo de análisis, pese a que estudios recientes han demostrado la relación de estas regiones sinténicas del genoma *core* entre cepas cercanas con bajos dN/dS y la recombinación homóloga (Castillo-Ramírez *et al*., 2011). De hecho, 4 de los 6 eventos de recombinación mayores de 100 kb detectados en *S.ruber* se detectaron entre las CFs 3, 4 y 5 e incluyeron la HRV1, una de las zonas más divergentes lo que favorecería la detección de eventos de recombinación entre CF cercanas y muestra el impacto de este mecanismo entre cepas cercanas. Dos de estos 4 eventos, de 210kb y 119 kb, se dieron desde la CF3 a la CF4, uno de 206 kb desde CF4 a 5 y el cuarto de 166 kb desde CF5 a CF4. Esto explicaría las semejanzas estructurales de estas fGI entre cepas como M8 y RM158 (véase apartado 4.1).

Se identificaron eventos de recombinación a lo largo de los cromosomas de las 8 cepas (**figura C2.22**) de acuerdo a los datos observados en los análisis filogenéticos (**figura C2.19**), aunque su distribución no fue homogénea. Se identificaron menos eventos cerca de las HRVs, algo que se observó en las regiones próximas de las fGI de *A. macleodii* (López-Pérez *et al*., 2014), en especial la HRV1. Por el contrario, las regiones sinténicas situadas entre ambas HRVs y la correspondiente a la región donde se ubica la CR (posiciones 2.583.792 pb y 2.969.363 en M8) acabar la frase concentraron una alta proporción estos eventos. Un 75% de los eventos de recombinación (719/ 935) contuvieron entre 1 y 9 genes, observando hasta más de 100 genes en los eventos de mayor tamaño. el análisis funcional de los genes situados en las regiones recombinadas reveló un empobrecimiento en genes de 10 categorías COG (Test de Fisher, $p<0.05$; $FDR<0.05$) (**anexo, tabla S2.6**): J (Traducción, estructura ribosómica y biogénesis), N (Motilidad celular), O (Modificación postranscripcional), y 7 de las 8 categorías relacionadas con el metabolismo celular C (Producción y conversión energética), transporte y metabolismo de aminoácidos (E), nucleótidos (F), coenzimas (H), lípidos (I), iones inorgánicos (P) y

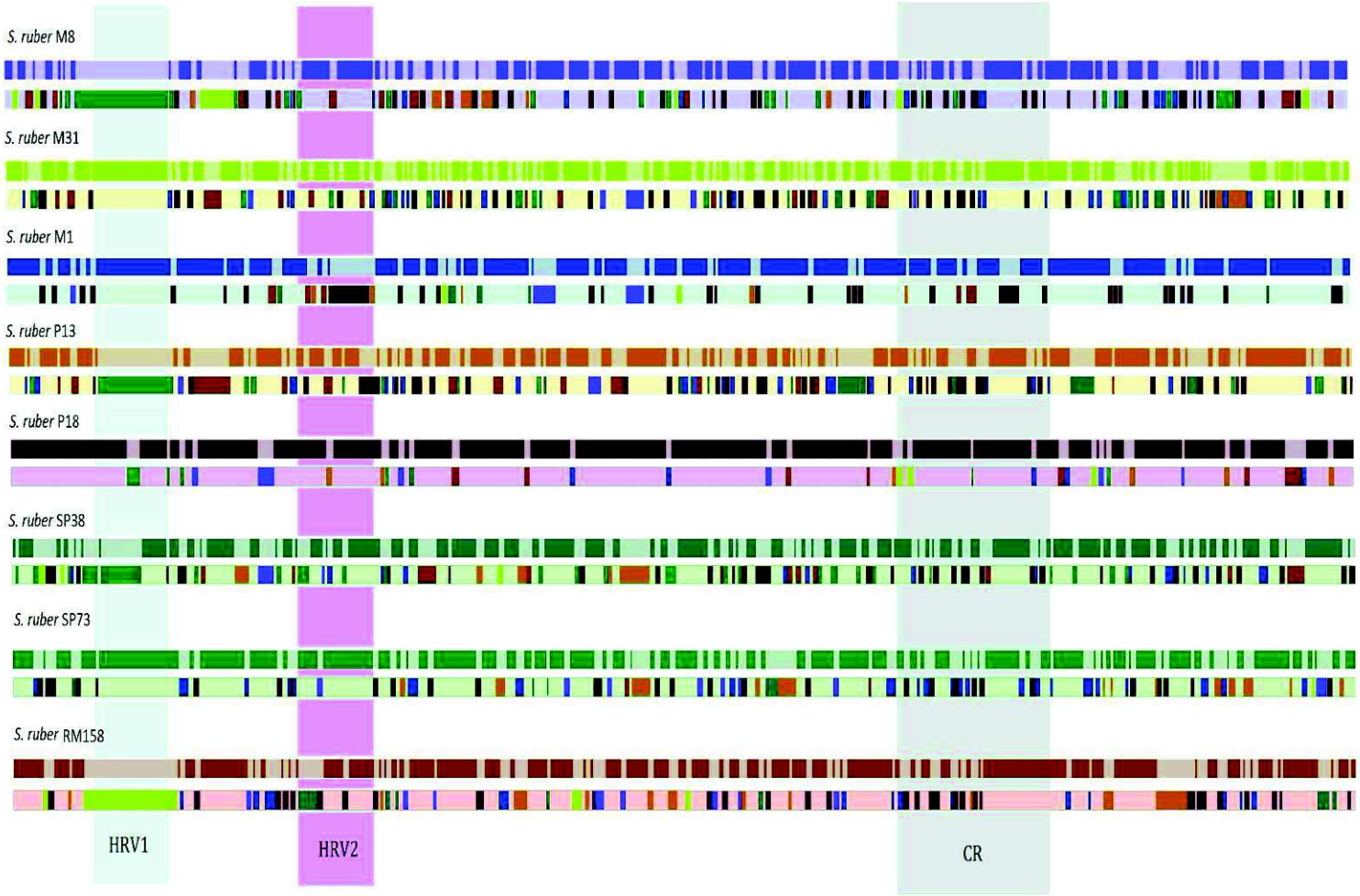


Figura C2.22. Representación de los eventos de recombinación identificados en los cromosomas de las 8 cepas secuenciadas de *S.ruber* con el programa RDP4. Cada cromosoma se identifica con un color. Para cada cepa, la barra horizontal superior muestra las regiones recombinadas en un color mas tenue, y la barra inferior en diferentes colores la cepa donadora del fragmento recombinado. Las bandas de color verticales indican la posición de las regiones hipervariables (HRV1 y HRV2) y la zona conservada (CR).

metabolitos secundarios (Q). Entre las categorías funcionales con una mayor representación en estas regiones encontramos 4: Replicación, recombinación y reparación (L), tráfico intracelular, secreción y transporte vesicular (U), transporte y metabolismo de carbohidratos (G), función general (R,) y de pobre caracterización (S).

Los resultados obtenidos para los COG J y L, involucrados en procesamiento básico de la información celular, fueron parecidos a los descritos en un análisis comparativo realizado con 6 cepas de *E.coli* (Mau *et al*., 2006) en donde genes implicados en la biosíntesis de proteínas y RNA (transcripción y traducción) no eran susceptibles de recombinar mientras que los implicados en reparación, replicación y recombinación mostraban una elevada tasa de sustitución alélica. Dentro de la categoría COG L observamos que un 58% (29/50) corresponden con genes que codifican para la proteínas XerD en los 8 cromosomas, la presencia de 6/8 genes codificantes para la proteína ComE (con una copia por cromosoma) y 5/8 copias de RecR, esta última implicada en procesos de reparación de DNA. La presencia de genes codificantes para XerD confirma su implicación en los procesos de recombinación homóloga específica de sitio tal como ya apuntaban las evidencias mencionadas en apartados anteriores al hablar de los procesos de dinámica dentro de fGI y plásmidos. Entre las COGs involucradas en procesos celulares y señalización y metabolismo celular encontramos un enriquecimiento en términos relativos a la glicosilación (dentro de la categoría COG M) y un enriquecimiento en COG G, que demuestra que la recombinación homóloga es uno de los mecanismos predominantes en el intercambio de *clusters* involucrados en la generación de diversidad de envueltas celulares, probablemente frente a la presión por virus ambientales. De esta manera mantendría la diversidad entre cepas observada en estudios metabolómicos previos (Roselló-Mora *et al*. 2008., Antón *et al*., 2013) y en la expresión diferencial en cultivo mixto en el capítulo 1 de esta tesis, mediante el barajado de alelos. Dentro de la categoría COG C, encontramos 9 de las 32 rodopsinas presentes en el genoma *core* de las 8 cepas. Los términos COG más abundantes dentro de las categorías COG T, G y H correspondieron a genes implicados en elementos de control del estrés oxidativo, tales como el gen de la proteína de estrés universal de unión a nucleótidos UpsA, transportadores de solutos compatibles como la glicín-betaína, y genes implicados en la biosíntesis de ubiquinonas/metaquinonas protectoras de envueltas celulares. Dentro de las categorías COG P y H, muchos de los eventos de recombinación tuvieron lugar en transportadores ABC para Fe o

grupos hemo, de forma similar a lo que se observó en un estudio de recombinación con *A. macleodii* (Gonzaga *et al*. 2012), y en algunos receptores TonB de membrana externa para cobalamina, lo que estaría indicando una fuerte selección positiva de los eventos de recombinación que proporcionan capacidades de transporte extra. Más del 70% de los genes que recombinaron en la categoría COG V estuvieron relacionados con sistemas de transporte multidroga o resistencia a antibióticos, los cuales presentaron también expresión diferencial entre M8 y M31 en cultivos mixtos y podrían estar involucrados en procesos de comunicación celular o competencia. En este sentido, encontramos 22 genes implicados en el transporte de bacteriocinas, que inhiben el crecimiento de cepas cercanas. La recombinación de estos genes, y su posible implicación en la comunicación intercelular como se vio al final del capítulo 1, podría modular la interacción entre cepas de *S.ruber* con diferentes capacidades metabólicas y fisiológicas.

En conjunto, los resultados obtenidos muestran el gran impacto que la recombinación homóloga tiene sobre las regiones sinténicas del genoma *core* de *S. ruber*. El aumento en el número de genomas secuenciados ha permitido observar grandes regiones sinténicas en otras especies que del mismo modo podrían estar sujetas a recombinación homóloga. Estos niveles de sintenia en bloques colineares conservados en otras especies podrían ser el resultado de procesos evolutivos complejos que optimizaran una arquitectura del pangenoma, puesto que diferencias intraespecíficas en la arquitectura genómica pueden influir en su fenotipo pese a tener el mismo repertorio génico afectando a los mecanismos reguladores (Mira *et al*., 2010). Si los niveles de recombinación son adecuados para dirigir la evolución del genoma *core* es una cuestión de que debe abordarse de manera individual en cada una de ellas, así como los factores que pueden estar afectando tales niveles, poniendo especial atención en los mecanismos barrera mencionados a lo largo de este capítulo.

Introducción

Objetivos

Materiales y métodos

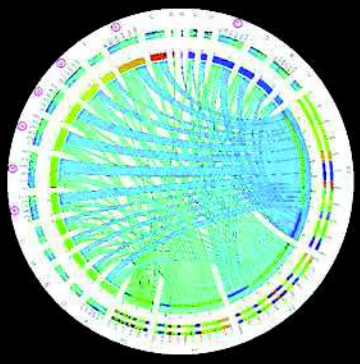
Resultados y discusión

Capítulo 1

Análisis de las diferencias transcripcionales e interacción de cepas cercanas de *S.ruber* mediante RNAseq.

Capítulo 2

Estudio de los mecanismos y estrategias de diversificación genómica en *S. ruber*



Capítulo 3

Impacto de la recombinación homóloga sobre la evolución de genomas *core* procariotas

Conclusiones

Bibliografía

Anexos

Resumen

Algunos de los estudios metagenómicos y genómicos en cepas de una misma especie llevados a cabo en los últimos años sugieren que la recombinación homóloga intraespecífica podría jugar un papel importante en la microevolución de muchos microorganismos, pudiendo llegar a ser éste, en ocasiones, el mecanismo predominante. Sin embargo, estos estudios son todavía escasos y de distribución heterogénea a lo largo de la filogenia procariota, y por lo que son muchas las incógnitas sobre su efecto en microorganismos con diferentes estrategias de vida.

Este capítulo aborda el estudio del impacto de la recombinación homóloga en 338 cepas de 54 especies procariotas abarcando una amplia distribución de grupos filogenéticos y estrategias de vida. Con el objetivo de elucidar los factores más relevantes que determinan el impacto y distribución de la recombinación homóloga se obtuvieron datos genómicos y ecológicos, presencia de elementos que favorecen o dificultan la transferencia horizontal como por ejemplo contenido en sistemas RM, CRISPR-Cas o genes *com*. Los resultados del análisis con genomas completos permitieron identificar un total de 16.000 eventos de recombinación, la mayoría de ellos, más del 80%, de longitudes menores a las 10Kb aunque algunos de tamaños considerables y similares a los de la CR de *S. ruber* y a tamaños documentados en estudios previos.

La ecología o modo de vida fue uno de los factores que más se asociaron a la distribución de estos fragmentos, encontrándose mayor número de eventos de recombinación en patógenos oportunistas qué en patógenos intracelulares y simbiontes. Los patrones de intercambio génico por recombinación homóloga mostraron una clara estructuración funcional acorde con las distintas estrategias ecológicas y procesos de adaptación. Además las especies competentes y con mayor contenido en sistemas MR recombinaron en mayor medida e intercambiaron fragmentos de mayor tamaño, detectándose una posible relación coevolutiva entre la competencia y la presencia de sistemas CRISPR-Cas. Por último este estudio ha permitido confirmar con genomas completos de un número considerable de cepas que la recombinación homóloga contribuye a la homogenización de los genomas *core*, manteniendo la cohesión poblacional e influyendo notablemente en la microevolución de algunas de estas especies.

1. Introducción.

Tras su mejora en calidad y abaratamiento de costes, las tecnologías de secuenciación de nueva generación (NGS) se han convertido en una herramienta potente y versátil a la hora de afrontar nuevos retos en el estudio de microorganismos procariotas. Durante la última década las bases de datos de secuencias procariotas han experimentado un crecimiento exponencial. En el caso de la genómica de procariotas ha sido muy notable el avance experimentado desde el año 2005, en el que se contaba con poco más de 200 genomas secuenciados, en su mayoría de patógenos relevantes. Más recientemente, la secuenciación de varios genomas de una misma especie está permitiendo desarrollar estudios de genómica comparativa completos, definiendo las características genómicas de diferentes especies, en la actualidad incluso en organismos de vida libre e importancia ecológica.

Estos avances han abierto nuevas perspectivas en el análisis genómico comparativo, presentando respuestas a cuestiones concretas como la descripción de los genomas *core* y accesorio de algunas especies, o el estudio microevolutivo de especies particulares. En este sentido, el incremento en el número de genomas completos secuenciados está permitiendo explorar estructuras poblacionales y su evolución, aspectos aún por abordar en la mayoría de especies procariotas (Polz y Hanage., 2013). Un ejemplo de este tipo de aproximación lo muestran estudios recientes que emplean la genómica comparativa a metagenomas y genomas completos para abarcar el estudio de la microdiversidad y la estructura poblacional a nivel de especie e incluso dentro de una misma población (Mira *et al*., 2010, Caro-Quintero *et al*., 2011).

Entre estos últimos, destacan los trabajos realizados con metagenomas de ambientes naturales acuáticos en especies como *Prochlorococcus marinus* (Coleman *et al*., 2006) o *Haloquadratum walsbyi* (Cuadros-Orellana *et al*., 2007, Tully *et al*., 2015), que demuestran que ambas especies presentan una estructura poblacional en *clusters*, constituidos por un conjunto de secuencias con una identidad elevada por encima de un umbral, normalmente entre 80-95% respecto a un genoma de referencia del mismo ambiente, por debajo del cual se observa una discontinuidad. Por otra parte y, aunque escasos todavía, los estudios basados en genomas completos de coaislados apuntan en el mismo sentido a la existencia de estos *clusters*,

cuantificando las diferencias dentro de genomas de la misma especie en un mismo en un rango de ANI ente menos del 1% y el 5% como mucho (Konstantinidis y DeLong 2008; Papke *et al*., 2004, 2007). Estos estudios muestran niveles de identidad de secuencia mayores que los observados al comparar los genomas *core* completos secuenciados de diversos lugares, donde se ha observado que hasta cerca del 30% de los genes pueden constituir el genoma accesorio (Caro-Quintero *et al*., 2011).

En conjunto, los datos anteriores revelan la coexistencia de cepas ecológicamente homogéneas, que presentan pequeñas diferencias genómicas a nivel del genoma *core*. Estos niveles de microdiversidad, observados en el genoma *core* de cepas cercanas de diversas especies, junto al genoma accesorio, constituyen un reservorio importante en procesos de adaptación y evolución tal y como se expone en **hipótesis del genoma *core*** (Lan y Reeves., 2001). Estas poblaciones discretas o *clusters* tienen por tanto una diversidad de secuencia y de contenido génico importantes, mucho mayor a la observada en líneas clonales. Este tipo de patrón o *clustering* se ha observado en varios grupos taxonómicos, entre los que se incluyen representantes de *Crenarchaea, Cyanobacteria y Proteobacteria*. Se han encontrado patrones de diversidad similares a los descritos en especies oceánicas también en otros ambientes como la microbiota humana asociada a la garganta (Qin *et al*., 2010), *biofilms* de bacterias reductoras de hierro (Tyson *et al*., 2004), surgencias termales (Bhaya *et al*., 2007), biorreactores de interés industrial (García Martin *et al*., 2006), especies halófilas extremas como se detalla en la introducción del capítulo 2, y comunidades microbianas dulceacuícolas (Oh *et al*., 2011). Así pues, parece que cuando un conjunto de organismos genéticamente homogéneos se mantienen en un nicho estable, y por tanto comparten la misma trayectoria ecológica, como sucede en los ambientes oceánicos y las especies acuáticas nombradas anteriormente, se generan *clusters* discretos que pueden aproximarse a especies. Estos *clusters* son unidades de diversidad microbiana presentes en una comunidad que comprenden genotipos claramente distinguibles de otros con los que cohabitan. Estudios metagenómicos en *Prochclorococcus* (Luo y Konstantinidis., 2011) y poblaciones de Crenarchaea marinas (Konstantinidis 2011) muestran que estas poblaciones o *clusters* no se limitan a un único emplazamiento geográfico, sino que

pueden abarcar *habitats* interconectados que se caractericen por tener propiedades fisicoquímicas similares (Caro-Quintero y Konstantinidis., 2011).

Una de las cuestiones a abordar en diferentes especies bacterianas es el mecanismo por el cual se generan este tipo de *clusters* y qué fuerzas llevan a la cohesión de los genotipos de una población. Entre las principales teorías encontramos el modelo neutral (Fraser *et al*., 2007) que destaca el papel de la recombinación homóloga en la convergencia y divergencia de *clusters,* y las que explican este tipo de *clustering* en el marco de la teoría de ecotipos, según la cual estos se generarían tras procesos de selección natural periódica y/o deriva génica dando lugar a grupos con características ecológicas similares (Cohan 2001; 2006) (véase introducción, apartado 1.2.2). Aunque existen otras teorías y modelos que tratan de explicar este fenómeno mediante procesos de cuello de botella o periodos de crecimiento poblacional y extinción periódicos (revisado en Vos 2009), estos sólo son aplicables a unos pocos *habitats*, situaciones y filotipos.

En base a los resultados mencionados en diversos *habitats*, las dos teorías anteriores cobran fuerza, aunque cabría determinar la contribución relativa de cada uno de los dos procesos, que quizás difiere en ambientes o clados dentro de una especie. La selección periódica, causada por perturbaciones ambientales, podría explicar diferencias en diversidad observadas en poblaciones sometidas a diferentes presiones ambientales. Por ejemplo, las poblaciónes de *Crenarchaea* de aguas superficiales del mar de los Sargazos en aguas superficiales presentan una diversidad mucho menor que las de 4000 metros de profundidad (98-100% de ANI reclutado frente a 90-100%) (Konstantinidis y de Long., 2008). Las perturbaciones en las condiciones fisicoquímicas en aguas superficiales, son mucho mayores que en las profundidades, donde el ambiente es muy estable y los procesos de selección periódicos podrían explicar el patrón de diversidad observado. Por otra parte existen diversos estudios que muestran niveles detectables de recombinación homóloga en distintas poblaciones terrestres naturales formadoras de *biofilms* (Tyson *et al*., 2004; Vergin *et al*., 2007) y planctónicas marinas (Vergin *et al*., 2007; Konstantinidis y DeLong 2008; López-Pérez *et al*., 2014). En muchos casos la recombinación homóloga no afecta a genes concretos bajo selección positiva sino a prácticamente todo el genoma, por lo que podría actuar como fuerza cohesiva y homogeneizadora de la población (Caro-Quintero *et al*., 2011). Esto, junto al hecho de que los niveles de recombinación homóloga

decrecen con la divergencia de secuencia mostrando un declive significativo para valores de ANI inferiores al 90-95% (Thomas y Nielsen 2005), podría indicar su papel importante en la evolución de poblaciones microbinanas.

Estudios recientes empleando genomas completos de *S. baltica* (Caro-Quintero *et al*., 2011) confirman que la recombinación homóloga actúa como mecanismo importante de cohesión. Trabajos similares en *Streptococcus pneumoniae* indican que es el principal mecanismo evolutivo que actúa sobre el genoma *core* durante los procesos adaptativos en infecciones policlonales, en las que cepas pertenecientes a diferente linajes participan en el proceso infectivo (Hiller *et al*., 2010). Sin embargo, hasta la fecha son escasos los análisis de recombinación homóloga de este tipo debido al número de genomas completos (**tabla C3.1**) y las limitaciones técnicas a la hora de detectar fragmentos recombinantes entre cepas muy próximas (Didelot y Martin 2010, Martin *et al*., 2011). La mayoría de análisis se basan en aproximaciones con MLSA considerando unos pocos *loci* (revisado en Vos y Didelot 2009) (**tabla C3.2**). Este hecho, junto a la falta de homogeneidad en los criterios y trabajos publicados, como ya indican algunas revisiones (Vos y Didelot 2009; Didelot y Maiden 2010), dificulta la

Tabla C3.1. Revisión de los valores de r/m y rho/theta en estudios previos con genomas completos. Entre paréntesis se muestra la clase ecológica correspondiente a la base de datos de este capítulo: (0): Simbiontes/patógenos intracelulares, (1): no patógenos (comensales y vida libre), (2): patógenos obligados y (3): patógenos oportunistas.

Especie	Filo/division	Nº genomas	r/m	95% CI	rho/theta	95% CI	Referencia
Bacillus cereus	Firmicutes	13	2,41	(2,37-2.45)	0,21	(0,20-0,23)	Didelot *et al.*, (2010)
Alteromonas macleodii	g-proteobacteria	6	0,6	(0,45-0,91)	0,01	(0,01-0,02)	López-Pérez *et al.*, 2014
Francisella tularensis	g-proteobacteria	12	0,8	----	0,07	----	Larsson *et al.*,(2009)
Chalmidia trachomatis	Chlamydiae	12	0,71	(0,56-1,01)	0,07	(0,05-0,11)	Joseph et al., (2011)
Bartonella sp.	a-proteobacteria	**	1	----	----	----	Guy *et al.*, (2012)
Ricketsia dadantii	e-proteobacteria	24	0,355	(0,025-0,037)	0,034	(0,325-0,514)	Hernández-López *et al.*, (2013)
Helicobacter pylori	e-proteobacteria	7	1,2-1,62	----	1,2-2,5	----	Kennemann *et al.*, 2010

** Empleo de regiones genómicas amplias aunque no genomas completes.

Tabla C3.2. Revisión de los valores de r/m obtenidos por MLSA en estudios previos. Entre paréntesis se muestra la clase ecológica correspondiente a la base de datos de este capítulo. Para cada estudio se muestra el número de genes empleado (n *loci*) y las secuencias de tipado diferentes (n STs).

Especie	Filo/division	Ecología	n STs	n loci	r/m	95% CI	Referencia
Bacillus cereus	Firmicutes	Patógeno oportunista (3)	13	6	0.7	(0.2–1.6)	Sorokin *et al.*, (2006)
Bacillus cereus	Firmicutes	Patógeno oportunista (3)	111	7	0.9-1.5	----	Didelot *et al.*, (2009)
Bacillus thuringiensis	Firmicutes	Patógeno obligado (2)	22	6	0.8	(0.4–1.3)	Sorokin *et al.*, (2006)
Bacillus weihenstephanensis	Firmicutes	Comensal - Vida libre (1)	36	6	2	(1.3–2.8)	Sorokin *et al.*, (2006)
Bartonella grahanii	a-proteobacteria	Simbiontes / Pat. Intracel. (0)	63	6	3.77	----	Buffet *et al.*, (2013)
Bartonella henselae	a-proteobacteria	Patógeno obligado (2)	14	0.1	0.1	(0.0–0.7)	Arvand et al., (2007)
Bartonella taylorii	a-proteobacteria	Simbiontes / Pat. Intracel. (0)	63	6	6.81	----	Buffet *et al.*, (2013)
Bordetella pertussis	b-proteobacteria	Patógeno obligado (2)	32	0.2	0.2	(0.0–0.7)	Diavatopoulos *et al.*, (2005)
Brachyspira sp.	Spirochaetes	Patógeno oportunista (3)	36	0.2	0.2	0.1–0.4	Rasback *et al.*, (2007)
Burkholderia pseudomallei	b-proteobacteria	Patógeno obligado (2)	106	22	4.5-8.5	----	Nandi *et al.*,, (2015)
Campylobacter insulaenigrae	e-proteobacteria	Patógeno oportunista (3)	59	7	3.2	(1.9–5.0)	Stoddard *et al.*, (2007)
Campylobacter jejuni	e-proteobacteria	Patógeno oportunista (3)	>1000	7	6.76	(6.08-8.09)	Yu *et al.*, (2012)
Campylobacter jejuni	e-proteobacteria	Patógeno oportunista (3)	110	7	2.2	(1.7–2.8)	pubmlst.,org
Campyobacter coli	e-proteobacteria	Patógeno oportunista (3)	>1000	7	1.01	(0.78-1.30)	Yu *et al.*, (2012)
Chlamydia trachomatis	Chlamydiae	Simbiontes / Pat. Intracel. (0)	14	0.3	0.3	(0.0–1.8)	Pannekoek *et al.*, (2008)
Clostridium difficile	Firmicutes	Patógeno oportunista (3)	34	0.2	0.2	(0.0–0.5)	Lemee *et al.*, (2004)
Enterococcus faecalis	Firmicutes	Patógeno oportunista (3)	37	0.6	0.6	(0.0–3.2)	Ruiz-Garbajosa *et al.*, (2006)
Enterococcus faecium	Firmicutes	Patógeno oportunista (3)	15	7	1.1	(0.3–2.5)	Homan *et al.*, (2002)
Escherichia coli ET-1 group	g-proteobacteria	Patógeno oportunista (3)	44	0.7	0.7	(0.03–2.0)	Walk *et al.*, (2007)
Flavobacterium psychrophilum	Bacteroidetes	Patógeno obligado (2)	33	7	63.6	(32.8–82.8)	Nicolas *et al.*, (2008)
Haemophilus influenzae	g-proteobacteria	Patógeno oportunista (3)	50	7	3.7	(2.6–5.4)	Meats *et al.*, (2003)
Haemophilus parasuis	g-proteobacteria	Patógeno oportunista (3)	79	7	2.7	(2.1–3.6)	Olvera et al., (2006)
Halorubrum sp.	Halobacteria (Archaea)	Comensal - Vida libre (1)	28	4	2.1	(1.2–3.3)	Papke *et al.*, (2004)
Helicobacter pylori	e-proteobacteria	Patógeno oportunista (3)	117	8	13.6	(12.2–15.5)	pubmlst.,org
Klebsiella pneumoniae	g-proteobacteria	Patógeno oportunista (3)	45	0.3	0.3	(0.0–2.1)	Diancourt *et al.*, (2005)
Lactobacillus casei	Firmicutes	Comensal - Vida libre (1)	32	0.1	0.1	(0.0–0.5)	Diancourt *et al.*, (2007)
Legionella pneumophila	g-proteobacteria	Simbiontes / Pat. Intracel. (0)	30	2	0.9	(0.2–1.9)	Coscolla y Gonzalez-Candelas (2007)
Leptospira interrogans	Spirochaetes	Patógeno oportunista (3)	61	0.02	0.02	(0.0–0.1)	Thaipadungpanit *et al.*, (2007)
Listeria monocytogenes	Firmicutes	Patógeno oportunista (3)	34	0.7	0.7	(0.4–1.1)	Salcedo *et al.*, (2003)
Listeria monocytogenes	Firmicutes	Patógeno obligado (2)	92	7	4.42	----	den Backer *et al.*, (2010)
Mastigocladus laminosus	Cyanobacteria	Comensal - Vida libre (1)	34	4	0.9	(0.5–1.5)	Miller *et al.*, (2007)
Microcoleus chthonoplastes	Cyanobacteria	Comensal - Vida libre (1)	22	2	0.8	(0.2–1.9)	Lodders *et al.*, (2005)
Microcystis aeruginosa	Cyanobacteria	Comensal - Vida libre (1)	79	7	18.3	(13.7–21.2)	Tanabe *et al.*, (2007)
Moraxella catarrhalis	g-proteobacteria	Patógeno oportunista (3)	50	8	10.1	(4.5–18.6)	web..mpiib-berlin..mpg..de/mlst
Mycoplasma hyopneumoniae	Firmicutes	Patógeno oportunista (3)	33	7	3	(1.1–5.8)	Mayor *et al.*, (2007)
Myxococcus xanthus	d-proteobacteria	Comensal - Vida libre (1)	57	5	5.5	(1.9–11.3)	Vos y Velicer (2008)
Neisseria lactamica	b-proteobacteria	Comensal - Vida libre (1)	180	7	6.2	(4.9–7.4)	pubmlst..net
Neisseria meningitidis	b-proteobacteria	Patógeno oportunista (3)	93	20	14.4-108.9		Didelot *et al.*, (2009)
Neisseria meningitidis	b-proteobacteria	Patógeno oportunista (3)	83	7	7.1	(5.1–9.5)	Jolley *et al.*, (2005)
Neisseria meningitidis	b-proteobacteria	Patógeno oportunista (3)	30	7	100-271	----	Feil *et al.*, (2001)
Oenococcus oeni	Firmicutes	Comensal - Vida libre (1)	17	0.7	0.7	(0.2–1.7)	de Las Rivas *et al.*, (2004)
Pelagibacter ubique (SAR 11)	a-proteobacteria	Comensal - Vida libre (1)	9	8	63.1	(47.6–81.8)	Vergin *et al.*, (2007)
Plesiomonas shigelloides	g-proteobacteria	Comensal - Vida libre (1)	58	5	7.1	(3.8–13.0)	Salerno *et al.*, (2007)
Porphyromonas gingivalis	Bacteroidetes	Patógeno obligado (2)	99	0.4	0.4	(0.0–3.4)	Enersen et al., (2006)
Pseudomonas syringae	g-proteobacteria	Patógeno obligado (2)	95	4	1.5	(1.1–2.0)	Sarkar y Guttman (2004)
Pseudomonas viridiflava	g-proteobacteria	Patógeno obligado (2)	92	3	2	(1.2–2.9)	Goss *et al.*, (2005)
Ralstonia solanacearum	b-proteobacteria	Patógeno obligado (2)	58	7	1.1	(0.7–1.6)	Castillo y Greenberg (2007)
Rhizobium gallicum	a-proteobacteria	Simbiontes / Pat. Intracel. (0)	33	0.1	0.1	(0.0–0.38)	Silva *et al.*, (2005)
Salinispora arenicola	Actinobacteria	Comensal - Vida libre (1)	19	7	3.26	----	Kelle *et al.*, (2013)
Salinispora pacifica	Actinobacteria	Comensal - Vida libre (1)	17	7	0.02	----	Kelle *et al.*, (2013)
Salinispora tropica	Actinobacteria	Comensal - Vida libre (1)	12	7	2.56	----	Kelle *et al.*, (2013)
Salmonella enterica	g-proteobacteria	Patógeno oportunista (3)	50	7	30.2	(21.0–36.5)	web..mpiib-berlin..mpg..de/mlst
Staphylococcus aureus	Firmicutes	Patógeno oportunista (3)	53	0.1	0.1	(0.0–0.6)	Enright *et al.*, (2000)
Staphylococcus aureus	Firmicutes	Patógeno oportunista (3)	25	7	24	----	Basic-Hammer *et al.*,(2010)
Streptococcus dysgalactiae	Firmicutes	Patógeno oportunista (3)	80	7	20.7	----	Mc Millan *et al.*, (2010)
Streptococcus pneumoniae	Firmicutes	Patógeno oportunista (3)	52	6	23.1	(16.7–29.0)	Hanage *et al.*, (2005)
Streptococcus pneumoniae	Firmicutes	Patógeno oportunista (3)	40	7	61	----	Feil *et al.*, (2001)
Streptococcus pyogenes	Firmicutes	Patógeno oportunista (3)	50	7	17.2	(6.8–24.4)	Enright *et al.*, (2001)
Sulfolobus islandicus	Thermoprotei (Archaea)	Comensal - Vida libre (1)	17	5	1.2	(0.1–4.5)	Whitaker *et al.*, (2005)
Vibrio parahaemolyticus	g-proteobacteria	Patógeno oportunista (3)	20	7	39.8	(27.4–48.2)	Gonzalez-Escalona *et al.*, (2008)
Vibrio parahemoliticus	g-proteobacteria	Patógeno obligado (2)	174	11	16.84	----	Yan *et al.*, (2010)
Vibrio vulnificus	g-proteobacteria	Patógeno oportunista (3)	41	5	26.7	(19.4–33.3)	Bisharat *et al.*, (2007)
Wolbachia b complex	a-proteobacteria	Simbiontes / Pat. Intracel. (0)	16	5	3.5	(1.8–6.3)	Baldo *et al.*, (2006)
Xylella fastidiosa	g-proteobacteria	Patógeno obligado (2)	25	7	3.29	----	Scally *et al.*, (2005)
Yersinia pseudotuberculosis	g-proteobacteria	Patógeno obligado (2)	43	7	0.3	(0.0–1.1)	web..mpiib-berlin..mpg..de/mlst

obtención de respuestas concluyentes. Tal como sugieren Konstantinos T. Konstantinidis y Alejandro Caro-Quintero, se deben analizar muchas más poblaciones y *habitats* para poder establecer el papel del ambiente y la recombinación homóloga en los procesos evolutivos (Konstantinidis *et al*., 2006; Caro-Quintero *et al*., 2011; Caro-Quintero y Konstantinidis 2012).

Los resultados observados para especies como *S. baltica* y los presentados en el capítulo 2 en esta tesis para *S. ruber* nos llevan a plantearnos si el papel predominante de la recombinación homóloga sobre la evolución de los genomas core constituye la norma o la excepción. De hecho existen interrogantes ya no sólo acerca del papel homogenizador o no en diversas especies sino de la magnitud con que los procesos de recombinación afectan a algunos grupos filogenéticos o ecotipos (Vos y Didelot 2009, Caro-Quintero y Konstintinidis 2012).

Análisis genómicos comparativos y basados en MLSA sugieren que especies de vida libre y patógenos oportunistas con grandes poblaciones efectivas serían candidatos a sufrir más procesos de recombinación homólogas, sin embargo tal como se indica en los mismos, son escasos aún los estudios llevados a cabo con genomas completos y menos aun los que exploran el impacto de este fenómeno en los procesos microevolutivos (Vos y Didelot 2009, Mira *et al*., 2010). Qué factores y en qué medida estarían controlando los niveles de recombinación homóloga intraespecífica y si afectan del mismo modo a los diferentes grupos filogenéticos son cuestiones que permanecen sin abordar de manera concluyente.

Aunque se han llevado a cabo estudios extensivos de intercambio genético entre genomas bacterianos de diversas especies (Kloesges *et al*., 2010; Popa *et al*., 2011; Smillie *et al*., 2011), la metodología empleada en los mismos, análisis de redes y bloques de elevada identidad (ANI > 97%) entre especies (rRNA 16S < 97%), limitó los flujos detectados a intercambios interespecíficos y entre *phyla*. Ante la dificultad planteada a la hora de detectar intercambios intraespecíficos, estos análisis exploraron principalmente procesos de HGT interespecíficos, correspondientes a eventos mediados por recombinación no homóloga en su mayoría (Kloesges *et al*., 2010; Popa *et al*., 2011).

En este capítulo se abordan algunas de estas incógnitas mediante el análisis exhaustivo del efecto de la recombinación homóloga en 54 especies procariotas. La metodología empleada permitió comparar los niveles de recombinación detectados en los diferentes grupos y abordar

cuestiones abiertas en revisiones y trabajos anteriores como la influencia de factores filogenéticos, presión ambiental y estilo de vida de las diferentes especies analizadas influyen en los procesos de recombinación homóloga en los genomas *core*. Este trabajo representa el primer estudio extensivo de este tipo con genomas completos, que abarca varias cepas de especies distintas bajo una metodología común.

2. Construcción de la base de datos y análisis de eventos de recombinación en genomas completos.

2.1- Características de las especies analizadas y variables de estudio.

Este estudio se inició motivado por la búsqueda *in silico* de regiones con características similares a las de la zona conservada (CR) con el objetivo de determinar el significado y origen de la misma. Algunos análisis genómicos comparativos relacionan el impacto directo de la recombinación homóloga con la presencia de regiones amplias de dN/dS próximo a cero entre cepas cercanas de especies como *Francisella tularensis* (Larsson *et al*., 2009) *Staphylococcus aureu*s o *Clostrudium difficile* (Castillo-Ramírez *et al*., 2011). La aproximación más plausible fue la de explorar la presencia de regiones altamente conservadas y potencialmente recombinantes dentro de genomas bacterianos completos. Este tipo de análisis se ha llevado a cabo en especies concretas *E.coli* (Mau *et al*., 2006) y patógenas como *Listeria monocitogenes* (Renato *et a*l., 2008) y *Chlamydia trachomatis* (Sandeep *et al*., 2011), pero nunca de manera masiva ni empleando genomas secuenciados completamente procedentes de diversas especies. Con el objetivo de analizar el impacto y distribución de los procesos de recombinación homóloga a lo largo de los distintos filotipos procariotas se construyó una base de datos que comprendió 54 especies de bacterias y arqueas. Se consideraron aquellas especies con al menos 3 genomas secuenciados completamente, seleccionando cepas de diferentes ambientes, grupos filogenéticos y estilos de vida (**anexo, tabla S3.1**). La **figura C3.1** muestra la distribución filogenética de los 338 genomas incluidos en este estudio, 325 pertenecientes al dominio bacteria y 13 del dominio arquea y su distribución en 4 grandes grupos según su estilo de vida: Patógenos

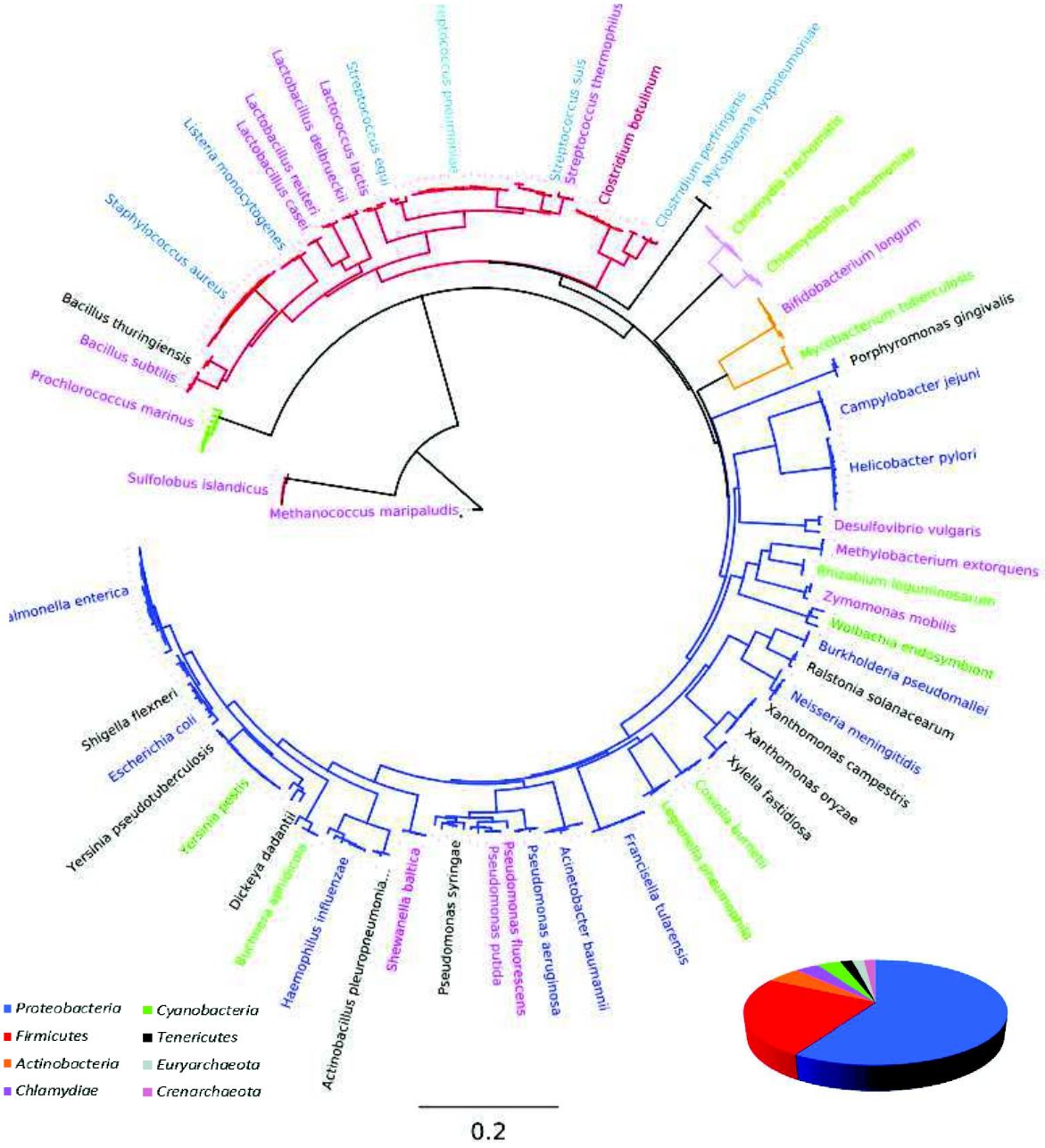


Figura C3.1. Árbol filogenético del gen del rRNA 16S realizado por *Neighbor-Joining* con un *Boostrapping* de 100 réplicas, que muestra la distribución filogenética de las 54 especies procariotas consideradas en este estudio. Las ramas principales representan los *phyla* en diferentes colores: (Azul= *Proteobacteria*; Naranja= Actinobacteria; Rosa=*Chlaydiae*; Negro=*Tenericutes*; Rojo=*Firmicutes*; Verde= *Cyanobacteria*). El nombre de las especies aparece coloreado en base a su estilo de vida: (0-Negro= Simbiontes/patógenos intracelulares, 1-Rosa= no patógenos (comensales y vida libre) , 2-Verde= patógenos obligados y 3-Azul= patógenos oportunistas).

Intracelulares y simbiontes, patógenos obligados, organismos de vida libre y patógenos oportunistas (véase introducción, apartado 2.3.7). Además se incluyeron 6 grupos externos que comprendieron cepas pertenecientes al mismo género pero no a la misma especie (ANI 16S <98.7%) (Achtman y Wagner 2008). Se comprobó la correspondencia de los genomas con su asignación filogenética en las bases de datos públicas (NCBI) comparando los valores de identidad ANI a nivel de la secuencia del gen del rRNA 16S 2 a 2. Empleando una matriz de identidades obtenida del alineamiento con Silva (Pruesse *et al*., 2007) (**anexo, tabla S3.3**) confirmamos que las cepas incluidas dentro de cada especie tenían un ANI 16S >98.7%. Entre las especies consideradas incluimos algunas patógenas y simbiontes para las cuales no se ha realizado hasta la fecha ningún análisis de recombinación. Dentro de los taxones estudiados, las *Gammaproteobacteria* (23 especies) y *Firmicutes* (14 especies) fueron los más representadas, abarcando especies con diferentes características genómicas y ecológicas (**figura C3.1**).

Se recuperaron los datos genómicos y ecológicos para cada una de las especies desde las bases de datos públicas o mediante análisis *in sílico* estableciendo las variables de estudio analizadas en este capítulo detalladas en material y métodos (véase introducción, **tabla 4M):** variables genómicas (tamaño del genoma e islas genómicas y contenido en transposasas) (**anexo tabla S3.1**), especialización ecológica (**anexo, tabla S3.5**), medidas de homología (proporción de genoma *core* recombinable y ANI entre cepas de una misma especie) (**anexo, tabla S3.4**), variables relativas a mecanismos que dificultan el intercambio de ADN, que denominamos "barrera" (abundancia e proteínas del sistema de MR y CRISPR-Cas), y que lo favorecen, que definimos como "movilidad" (% HGT, genes involucrados en mecanismos de reparación, conjugación y competencia) (**anexo, tabla S3.6**).

2.2- Estimación de la distribución de las regiones recombinantes en genomas completos.

La identificación de eventos de recombinación comprendió un alineamiento inicial de los cromosomas completos de cada especie para delimitar bloques colineares (**figura C3.2**; **anexo, figura S3.2**) y la posterior identificación y selección de posibles eventos de recombinación con el programa RDP4, tal como se hizo en el capítulo 2 con *S. ruber* (**anexo, figura S3.1**). El alineamiento de genomas también permitió calcular la longitud del genoma *core* susceptible de

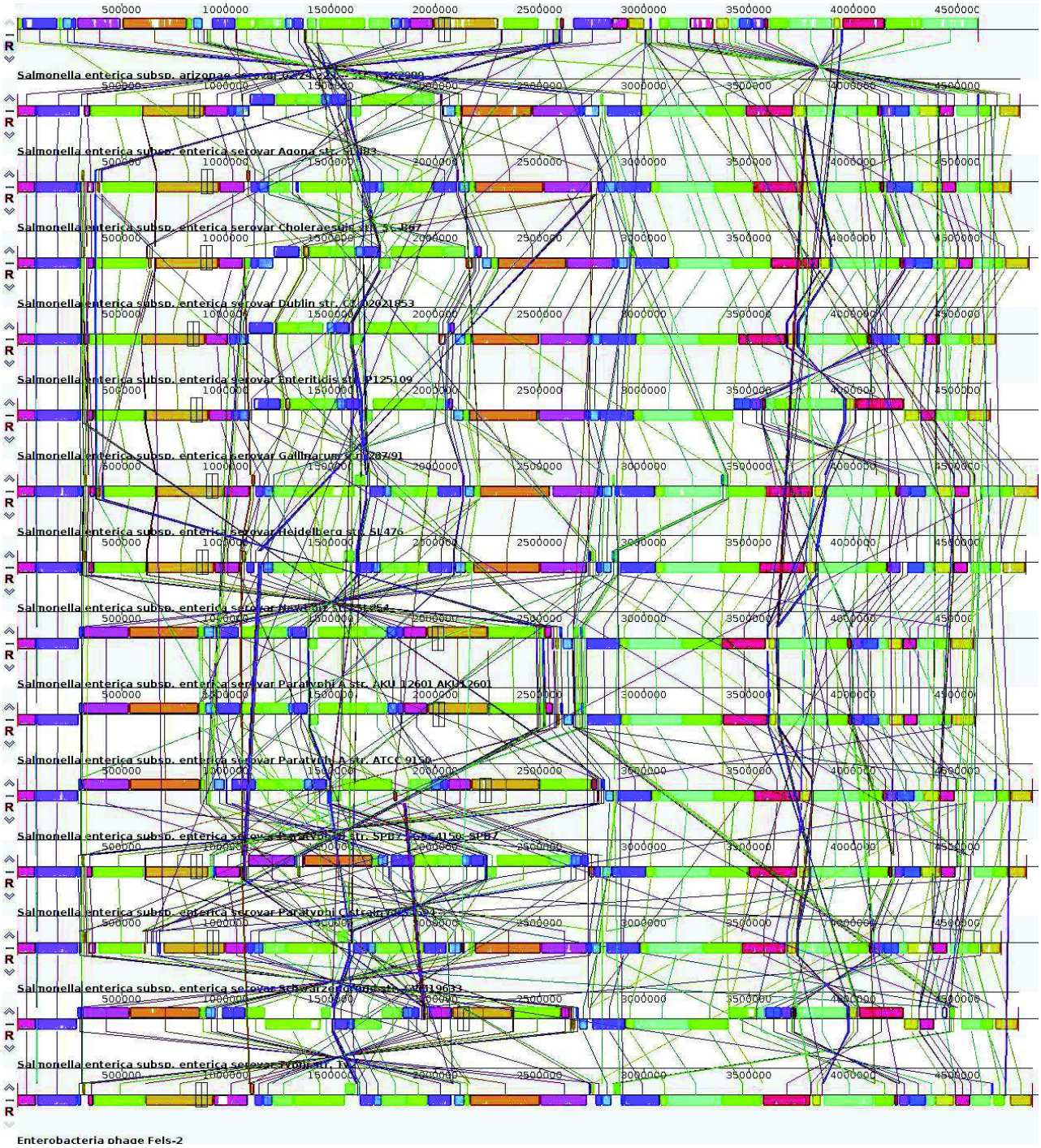


Figura C3.2. Alineamiento de las 15 cepas de *Salmonella enterica* con el programa ProgressiveMauve. La imagen muestra los bloques colineares identificados que representan regiones sinténicas compartidas por al menos 2 cepas (genoma accesorio) o por todas ellas (genoma *core*). Las líneas identifican la posición del bloque en las diferentes cepas. Aquellos boques situados bajo la línea horizontal corresponden a regiones genómicas invertidas.

sufrir eventos de recombinación. En total se identificaron y caracterizaron 16.300 eventos de recombinación distribuidos entre las 54 especies analizadas aunque no de manera homogénea (**anexo, tabla S3.7**), ya que en algunas especies como *E.coli* la recombinación afectó a más del 20% del genoma, un nivel similar a los observados en estudios previos con organismos de vida libre como *Francisella novicida* o *Francisella philomiragia* (Larsson *et al*., 2009) y *S. baltica* (Caro-Quintero *et al*., 2011) o patógenos oportunistas considerados tradicionalmente muy recombinantes como *Streptococcus pneumoniae* (Hiller *et al.*, 2010) o *Neisseria meningitidis* (Joseph *et al*., 2011). Por otra parte, en otras especies patógenas intracelulares como *Francisella tularensis* o *Mycobacterium tuberculosis* la recombinación afectó a menos del 4% del genoma, de acuerdo con las observaciones anteriores en organismos con este estilo de vida (Vos y Didelot 2009). Aunque en algunos casos observamos diferencias en las estimaciones de r/m y genoma afectado por recombinación, en la mayoría de casos los valores fueron parecidos a los estimados previamente mediante MLSA (**tabla C3.2**) o genomas completos (**tabla C3.1**). Estas diferencias atribuibles a la aproximación metodológica, aspecto comentado en la introducción, se han detectado anteriormente en especies como *Xanthomonas campestris*, en donde los niveles de recombinación detectados recientemente con genomas completos fueron muy superiores a los derivados de estudios MLSA (Huang *et al*., 2015), sucediendo algo similar en el caso de *S. islandicus* (Cadillo-Quiroz *et al*., 2010). Asimismo se han observado diferencias en las estimaciones de análisis de MLSA de estudios previos con diversas especies patógenas (Hanage *et al*., 2006, Pérez-Losada *et al.*, 2006; Vos y Didelot 2009), en el caso de *N. meningitidis* de hasta dos órdenes de magnitud en los valores de r/m.

Alrededor de un 60% de los fragmentos putativamente recombinados tuvieron longitudes menores de 3kb y contuvieron al menos 5 genes (**figura C3.3, figura C3.4, anexo tabla S3.7**), acorde con los resultados mostrados por diversos estudios de MLSA en *Neisseria meningitidis* (Jolley, 2004), *Campylobacter jejuni* (Fearnhead *et al.*, 2005; Wilson *et al*., 2009) y *Bacillus cereus* (Didelot e*t al.*, 2009) que muestran que la mayoría de eventos de recombinación afectaron regiones de longitud comprendida entre unos cientos pares de bases y 3-4 kb. Estos valores se sitúan entre los detectados en estudios anteriores con genomas completos de especies como *E.coli*, donde el 80% de los eventos de recombinación encontrados fueron menores de 2 kb

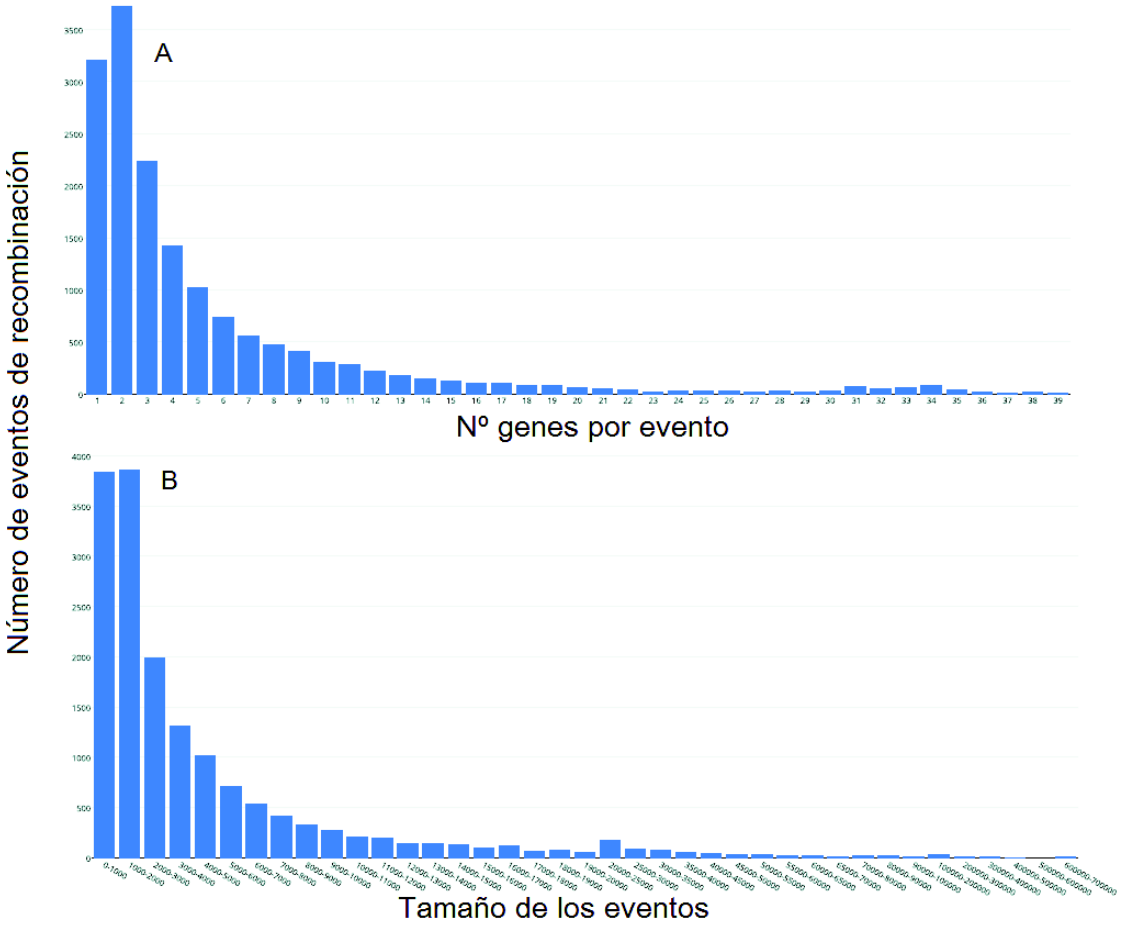


Figura C3.3. Distribución de eventos de recombinación en las 54 especies estudiados en función del número de genes contenidos (figura A) y de su tamaño (figura B). La mayoría de eventos presentaron tamaños menores de 3Kb y menos de 8 genes.

(Mau *et al*., 2006), en el patógeno intracelular *Burholderia pseudomallei*, con un promedio de 5kb y eventos en el rango de 3-71 kb (Nandi *et al*., 2015) y en el patógeno oportunista *Salmonella enterica* (Didelot *et al*., 2007). Un 11% (1.799) de los fragmentos detectados fueron mayores de 10 kb y 4,5% (75) mayores de 80 kb. En algunas especies la proporción de regiones de gran tamaño fue mayor, de hecho cada especie presentó una distribución distinta como se muestra en la **figura C3.4,** aunque con predominio de regiones menores de 2 kb. Se cree que la presencia de eventos de recombinación de gran tamaño, aunque no es tan frecuente en la naturaleza, está asociada a mecanismos de conjugación y selección positiva (Thomas y

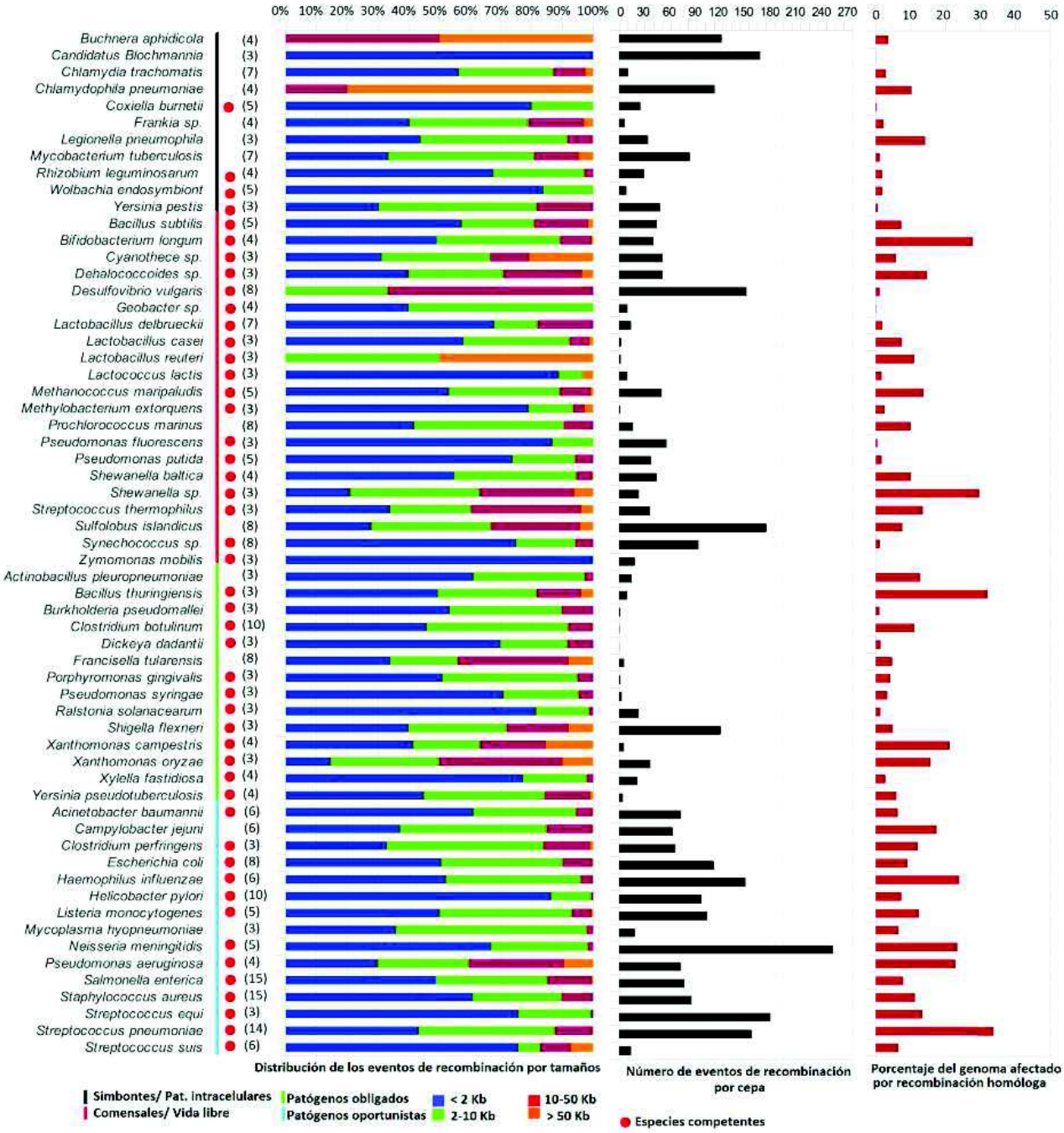


Figura C3.4. Distribución de la recombinación homóloga en las 54 especies analizadas y distribuidas en 4 estrategias ecológicas. El número de cepas incluidas en cada especie se muestra entre paréntesis y las especies competentes se identifican con un punto rojo. Para casa especie se muestra la distribución de eventos en función del tamaño y número y la proporción de genoma afectado por recombinación homóloga.

Nielsen 2005; Didelot y Maiden 2010; Castillo-Ramírez *et al*., 2011) como se discutirá más adelante (apartado 3.2). Identificamos regiones extensas similares a las detectadas en trabajos previos en extremófilos como *S. islandicus* (Cadillo-Quiroz *et al*., 2012) o *Haloferax sp*. (Naor *et al*., 2012), en patógenos como *S. pneumoniae* (Hiller *et al*., 2010), *S. aureus* (Castillo-Ramírez *et al*., 2011) y *L. pneumophila* (Gómez-Valero *et al*., 2011) y en organismos como *E. coli* (Mau *et al*., 2006. Además, el hecho de que las regiones recombinadas tengan un contenido en GC similar al del resto del genoma (**figura 3.4**) ($p>0,05$, ANOVA comparación inter-grupos), apoya que las transferencias detectadas sean intraespecíficas y entre cepas cercanas, ya que en el caso de tratarse de eventos de transferencia interespecíficos, las diferencias de GC entre donador y receptor podrían ser de hasta un 5% (Popa *et al*., 2011). Además estos resultados nos indican que la recombinación homóloga no actúa preferentemente sobre regiones de GC distinto al del promedio del genoma (**figura C3.5 B**). Estos resultados son concluyentes ya que además la distribución de genomas incluidos fue representativa al considerar organismos con diversas estrategias que presentaron una correlación de tamaños genómicos y contenido en GC ($r^2=0,688$) (**figura C3.5 A**) similar a la observada previamente con 364 genomas bacterianos (Wu *et al*., 2012).

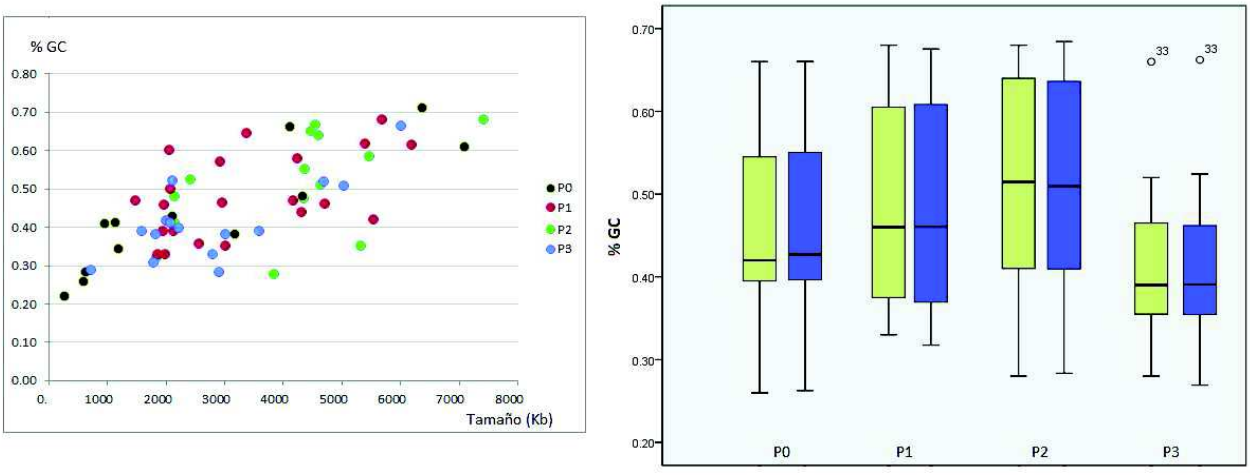


Figura C3.5. Contenido en GC promedio de los genomas de las 54 especies analizadas en este estudio distribuidas en 4 estrategias ecológicas: (P0 (negro) = Simbiontes/patógenos intracelulares, P1 (negro)= no patógenos (comensales y vida libre), P2 (azul) = patógenos obligados y P3 (verde)= patógenos oportunistas).
Figura A: Correlación entre el tamaño promedio del genoma de cada especie y su contenido en GC.
Figura B: *Box plot* comparando los promedios de contenido en GC de los genomas de las especies distribuidos en las 4 clases ecológicas (gris) y de sus regiones recombinadas (azul).

3. Factores que afectan a la incidencia de la recombinación homóloga.

3.1- El efecto de la filogenia y la especialización ecológica.

En vista de los patrones heterogéneos de recombinación homóloga en términos de porcentaje de genoma recombinado, eventos por cepa y relación r/m, observados en las especies analizadas, decidimos estudiar si esta seguía algún patrón general de distribución. Con el objetivo de explorar la influencia de la distribución filogenética en las variables de recombinación mencionadas, incluimos en el estudio los denominados grupos control: genomas con relación filogenética a nivel de género (ANI 16S <98.7%) (Achtman y Wagner 2008) (anexo, tabla S3.2) y un ANI global inferior al 95%, lo que corresponde con una hibridación DNA-DNA entre cepas del 70%, por encima del cual dos cepas se consideran de la misma especie (Konstantinidis y Tidje 2005; Konstantinidis *et al*., 2006; Goris *et al*., 2007; Caro-Quintero y Konstantinidis 2011). Entre las 54 especies estudiadas, incluimos algunas de un mismo género con distinto estilo de vida. Trabajos anteriores indican que la incidencia de la recombinación homóloga decae con la divergencia de secuencia (Majewski 2001, Thomas y Nielsen., 2005, Fraser *et al*., 2007; Polz y Hanage 2013) y decrece enormemente entre fragmentos de DNA que divergen más de un 25-30% (Thomas y Nielsen., 2005), siendo más común entre cepas de una misma especie e incluso dentro de una misma especie, entre linajes cercanos geográficamente. Este último es el caso de *clusters* y líneas clonales en especies patógenas como *Listeria monocitógenes* (den Bakker *et al*., 2008), *N. meningitidis* (Orsi *et al*., 2008) o *S. enterica* (Didelot *et al*., 2011) o linajes patogénicos de *E.coli* (Liu *et al*., 2006), o entre *clusters* poblacionales de *S. islandicus* (Cadillo-Quiroz *et al.*, 2012) en proceso de diferenciación poblacional. Aunque menos frecuentes, existen trabajos previos en los que se han detectado niveles de recombinación homóloga relevantes entre especies cercanas de *Haloarchaea* (Naor *et al*., 2012) o entre especies del género *Streptococcus* como *S. zooepidermicus* y *S. equi* y *S. pyogenes* (Holden *et al*., 2009), entre las especies *S. pneumoniae* , *S. pseudopneumoniae, S. mitis* (Do *et al.*, 2009; Donati *et al*., 2010). Destaca el caso de las especies patógenas gastrointestinales *Campylobacter jejuni* y *Campylobacter coli* en proceso de reunificación tras el incremento en las tasas de recombinación entre ambas (Sheppard *et al*.,

2008; Caro-Quintero *et al*., 2009; Sheppard y Maiden 2015). En nuestro caso, los análisis de recombinación para 4 de los 5 grupos control *Cyanothece* sp. *Candidatus sulcia*, *Synechococcus* sp. *Frankia* sp. y *Candidatus blochmania* indican que recombinaron menos del 2% de su genoma, acorde con lo expuesto y confirmando que los datos observados en el resto de especies son atribuibles a recombinación homóloga intraespecífica, siendo los intercambios homólogos interespecíficos mucho más limitados.

Distribución filogenética del impacto de la recombinación homóloga.

El efecto de la filogenia y su relación con los de mecanismo de recombinación ha sido ampliamente analizada y revisada en estudios previos (Vos y Didelot 2009; Didelot y Maiden 2010) en los que se han recopilado valores de relación r/m procedentes de estudios fundamentados en MLSA para varias especies de un mismo género, caso de *Vibrio* (González-Escalona *et al*., 2008, Bisharat *et al*., 2007), *Streptococcus* (Hanage *et al*., 2005, Enright *et al*., 2001), *Neisseria* (Jolley et al., 2005, pubmlsd.net), *Psedomonas* (Goss *et al*., 2005, Sarkar y Gutman 2004) y Bacillus (Sorokin *et al*., 2006) (**tabla C3.2**). Estas revisiones indican que las tasas de recombinación podrían ser similares entre especies filogenéticamente cercanas, argumentando la posibilidad de que exista una evolución o constricción de las tasas r/m a nivel de género. Sin embargo estos estudios son limitados debido a que la mayoría de géneros analizados contuvieron sólo dos especies que compartían estrategia de vida, lo que impidió comprobar si la asociación se debe más a este último factor. Para comprobar esto, sería necesario ampliar las comparaciones a un número de especies mayor y diversas. En nuestro estudio se incluyeron 7 géneros con más de una especie analizada, y en ocasiones con distintas estrategias ecológicas: *Bacillus* (2), *Lactobacillus* (3), *Streptococcus* (4), *Xanthomonas* (3), *Pseudomonas* (4), *Yersinia* (2) (**SMF2.2**). Observamos niveles de recombinación diferentes entre muchas de ellas como muestran las dos especies consideradas del género *Yersinia*, el **patógeno monomórfico** intracelular *Y. pestis* y el oportunista *Y. pseudotuberculosis*, la segunda con una tasa r/m muy superior y 10 veces más eventos de recombinación por cepa. Es posible apreciar una situación similar entre las especies de los géneros *Streptococcus*, donde *S. pneumoniae*

presentó valores muy superiores en tasas de recombinación y porcentaje de genoma recombinado que *S. suis*, *S. thermophilus* y *S. Equi*. Además detectamos que los niveles de recombinación fueron inferiores en las especies de vida libre del *género Pseudomonas, P. fluorescens y P. putida* respecto a las patógenas *P. aeruginosa* y *P. syringae*. Acorde con nuestro resultados, estudios anteriores mostraron diferencias importantes en las tasas de recombinación entre poblaciones de especies muy cercanas del *género Francisella*, *F. tularensis* y *F. novicida* (Larsson *et al*., 2009), la primera de ellas patógena intracelular con nicho restringido, e incluso entre linajes de una misma especie como es el caso de *L. monocitogenes* (Den Bakker *et al*., 2010), *Bacillus cereus* (Didelot *et al*., 2009). Estos estudios junto con los resultados obtenidos en nuestro

Niveles de recombinación homóloga asociados a la especialización ecológica.

En conjunto los datos observados indican que especies de bacterias cercanas filogenéticamente con diferentes estilos de vida podrían presentar niveles de recombinación muy distintos. El efecto del estilo de vida sobre la transferencia horizontal interespecífica ha sido caracterizado ampliamente con el uso de genomas completos determinando que los niveles de flujo entre organismos endosimbiontes y patógenos intracelulares son mucho menores que los detectados en organismos de vida libre o patógenos oportunistas. Estos mismos estudios detectaron mayor recombinación interespecífica entre especies competentes del *phylum Proteobacteria* (Kloesges *et al*., 2010; Popa *et al*., 2011). La falta de resolución en ocasiones y la heterogenidad en las estimaciones llevadas a cabo por MLSA, en especial en especies comensales y patógenas obligadas y oportunistas debido a las fluctuaciones asociadas a procesos de virulencia, han dificultado el establecimiento de un patrón general. En un intento de explorar el efecto de las diferentes estrategias de vida sobre el impacto de la recombinación homóloga, distribuimos las 54 especies analizadas en 4 grandes grupos. La comparación de medias para las variables de recombinación porcentaje de genoma recombinado y eventos por cepa mostró diferencias significativas entre las clases analizadas ($p<0,05$, test de Kruskal-Wallis y Jonkheere-Tepstra) (**figura C3.6**), confirmando la influencia del estilo de vida como factor determinante que explica el 27% de la variabilidad de los datos ($p<0,05$; modelo lineal).

Los resultados indican que los endosimbiontes y patógenos intracelulares son los menos recombinantes, seguidos de organismos de vida libre y comensales, patógenos obligados y patógenos oportunistas. El orden establecido con el mostrado en estudios recientes (García-González *et al*., 2013), en los que se comparó el contenido en genes de reparación y recombinación entre estas cuatro estrategias de vida, aspecto se discutirá más adelante (apartado 3.2). Esta relación es una muestra de que el grado de especialización y adaptación a menudo viene acompañado por un cambio en las características genómicas generales, como puede ser el contenido en genes implicados en procesos de competencia o reparación de DNA y recombinación o la reducción de y genoma accesorio GIs (**figura C3.7**). Un ejemplo evidente

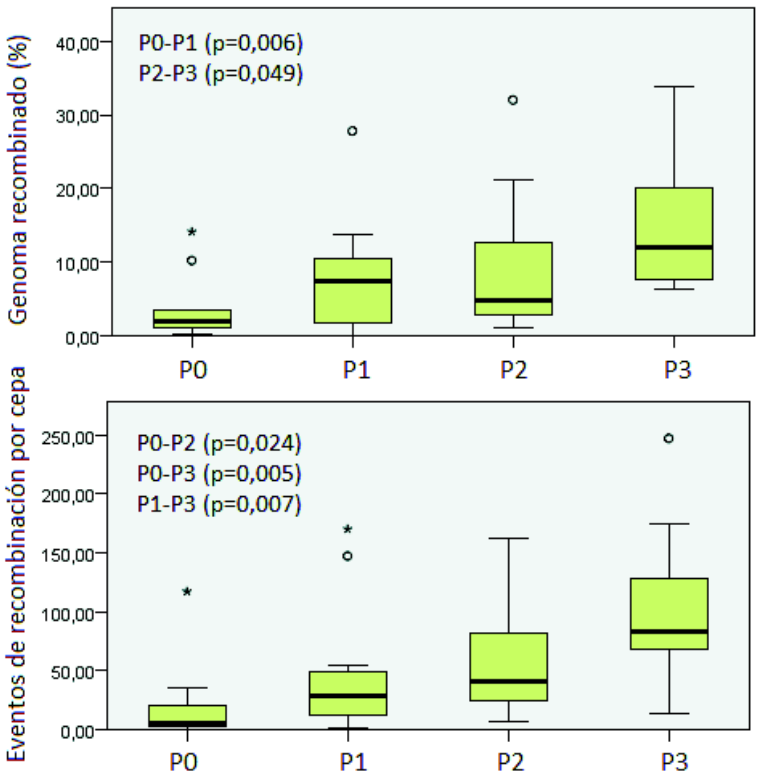


Figura C3.6. Diferencias en los niveles de recombinación homóloga por clases de especialización ecológica. Los diagramas de caja muestran las diferencias en los promedios de eventos de recombinación y porcentaje del genoma recombinado. En cada caso se muestran los p-valores significativos entre diferentes clases: Simbiontes/patógenos intracelulares (P0), no patógenos (comensales y vida libre) (P1), patógenos obligados (P2) y patógenos oportunistas (P3) ($p<0,05$, test de Kruskal-Wallis y Jonkheere-Tepstra).

son los casos de especialización máxima como la reducción del genoma en el caso de los organismos simbiontes o patógenos intracelulares (Ochman *et al*., 2001; Yu *et al*., 2009) o la incorporación de GIs en organismos de vida libre patógenos oportunistas y obligados, aspectos observadas en al comparar los genomas de las especies consideradas en nuestro estudio (**figura 3.7**).

Los organismos endosimbiontes y patógenos intracelulares, ente los que se incluyeron algunos de los más relevantes como *Buchnera aphidicola, M. tuberculosis, Y. pestis, C. trachomatis* y *C. pneumoniae*, confirmaron su naturaleza típicamente clonal (Achtman 2008), ya que en ninguna de ellas se detectaron más de 5 eventos por cepa y

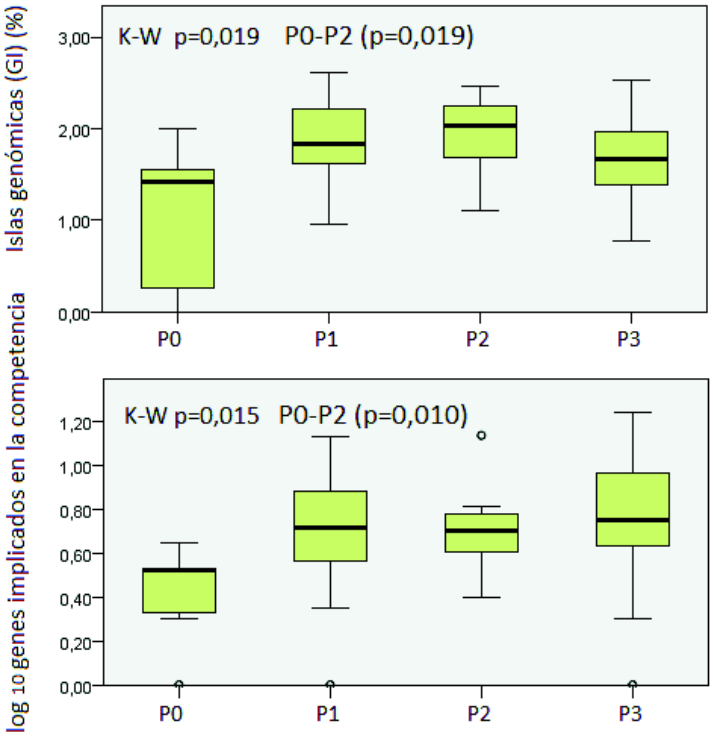


Figura C3.7. Diferencias en los contenidos en la capacidad de competencia y tamaño de GIs según la estrategia ecológica. homóloga por clases de especialización ecológica. Diagramas de caja muestran las diferencias en los promedios de eventos de recombinación y porcentaje del genoma recombinado. En cada caso se muestran los p-valores significativos entre diferentes clases: Simbiontes/patógenos intracelulares (P0), no patógenos (comensales y vida libre) (P1), patógenos obligados (P2) y patógenos oportunistas (P3) ($p<0,05$, test de Kruskal-Wallis y Jonkheere-Tepstra).

valores r/m mayores de 0,7 (**figura C3.4; anexo, tabla S3.8**). Estos dos grupos de organismos parecen tener una estructura poblacional clonal, caracterizada por los bajos niveles de diversidad, que dificulta mucho tanto los análisis filogeográfico y la detección de eventos de recombinación homóloga (Achtman 2008) De hecho hasta el momento sólo han estimado las tasas de r/m para 4 de las 9 especies incluidas en la categoría de patógenos intracelulares y simbiontes incluidas en este capítulo tal como se muestra en revisiones recientes (Vos y Didelot 2009; Didelot y Maiden 2010). Además estas especies presentan niveles de intercambio interespecífico muy reducidos debido a la restricción de nicho y aislamiento (Popa *et al*., 2011). Entre los patógenos monomórficos incluídos en nuestro estudio encontramos ejemplos de especies con elevada especialización ecológica y niveles de recombinación inferiores a los detectados en especies del mismo género menos especializadas como en el caso de *Y. pestis (Achtman et al., 2004*, esta última clon monomórfico de *Y. pseudotuberculosis* tras divergir y adaptarse a un nuevo nicho coincidiendo con la incorporación de de plásmidos, (Achtman *et al*., 1999). Los tres organismos simbiontes incluidos en esta tesis, *B. aphidicola, Rhizobium leguminosarum* y *Wolbachia endosymbionti*, simbiontes de áfidos, leguminosas y nematodos, respectivamente presentaron niveles de recombinación muy bajos, acorde con los valores obtenidos para el resto de organismos especializados de este estudio. Nuestras estimaciones con las especies *W. endosymbionti y R. leguminosarum* fueron más conservadoras que las obtenidas por MLSA en estudios previos (Baldo *et al*., 2006; Kumar *et al.*, 2015), considerando estos estudios niveles de recombinación moderados en ambas que contribuirían a la homogenización y cohesión de los genomas *core* pese a que en algunas **genoespecies** se detectaron valores de r/m superiores a 100 (Kumar *et al.*, 2015).

En el caso de los organismos de vida libre y comensales encontramos niveles de recombinación variables (**figuras C3.4 y C3.6; anexo, tabla S3.8**). Entre los *Firmicutes* del género *Lactobacillus* incluidos en este estudio, *L. casei, L. delbruecki, L. reuteri* y *L. lactis* estimamos un bajo impacto de la recombinación homóloga, menos de 50 eventos de recombinación por cepa. Nuestros datos confirman los niveles de clonalidad en este género tal como indican estudios con MLSA en *L. casei* (Diancourt *et al*., 2007) y *L. fermentum* (Dan *et al*., 2015), esta última considerada la especie más cercana filogenéticamente a *L. reuteri*. Los

organismos acuáticos de vida libre presentaron niveles de recombinación mayores a los organismos mencionados anteriormente, aunque intermedios teniendo en cuenta el conjunto de especies analizadas en este estudio y revisadas anteriormente (Didelot *et al.*, 2009). Tanto el elevado número de eventos de recombinación como los valores de r/m en todos los casos se situaron en el rango entre 0,25 y 4, considerado como el indicado para la formación de poblaciones estables en las que se daría la mezcla y cohesión de linajes y de genomas *core* por recombinación homóloga, y dentro del cual se encuentran *A. macleodii* (López-Pérez *et al*., 2013), *Halorrubrum* sp. (Papke *et al*., 2004, 2007) *S. ruber* (analizada en el capítulo 2). Además de en estos microorganismos acuáticos, en otros como *S. islandicus* , *S. baltica* o *P. marinus*, con cerca 10% del genoma afectado por recombinación las dos primeras, la recombinación homóloga es el mecanismo principal de evolución de sus genomas *core* (Whittaker *et al.*, 2005, Caro-Quintero *et al*., 2011). Nuestras estimaciones fueron más conservadoras que las aportadas por estudios previos en los que se compararon genomas completos (Caro-Quintero *et al*., 2011; Cadillo-Quiroz *et al*., 2012). La presencia de recombinación homóloga entre organismos acuáticos de vida libre fue mayor que en los terrestres confirmando las hipótesis planteadas en estudios anteriores que atribuyen tales diferencias a factores ambientales como la mayor presencia de virus en ambientes acuáticos, que seleccionarían aquellas cepas con elevadas tasas de recombinación fruto de los elevados niveles de coevolución entre virus-hospedador (Weinbauer 2004; Vos y Didelot 2008; Vos 2009).

Nuestros resultados indican que entre los organismos patógenos oportunistas y obligados existe una amplia variabilidad de frecuencias de recombinación (**figuras C3.4 y C3.6**), probablemente afectadas por factores como la restricción del rango de hospedador y el grado de especialización, mayor en patógenos obligados, y cambios ambientales derivados de la interacción del sistema inmune del hospedador (Vos y Didelot 2008; Michod *et al*., 2008). En el caso de bacterias patógenas, la constricción del ambiente del hospedador es mayor que la de ambientes acuáticos y de vida libre, por lo que la interacción y densidad bacterianas dentro de ambientes como en intestinal, el respiratorio o en *biofilms* es mucho mayor (Dobrind *et al*., 2004). La recombinación homóloga supone una estrategia de adaptabilidad patogénica que permite incrementar la dinámica, plasticidad genómica y diversidad poblacional durante

procesos de infección policlonal (Roberts *et al*., 2014; Hiller *et al*., 2010, Seitz y Blokesch 2012). La presencia de GIs de patogenicidad contribuye a la dinámica e incorporación de factores de virulencia y otros genes determinantes que favorecen el acceso al *pool* génico y son más frecuentes en especies patógenas que en otras cercanas filogenéticamente con estrategias de vida distintas (Hacker y Carniel 2001; Dobrind *et al*., 2004). La presión selectiva ambiental resulta un factor determinante, ya que a menudo se ha asociado la presencia de sistemas de recombinación diversos y completos en bacterias cuyo material genético se encuentra sometido a condiciones ambientales estresantes que afecten a la estabilidad del DNA (Seitz y Blokesch 2012; Van Wolferen *et al*., 2013). *Helicobacter pylori* y *S. pneumoniae* se consideran dos claros ejemplos al estar sometido al estrés oxidativo extremo por acción del sistema inmune compensado con sistemas completos de recombinación (Michod *et al*., 2008).

A menudo los patógenos se enfrentan a incrementos en el estrés oxidativo debido a la respuesta inmune del hospedador, durante la cual se ha descrito un incremento en las tasas de recombinación y mutagénesis adaptativa, secreción activa de DNA al medio, inducción de competencia y sistemas de reparación, como en el caso de *S. pneumoniae* (Prudhome *et al*., 2006; Claverys *et al*., 2006) y *H. pylori* (Faluch *et al*., 2001; Dorer *et al*., 2010; Seitz y Blokesch 2012; Didelot *et al*., 2013), consideradas especies altamente recombinantes, y formación de *biofilms*, estos últimos considerados un mecanismo barrera de defensa. A menudo las elevadas tasas de recombinación se asocian con la capacidad virulenta de las cepas ya que es el principal mecanismo que media la variación antigénica y que, tras la incorporación en una de las cepas y su estabilización, permitiría su acceso al resto de la población (Bruen *et al*., 2006; Vink *et al*., 2012).

Entre los patógenos obligados incluidos en nuestro estudio encontramos patógenos de plantas que muestran una estructura poblacional claramente clonal como es el caso de *P. syringae* (Yan *et al*., 2008), *Xylella fasidiosa* (Scally *et al*., 2005; Nunney *et al*., 2014), *Ralstonia solanacearum* (Castillo *et al*., 2007) y del patógeno periodontal *Porphyromonas gingivalis* en las que detectamos proporciones de genoma recombinado inferiores al 4% y menos de 40 eventos de recombinación por cepa, resultados que confirman su carácter clonal. Los niveles sobreestimados para las tasas r/m y rho/theta en estos casos puntuales se deben al reducido número de SNPs

entre cepas. Por tanto no implican que el impacto por recombinación homóloga sea elevado, sino simplemente que su aporte es muy superior a los SNPs generados por mutación. En el caso del patógeno *Xhanthomonas oryzae*, los niveles de recombinación detectados,15% del genoma *core* recombinado, fueron superiores a los detectados recientemente en un estudio llevado a cabo con genomas completos en el que se detectaron niveles de recombinación en el 10% del genoma *core* (Huang *et al*., 2015).

Entre los patógenos oportunistas encontramos los niveles de recombinación mayores (**figura C3.4 y C3.6**), tanto en número de eventos como en porcentaje de genoma recombinado, muchos de ellos especies que habitan el tracto gastrointestinal, el tracto respiratorio o la piel. Las especies que presentaron más eventos de recombinación fueron *N. meningitidis, H. pylori, S. pneumoniae y Streptococus equi y C. jejuni* con más de 100 eventos de recombinación por cepa y valores de r/m mayores de 100 en ocasiones (**anexo, tabla S3.8)**. Estos datos reflejan la flexibilidad y adaptabilidad de los genomas de estas especias, donde la recombinación homóloga se postula como el mecanismo predominante en la evolución de sus genomas *core.* Aunque ya se llegó a esta conclusión en publicaciones previas con *N. meningitidis, H. pylor y, S. pneumoniae* (Feil *et al*., 2000), los niveles de recombinación detectados en nuestro estudio son mayores que los estimados entonces, alcanzando el nivel necesario para mantener la cohesión poblacional y evitar la divergencia de líneas clonales (Fraser *et al*., 2007). Destacan los niveles encontrados en las dos especies del género *Streptococcus* y en *H. pylori*, mayores que en estudios precedentes (**anexo, tabla S3.8**). En ambas especies se incluyeron aislados simultáneos de pacientes de la misma procedencia, en el caso de *S. pneumoniae* varios aislados clínicos del hospital de Pittsburg (USA) y Reino Unido (Hiller *et al.*, 2007) y para *H. pylori* procedentes de la población de Nariño (Colombia) (Kenneman *et al*., 2011). Los resultados obtenidos en nuestro estudio apoyarían la hipótesis planteada en revisiones anteriores según la cual los procesos de recombinación acompañan procesos de adaptación e incremento de virulencia en episodios epidémicos (Didelot y Maiden 2010; Vink *et al*., 2012) son mayores en regiones endémicas como en el caso del patógeno beta-hemolítico nasofaríngeo *Streptococcus dysgalactiae* (McMillan *et al*., 2010).

Relación entre los patrones funcionales de recombinación y adaptación y la especialización ecológica.

Aunque en muchos análisis de MLSA se han observado diferentes tasas de recombinación para distintos genes *housekeeping*, sólo con el aumento de estudios genómicos comparativos se ha podido confirmar que de los genes involucrados en eventos de recombinación homóloga, una fracción considerable se encuentra bajo presión selectiva positiva, tal como se ha observado en especies como *L. monocitogene*s (Orsi *et a*l., 2008), *C. trachomati*s (Joseph et al., 2011) y el género *Streptococcus* (Lefebure *et al*., 2005), lo cual implica una contribución en los procesos de adaptación al ambiente. Sin embargo hasta la fecha no se ha observado de manera sistemática un grupo de genes o categorías que ofrezcan patrones de tasas de recombinación homóloga elevadas o bajas en un rango amplio de especies, ni una distribución comunes (revisado en Didelot *et al*., 2010). Por esta razón, y para una mejor comprensión de los patrones de recombinación homóloga entro de los genomas *core*, estudiamos el contenido génico de las regiones recombinantes a lo largo de las 54 especies. La **figura C3.8** muestra el número de especies que experimentaron enriquecimientos y empobrecimientos significativos en genes recombinados según las categorías funcionales COG ($p<0,05$; corrección FDR <10%, test de Fisher). Las categorías enriquecidas en un mayor número de especies fueron aquellas implicadas en mecanismos de defensa (COG V), motilidad celular y secreción (COG N) y procesamiento y modificación de RNA (COG A), mientras que la mayoría de categorías relacionadas con funciones metabólicas estuvieron empobrecidas en al menos 10 especies, entre ellas metabolismo y transporte de aminoácidos (COG E), conversión y producción de energía (COG C) y metabolismo y transporte de carbohidratos (COG G) (**figura C3.8**). Precisamente estas tres categorías COG son las más frecuentes entre los intercambios interespecíficos mediados por recombinación ilegítima, donde predominó la presencia de genes relacionados con metabolismo, seguido se procesos celulares, que en el caso de la recombinación homóloga contuvieron una proporción importante de enriquecimientos, y por último aquellas categorías relacionadas con el procesamiento y almacenamiento de información (Popa *et al*., 2011).

Este patrón se confirmó por medio de una matriz de similitud generada con los datos de enriquecimientos y empobrecimientos para las diferentes categorías COG (**figura C3.9**) que sustenta un agrupamiento claro de las categorías relacionadas con el metabolismo celular y entre las más enriquecidas aquellas relacionadas con procesos celulares. Además se observa un agrupamiento de patógenos obligados y oportunistas, estilos de vida en los que la adquisición de elementos móviles en islas de patogenicidad juega un papel importantísimo en los procesos de adaptación rápida al hospedador (Haecker y Carniel 2001). También se agruparon especies del mismo género cuando estas comparten una misma estrategia ecológica, como es el caso de las

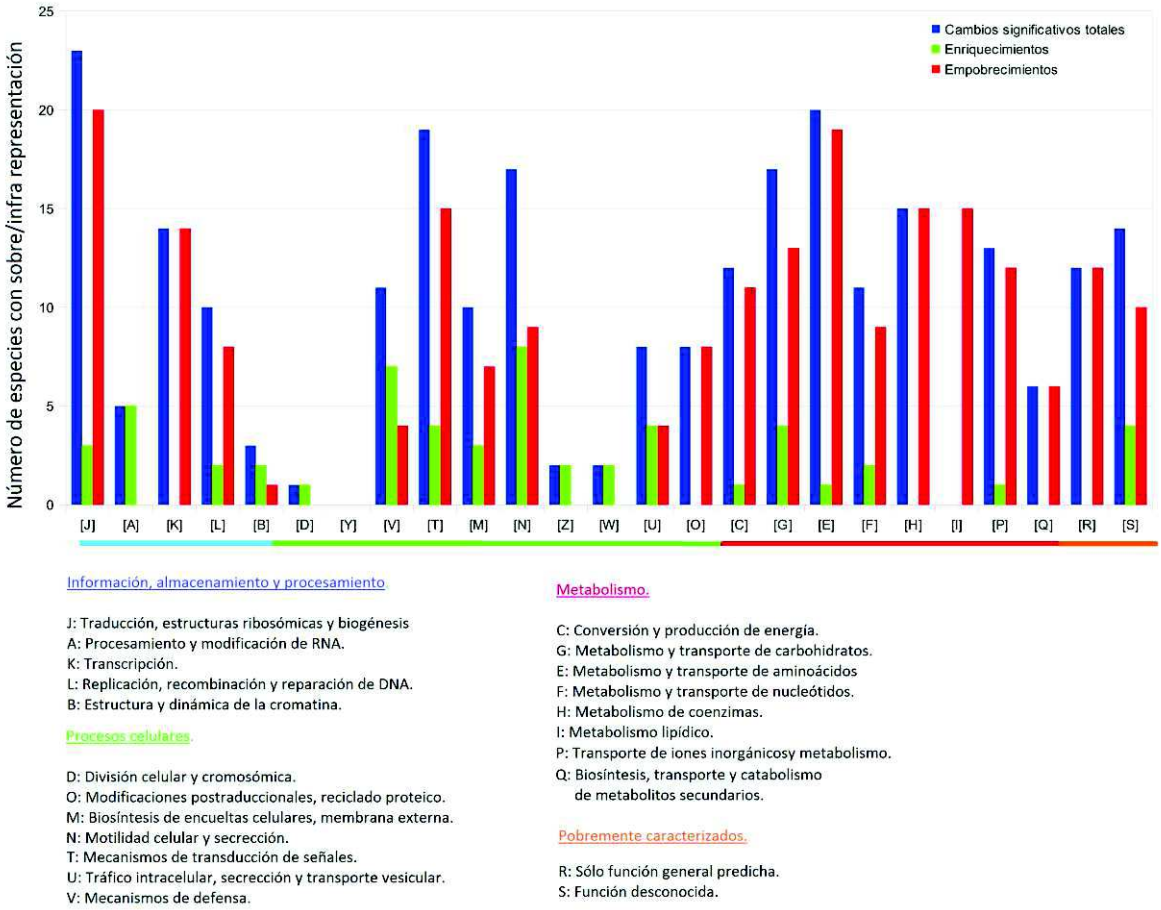


Figura C3.8. Distribución del número de especies que presentaron enriquecimientos (verde) o empobrecimientos (rojo) en sus eventos de recombinación para las distintas categorías COG. En azul se muestra la suma de los enriquecimientos y empobrecimientos detectados.

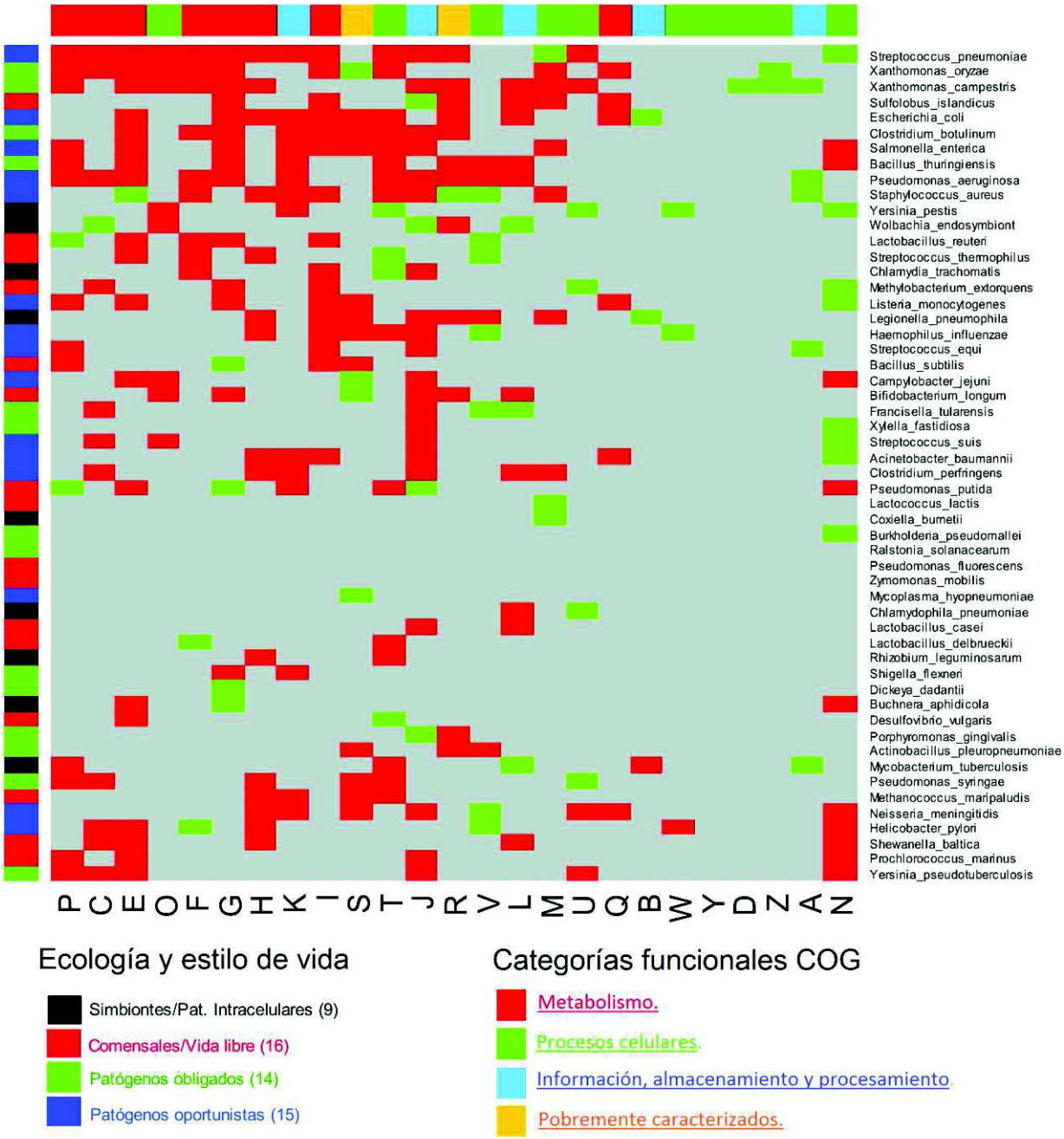
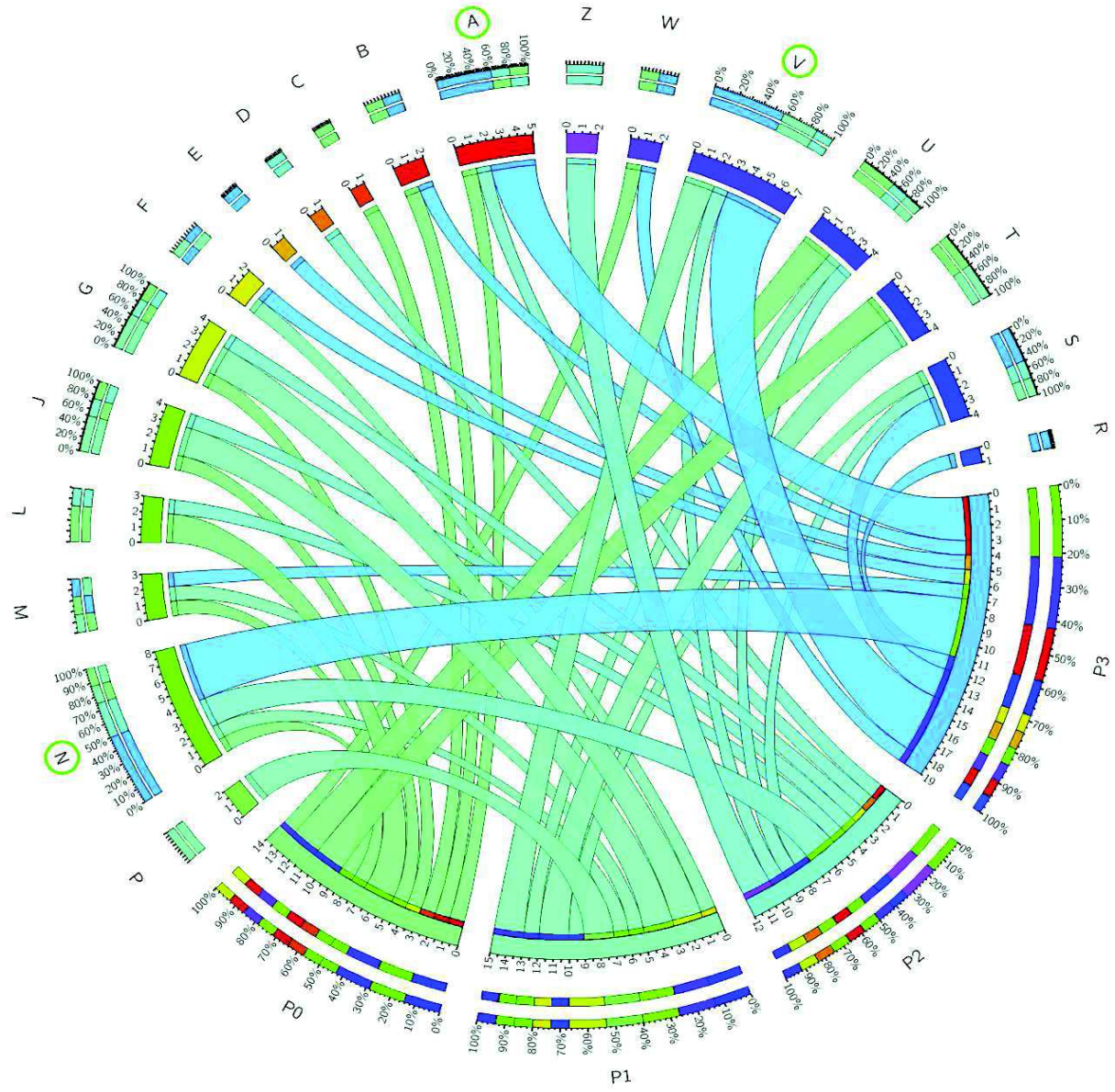


Figura C3.9. *Heatmap* para la matriz de similitud que muestra la agrupación de las 54 especies analizadas en base a su perfil de categorías COG enriquecidas (verde)/empobrecidas (rojo) significativamente ($p < 0,05$, Test de Fisher; Corrección FDR < 0,1). En el eje de las X se muestra la disposición de las categorías funcionales y en el de la Y la de las especies analizadas según patrones de similitud.

dos especies del género *Xanthomonas* y dos de las del género *Lactobacillus*, *L. casei* y *L. delbruecki*. Especies emparentadas filogenéticamente y con un mismo estilo de vida presentaron patrones similares, como es el caso de *E. coli* y *S. enterica*. Entre las bacterias de vida libre, *S. baltica* y *P. marinus*, ambas de ambiente marino y con valores de r/m y genoma recombinado casi idénticos, mostraron perfiles similares. En conjunto, los datos sugieren un patrón de flujo y recombinación homóloga dependiente de la estrategia ecológica.

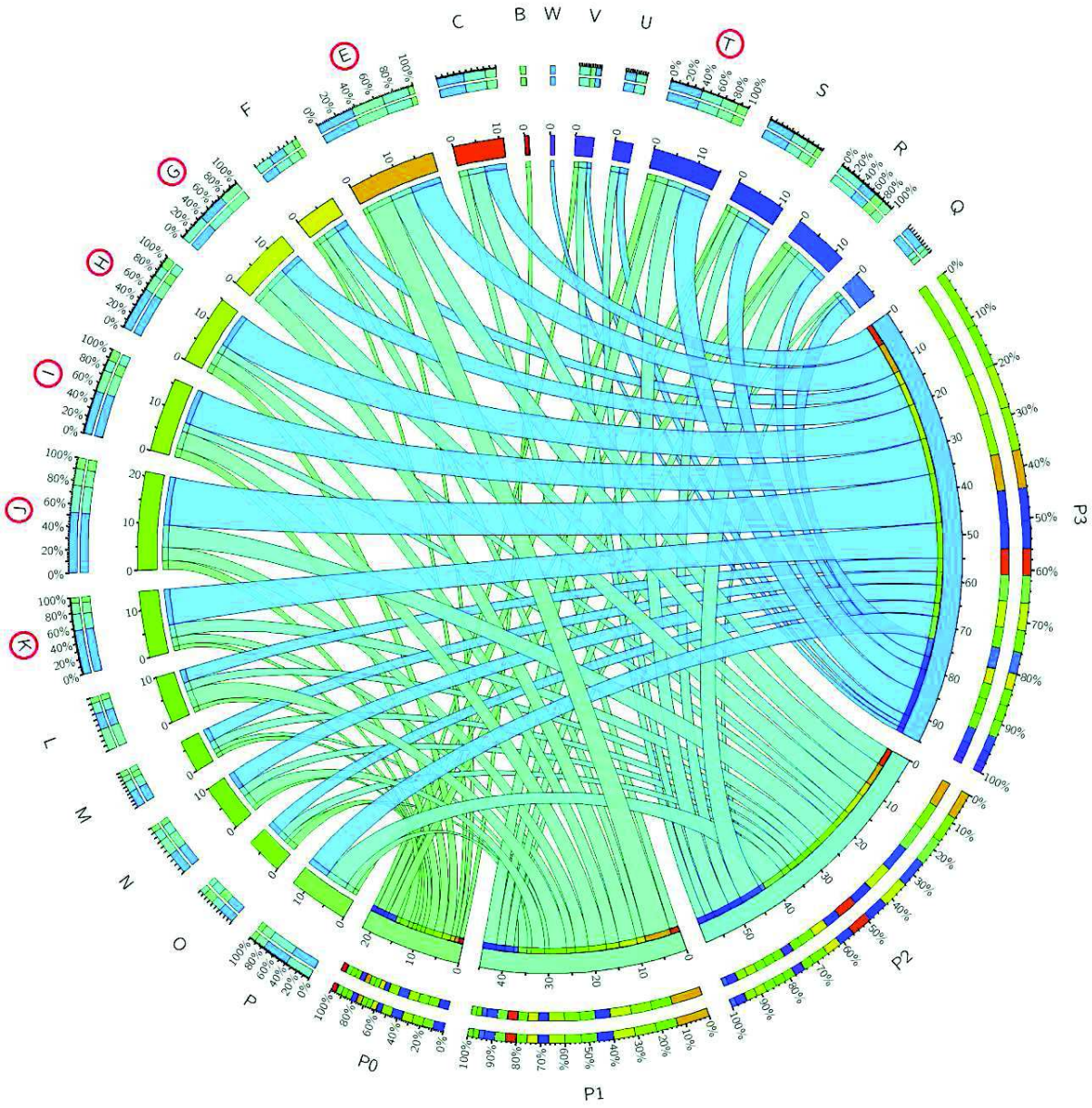
La distribución de enriquecimientos y empobrecimientos funcionales en las cuatro estrategias definidas permitió explorar los flujos de intercambio y la relevancia en cada una de ellas (**figuras C3.10 y 3.11**). Entre los simbiontes e intracelulares encontramos un enriquecimiento mayor en las categorías tráfico intracelular y secreción vesicular (COG U) y mecanismos de transducción de señales (COG T), mientras que las categorías más enriquecidas en comensales y organismos de vida libre fueron mecanismos de defensa (COG V) y mecanismos de transducción de señales (COG T). En el caso de los patógenos obligados la categoría que en más especies apareció enriquecida fue (COG N), también enriquecida en varias especies patógenas oportunistas como *Xylella fastidiosa, S. suis, Acinetobacter baumannii, L. monocitogenes* o *S. pneumoniae* (**figura 3.9**), junto a mecanismos de defensa (COG V) y (COG A). En estas últimas, patógenos obligados y sobre todo oportunistas, se encontraron gran parte de los empobrecimientos significativos en las categorías transcripción (COG K), mecanismos de transducción de señales (COG T) y categorías relativas al metabolismo celular: metabolismo y transporte de aminoácidos (COG E), metabolismo de conzimas (COG H), metabolismo y transporte iónico (COG I), metabolismo de carbohidratos (COG G).

Entre las categorías relativas a la información, procesamiento y almacenamiento, observamos un empobrecimiento en aquellas relativas al metabolismo de RNA y proteínas (COG J y COG K), mientras que se dieron más enriquecimientos en las relativas al mantenimiento, reparación y replicación de DNA (COG L) y procesamiento de RNA (COG A) (**figura C3.8**). Esta misma distribución se detectó en estudios de recombinación llevados a cabo en *E. coli* (Mau *et al*., 2006) y responde a la teoría de la complejidad (Jaint *et al*., 1999). Las observaciones sugieren que esta misma tendencia podría ser común en varias especies, donde genes informacionales que participan en procesos de interacción macromolecular, como es el caso de



Clase ecológica	J	A	K	L	B	D	Y	V	T	M	N	Z	W	U	O	C	G	E	F	H	I	P	Q	R	S
P0= Simbiontes/Pat. Intracelulares (9)	1	1	0	2	1	0	0	0	2	1	1	0	1	2	0	1	1	0	0	0	0	0	0	0	0
P1= Comensales/Vida libre (16)	2	0	0	0	0	0	0	2	2	1	1	0	0	1	0	0	2	0	1	0	0	2	0	0	1
P2= Patógenos obligados (14)	1	1	0	1	0	1	0	1	0	0	2	2	0	1	0	0	1	0	0	0	0	0	0	0	1
P3= Patógenos oportunistas (15)	1	3	0	0	1	0	0	4	0	1	4	0	1	0	0	0	0	1	1	0	0	0	0	1	2

Figura C3.10. Distribución de enriquecimientos para cada una de las cuatro estrategias ecológicas estudiadas (P0 a P3) definidas en la leyenda. El diagrama superior muestra la proporción de especies enriquecidas en cada COG dentro de cada estrategia ecológica. En el caso de patógenos oportunistas se aprecia una contribución importante de las categorías COG N, COG V y COG A, rodeadas en verde. La tabla inferior detalla el número de especies con enriquecimientos por estrategia ecológica y COG.



Clase ecológica	J	A	K	L	B	D	Y	V	T	M	N	Z	W	U	O	C	G	E	F	H	I	P	Q	R	S
P0= Simbiontes/Pat. Intracelulares (9)	2	0	1	1	1	0	0	1	2	1	1	0	0	0	2	0	0	1	1	2	2	1	0	2	1
P1= Comensales/Vida libre (16)	3	0	2	4	0	0	0	0	3	1	3	0	0	0	1	2	4	6	2	3	4	2	1	2	2
P2= Patógenos obligados (14)	5	0	3	2	0	0	0	2	4	2	2	0	0	2	2	4	5	5	3	2	2	5	1	5	3
P3= Patógenos oportunistas (15)	1	0	8	3	0	0	0	1	6	3	4	0	1	2	3	5	5	7	3	8	8	5	4	3	5

Figura C3.11. . Distribución de empobrecimientos para cada una de las cuatro estrategias ecológicas estudiadas (P0 a P3) definidas en la leyenda. El diagrama superior muestra la proporción de especies enriquecidas en cada COG dentro de cada estrategia ecológica. En el caso de patógenos oportunistas se aprecia una contribución importante de las categorías COG K , COG J, COG K, COG G, COG E y COG T, rodeadas en rojo. La tabla inferior detalla el número de especies con empobrecimientos por estrategia ecológica y COG.

los procesos de transtripción y traducción estarían menos sujetos a transferencia horizontal mientras que los operacionales, como podría ser el caso de enzimas citosólicas, estarían más involucrados. Sin embargo, más del 2,5% de los genes anotados se relacionaron directamente con la síntesis de RNA, de los cuales un 0,5 % correspondieron con rRNA y un 2% con tRNAs, mostrando que la transferencia de estos genes entre cepas de la misma especie no tiene por qué ser siempre deletérea acorde con los ejemplos mostrados en un número cada vez más creciente de estudios (Boucher *et al*., 2004; Papke *et al*., 2007; Zhaxybayeva *et al*., 2009; Williams *et al*., 2012).

Dentro de la misma categoría COG L, el término más abundante fue el de la serin recombinasa XerD (COG4974) con casi un 4% del total de genes de la categoría dentro de eventos de recombinación, como se observó en el caso de *S. ruber* (capitulo 2) o *E.coli* (Mau *et a*l., 2006). Además 7 de los 10 términos más abundantes para la categoría COG L correspondieron a elementos transponibles, los cuales representaron un 25% (1097/4411) de los términos de la categoría (**figura C3.12**). Su enriquecimiento significativo (P<0,05, pFDR<0,05, test Fisher) junto a la relación con elementos integrativos conjugativos, podría sugerir una importante contribución de los intercambios mediados por conjugación como se discutirá en el apartado 3.2. La abundancia de elementos transponibles se ha relacionado en estudios de intercambio interespecífico entre diversas especies procariotas con un papel relevante de los mecanismos conjugativos (Popa *et al*., 2012., Kloesges *et al*., 2010). Tanto elementos transponibles como serin recombinasas, comunes en GIs y que a menudo acompañan genes de virulencia (Dobrind *et al.*, 2004; Fernández-Gómez *et al*., 2012; Bellanger *et al*., 2013), son responsables de transferencias e incorporación de elementos accesorios a regiones sinténicas como se ha observado en especies del género *Streptomyces* (Ikeda *et al*., 2003) o en *S. ruber* (capítulo 1, apartado 4 y capítulo 2, apartado 4.2). A menudo se localizan flanqueando GIs (Hsiao *et al*., 2005) y participando en la incorporación de factores de virulencia como sistemas de secreción de tipo IV (Ambur *et al*., 2009). En el caso de *S. ruber* detectamos un enriquecimiento de elementos transponibles en las HRVs (capítulo 2 apartado4.1), participando activamente en el intercambio genético entre GIs y el resto del genoma. Para esta misma especie el gen *Xer*D, presente en la mayoría de replicones incluyendo algunos plásmidos, contribuyó al

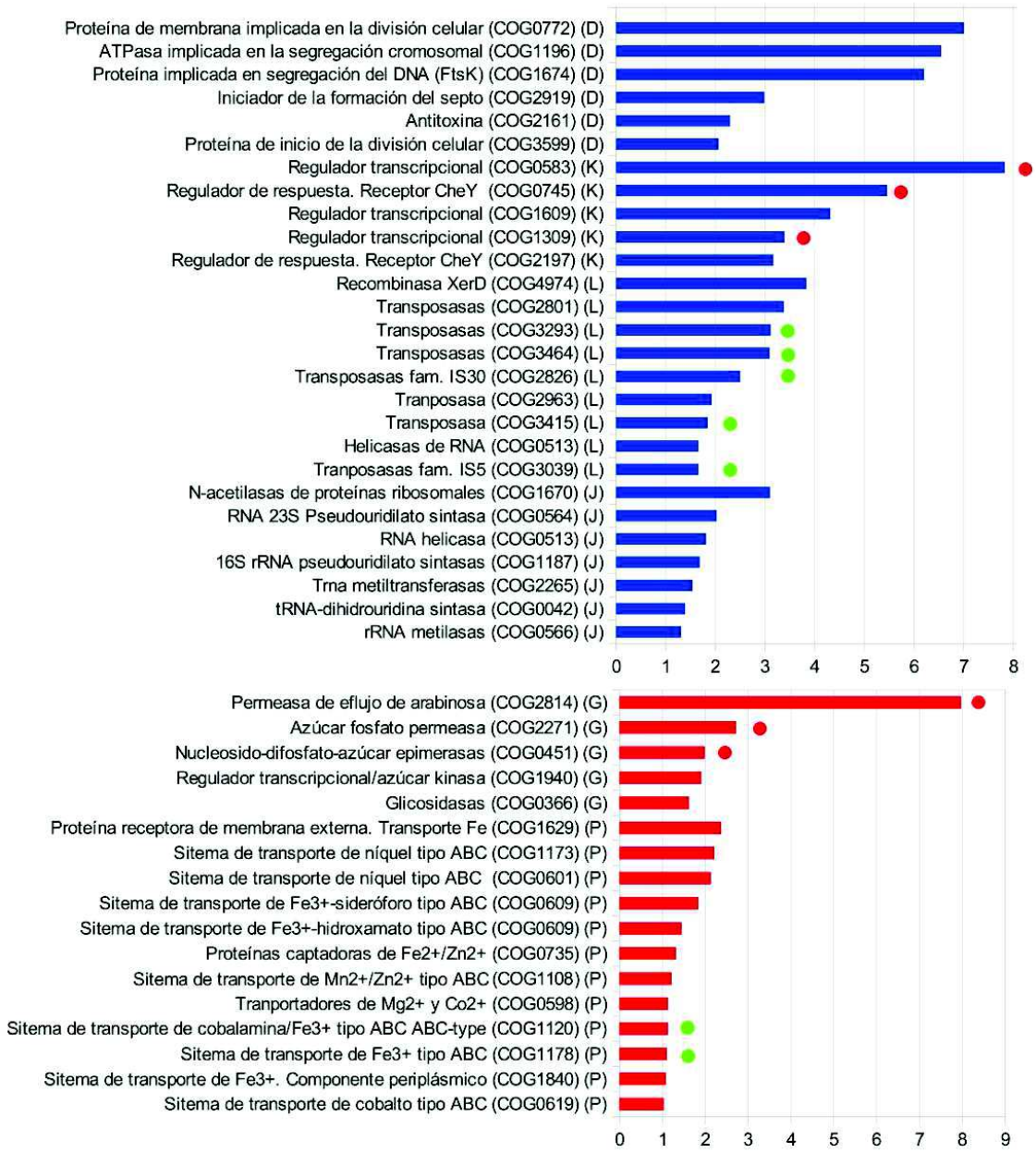


Figura C3.12. Distribución porcentual de genes en los eventos de recombinación de las 54 especies analizadas para los términos COG más abundantes dentro de las categorías división celular y cromosómica (COG D), transcripción (COG K), replicación, recombinación y reparación de DNA (COG L) y traducción, estructuras ribosómicas y biogénesis (COGJ) pertenecientes a procesos celulares y metabolismo; y transporte de carbohidratos (COG G) y transporte y metabolismo de iones inorgánicos (COG P) dentro de la categoría metabolismo. Aquellos términos enriquecidos/empobrecidos se marcan con puntos verdes y rojos respectivamente.

intercambio e inserción entre elementos del genoma accesorio, GIs y plásmidos, y regiones sinténicas del genoma core generando *indels* (capítulo 2 apartado4.2). La sinergia entre los elementos transponibles simples IS y la recombinación homóloga reside en que, pese a que este tipo de recombinación disminuye rápidamente con la divergencia de secuencia, tan sólo requiere de regiones de entre 25-200 pb flanqueando el inicio y final de la región transferida (Hacker y Carniel 2001; Thomas y Nielsen., 2005; Mau 2006). Dos elementos transponibles simples situados en el genoma a cierta distancia pueden actuar como extremos conservados contribuyendo a la movilización de los genes situados entre ellos, e incorporando nuevos genes a la especie, que ya pueden transferirse a otras cepas en procesos de adaptación al rodearse de regiones sinténicas (Mavingui *et al*., 2002; Thomas y Nielsen., 2005; Polz y Hanage 2013).

Aunque no mostraron un enriquecimiento significativo, entre los genes más transferidos de la categoría COG K, más de un 8% estuvieron relacionados con la transducción de señales en procesos de movilidad celular, precisamente algunos de ellos como *CheY* implicados en los mecanismos quimiotácticos. En el caso de la categoría COG J, aunque en ningún caso enriquecido significativamente, entre los términos más abundantes se encuentran metilasas y uridilasas de rRNA y acetilasas de proteínas ribosómicas, que engloban más del 8% de los genes trasferidos por recombinación homóloga en esta categoría. Estas últimas enzimas modifican la procesividad ribosomal y favorecen la regulación de la síntesis de proteínas (Kamita *et al*., 2011; Hassan *et al*., 2012). La metilación de rRNAs se considera un mecanismo de resistencia a beta-lactámicos, que tienen como diana de acción las subunidades 30S y 50S ribosomales (Kotra *et al*., 2000; Benítez-Páez *et al*., 2014). Se ha observado la adquisición de resistencia a estos antibióicos derivada de la metilación de rRNA 16S en diferentes especies de *Enterobacteriaceae* como *Kleibsiella pneumoniae*, *E. coli* y *Enterobacter cloaceae* por la incporporación de plásmidos (Wu *et al*., 2008). Resultados similares en otras bacterias del género *Acinetobacter* y en *P. aeruginosa* asocian la adquisición de resistencias multidroga y la capacidad de coproducción de beta-lactamasas y metalo-lactamasas a la adquisición de estas metilasas desde plásmidos gracias a eventos de transposición (Doi y Arakawa 2007). Los resultados obtenidos en este capítulo muestran que la adquisición de este mecanismo por recombinación homóloga puede ser una estrategia empleada por diversas especies procariotas.

Entre las categorías relativas a procesamiento celular revisamos la abundancia de términos dentro de las categorías mecanismos de defensa (COG V), mecanismos de transducción de señales (COG T), motilidad celular y secreción (COG N) y biosíntesis de envueltas celulares (COG M) (**figura C3.13**). Los términos más abundantes de la categoría COG T reflejan la elevada recombinación de genes implicados en la regulación quimiotáctica acorde con los más abundantes dentro de la categoría COG K. Entre los genes sobre representados en regiones recombinadas encontramos los que codifican la proteína de estrés universal de unión a nucleótidos UpsA, implicada en el control del estrés oxidativo y también identificada como una de las que más recombinó en *S.ruber* (capítulo 2, apartado 5). En el caso de las categorías COG N y COG V, encontramos gran cantidad de proteínas relacionadas con mecanismos de resistencia y patogenicidad, vinculadas en muchos casos a plásmidos y GIs (Hacker y Carniel 2001). Dentro de la categoría COG V los dos términos más abundantes, representando más del 50% de los genes recombinados en esta categoría, correspondieron con sistemas de transporte multidroga o resistencia a antibióticos y proteínas de resistencia a beta-lactámicos, muchos de ellos enriquecidos ($P<0,05$, $pFDR<0,05$, test Fisher). Ambos términos aparecieron enriquecidos entre los eventos de recombinación de *S. ruber* (capítulo 2, apartado 5) y presentes en la bacteria patógena de plantas *Xantomonas camprestris* (Huang *et al*., 2015). Además de su relevancia en procesos de adaptación y evasión en especies patógenas, este tipo de transportadores, junto a las proteínas de resistencia a bacteriocinas, podrían mediar procesos de comunicación celular o competencia entre cepas de una misma especie como se comentó al final del capítulo 1. Además encontramos un enriquecimiento significativo de componentes de sistemas de MR tipo I, lo que apoya la ubicuidad de la transferencia y heterogeneidad intraespecífica tal como se ha observado en *H. pylori* (Corvaglia *et al*., 2012) o en *S. ruber* (Peña *et al*., 2010). Este tipo de sistema MR es además el más común en GIs de bacterias marinas (Fernández-Gómez et al., 2012). Entre los términos más abundantes dentro de la categoría COG N encontramos la presencia de elementos de los sistemas de transporte tipo II y tipo IV, habitualmente presentes en plásmidos y GIs de bacterias marinas de vida libre (Persson *et al*., 2009; Fernández-Gómez *et al*., 2012) y patógenas (Kado *et al*., 2009). Los sistemas de secreción de tipo II y IV están involucrados directamente en procesos de interacción virus-hospedador, formación de *biofilms,* secreción de factores de

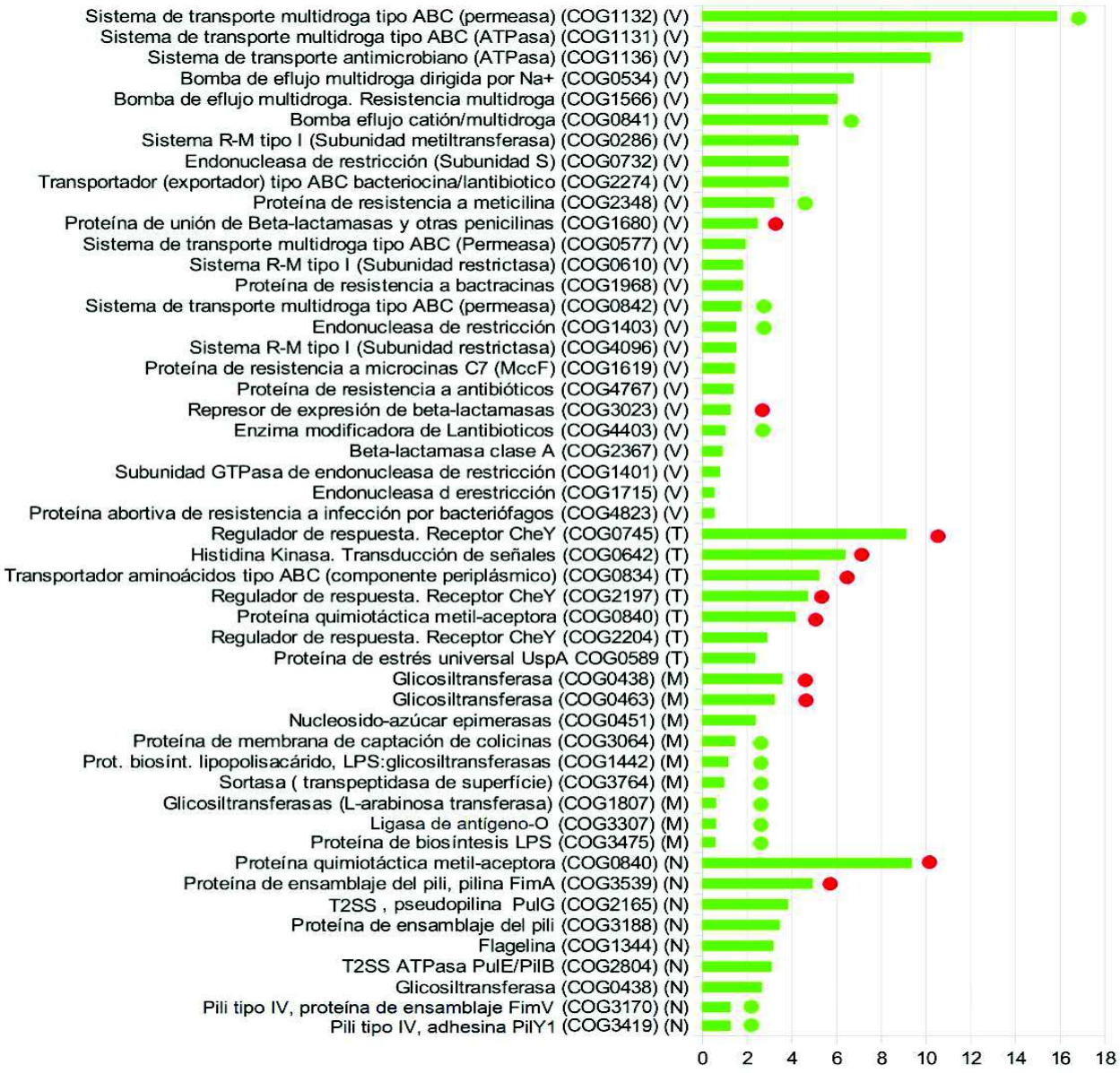


Figura C3.13. Distribución porcentual de genes en los eventos de recombinación de las 54 especies analizadas para los términos COG más abundantes dentro de las categorías mecanismos de defensa (COG V), mecanismos de transducción de señales (COG T) y motilidad celular y secreción (COG N), pertenecientes a procesos celulares. Aquellos términos enriquecidos/empobrecidos se marcan con puntos verdes y rojos respectivamente.

virulencia y adhesión a superficies, en el caso del sistema tipo IV además en procesos de transferencia de DNA mediante conjugación (Melville y Craig 2013). Se ha descrito la presencia de genes de sistema de secreción tipo IV en eventos de recombinación de especies como el patógeno humano *Bartonella henselae* (Nandi *et al*., 2015) y *B. pseudomallei* (Guy *et al*., 2012), en este último caso relacionado directamente con la capacidad virulenta de la especie. Además el gen *Fim*A, el más representado con diferencia dentro en la categoría COG N, es el principal factor de virulencia asociado a *Streptococcus parasanguis* (Yi-Ywan *et al*., 2011), patógeno que genera endocarditis infectiva, a *Streptocossus sanguinis* (Xu *et al*., 2007) y a *P. gingivalis*, patógeno bucal que causa periodontitis y en el cual resulta esencial en los procesos de adhesión y colonización (Amano *et al*., 2004; Moreno y Contreras 2013).

Dentro de la categoría COG M encontramos una gran proporción de glicosiltranferasas y epimerasas, la mayoría de ellas enriquecidas significativamente ($P<0,05$, $pFDR<0,05$, test Fisher) encargadas de la transferencia de residuos glucídicos a lipopolisacáridos y proteínas de membrana y transpeptidasas. Este tipo de proteínas se encuentran generalmente codificadas en GIs dentro del cluster de biosíntesis del antígeno-O tanto de organismos de vida libre como patógenos (Dobrindt *et al*., 2004; Fernández-Gómez *et al*., 2012; Gonzaga *et al*., 2012; Martín Cuadrado *et al*., 2015). Además de estas glicosiltransferasas, entre los términos enriquecidos encontramos ligasas del antígeno-O y biosíntesis de lipopolisacáridos. Tal como se observó en el caso de *S. ruber* (apartado 5 de los capítulos 1 y 2), la recombinación homóloga juega un papel relevante como mecanimo de transferencia de genes implicados en la generación de envueltas celulares que además, en el caso de las glicosilasas, participaron en la respuesta específica de cepa al interaccionar M8 y M31. En algunos organismos de vida libre, la diversidad del antígeno-O se atribuye a la presión ambiental generada por virus y en bacterias patógenas oportunistas u obligadas a mecanismos adaptativos para eludir la respuesta inmune (Schmidt y Riley, 2003; Wang y Quinn 2010).

Por último, y pese al empobrecimiento observado en algunas categorías relacionadas con procesos metabólicos, observamos la presencia relevante de algunos términos COG dentro de las categorías transporte y metabolismo de carbohidratos (COG G) y de iones inorgánicos (COG P) (**figura 3.12**). En el caso de COG G encontramos una buena representación de términos

relacionados con el empleo de azúcares para la modificación de elementos de superficie y envueltas celulares mencionado anteriormente. Dentro de la categoría COG P, los términos más representados se relacionaron con sistemas de transporte de Fe y Ni en su mayoría, los cuales se consideran que contribuyen a la adaptación en ambientes en condiciones limitantes de hierro, y se localizan frecentemente tanto en GIs de patogenicidad asociadas a cepas con una mayor virulencia (Haecker y Carniel 2001) como en GIs de organismos de vida libre, tal como se observó en *S. ruber*. Este tipo de transportador se ha encontrado de manera abundante en eventos de recombinación de bacterias acuáticas de vida libre *A. macleodii* (Gonzaga *et al*., 2012) y *S. ruber* (capítulo 2 apartado 5) y la patógena humano *B. henselae* (Nandi *et al*., 2015).

En conjunto, los resultados obtenidos para diferentes categorías funcionales apoyan la relevancia de la recombinación homóloga en los procesos de transferencia alélica entre genomas *core* en procesos de patogénesis y adaptación tanto a hospedador como a variaciones ambientales en bacterias de vida libre.

3.2- Factores genómicos relacionados con la especialización ecológica y estrategias de intercambio de DNA intraespecífico.

Los procesos de adaptación, especialización y adquisición de estrategias ecológicas comportan cambios genómicos, algunos discutidos en el apartado anterior, entre los que podemos encontrar estrategias de captación de DNA. Los procesos de especialización ecológica a menudo se asocian a eventos de ganancia o pérdida de capacidad competente, cambios en la maquinaria de reparación y recombinación y variación de la diversidad y versatilidad del genoma accesorio, mayor en organismos comensales y patógenos. Un ejemplo sería incorporación de GIs de patogenicidad, consideradas puntos calientes de incorporación de elementos del *pool* génico ambiental que favorecen procesos de evolución rápida frente a cambios ambientales, ya que pueden transferise en bloque permitiendo aumentar el *fitness* y adaptación de manera más frente a cambios ambientales o de especialización ecológica en el acceso a nichos con características diferentes (Hacker y Carniel 2001; Kado *et al*., 2009). Por tanto, ya sea por ganancia o pérdida de genes mediante HGT, la estructura genómica refleja el estilo de vida de un microorganismo

(Dobrind *et al*., 2004). En este sentido, y debido a la relación íntima con la adaptación y especialización ecológica nos resultó interesante explorar que aspectos genómicos acompañan a los patrones de recombinación y especialización ecológica observados anteriormente.

Sistemas de recombinación y reparación homóloga. Como se comentó en la introducción de esta tesis, la principal función de los procesos de recombinación homóloga mediada por RecA es la reparación, replicación y mantenimiento de la integridad del material genético. Pero además la recombinación homóloga juega un papel clave en la diversificación de genomas procariotas al actuar como vía dependiente de identidad de secuencia para la incorporación de DNA adquirido mediante conjugación y transformación (Thomas y Nielsen., 2005; Rocha *et al*., 2005; Michod *et al*., 2008). La distribución de 21 genes implicados en reparación, recombinación y mantenimiento de DNA y la correlación entre el contenido en GC y tamaño genómico ha sido estudiada en más 900 genomas de especies procariotas (Rocha *et al*., 2005; García-González *et al*., 2013). Según la teoría de la constricción proteómica (García González *et al*., 2012, 2013; Massey 2008, 2013), el incremento en contenido génico y secuencias codificantes conlleva un mayor efecto de la presión selectiva ambiental ya que la mayoría de mutaciones resultan potencialmente deletéreas, lo cual conduciría a la disminución drástica del tamaño efectivo poblacional y de la diversidad de la misma. De este modo, genomas de mayor tamaño requieren de una batería de genes implicados en procesos de recombinación y reparación de DNA, entre ellos los genes *Rec* y *Mut*. La ausencia de estos últimos, entre los que destacan *mut*M y *mut*Y, implicados directamente en corregir transversiones de GC a AT, se ha correlacionado en estudios previos con una disminución en el contenido en GC y tamaño genómico (González-García *et al*., 2012; Wu *et* al., 2012), representada también por los genomas incluidos en este capítulo (**figura 3.5**). Además en estos estudios se observó una clara relación entre el contenido en genes *Rec* y la estrategia ecológica en varias especies procariotas (García-González *et al*., 2013), detectando diferencias significativas entre organismos simbiontes y patógenos intracelulares, con ausencia de muchos de estos genes *Rec* de acuerdo a su estrategia reduccionista comentada anteriormente y a la teoría de la constricción proteómica, respecto a patógenos oportunistas, patógenos obligados, comensales y organismos de vida libre. También en estos mismos estudios se apreciaron diferencias significativas entre comensales y organismos

de vida libre y el conjunto de patógenos no intracelulares (obligados y oportunistas), donde estos últimos en general mostraron una mayor abundancia y colocalización de genes *Rec,* en especial los genes presinápticos que constituyen los sistemas RecBCD y RecFOR y postsinapticas,. Entre estos dos últimos grupos de patógenos no se observaron diferencias significativas.

El patrón y orden establecido entre estas estrategias ecológicas por García-González y colaboradores en base a la abundancia de genes *Rec* coincide con el establecido en base a la abundancia de eventos de recombinación detectados en este capítulo (apartado 3.1). Los análisis de comparación de medias para los eventos de recombinación mostraron un mayor número de eventos en patógenos oportunistas y obligados, seguidos de organismos de vida libre, como se comentó en el apartado anterior ($p< 0,05$, test de Jonckheere-Terpstra) (**figura 3.6**). Además se observaron diferencias significativas en el número de eventos de recombinación distribuidos en los intervalos mostrados en la figura 3.4 ($p< 0,05$, test de Jonckheere-Terpstra). En estos últimos se detectaron niveles de recombinación homóloga intermedios a los de patógenos y organismos intracelulares y simbiontes. Parece pues que podría establecerse una relación directa entre los niveles de recombinación detectados y la funcionalidad de la batería de genes *Rec* disponible para la célula. Para profundizar en esta relación funcional se analizó la relación entre la densidad de secuencias Chi conocidas y presencia de eventos de recombinación y genoma recombinado, cuestión abierta y planteada en revisiones anteriores (Vos *et al*., 2009). Estas secuencias juegan un papel clave en los procesos de recombinación homóloga ya que las proteínas presinápticas interaccionan con RecA tras alcanzar una secuencia Chi, proceso que conduce a la formación posterior de la estructura conocida como *Holliday junction* y finalmente el complejo postsináptico (Rocha *et al*., 2005; revisado en Dillingham y Kowalczykowski 2008). Empleamos las secuencias Chi consenso descritas anteriormente para 7 de las especies consideradas en este estudio *E. coli* (Ponticelli *et al*., 1985, El Karoui *et al*., 1999; Halpern *et al*., 2007), *S. aureus* (Halpern *et al*., 2007) *B. subtilis* (Chédin *et al*.,1998; El Karoui *et al*., 1999), *S. thermophilus* (Halpern *et al*., 2007), *S. pneumoniae* (Halpern *et al*., 2007), *L. lactis* (Biswas *et al.*, 1994; El Karoui *et al*., 1999) y *H. influenzae* (Sourice *et al*., 1998; El Karoui *et al*., 1999) para calcular la densidad de cada una de ellas de manera individual en cada una de sus cepas. Los valores obtenidos, que miden la frecuencia de aparición de las secuencias Chi en el genoma

(Kb/secuencia), fueron similares a los establecidos previamente y correlacionaron positivamente con los eventos/cepa detectados (p< ; r^2= 0,57) (**figura C3.14**). Por tanto, la presencia y densidad de estas secuencias favorece la incidencia de eventos de recombinación como ya se sugirió en el caso de *R. solanacearum* (Fall *et al.*, 2007). Trabajos previos sugieren la evolución independiente de estas secuencias en diferentes especies atendiendo a las presiones ambientales, especialmente aquellas que afectan a la estabilidad del material genético de la célula (El Karoui *et al*., 1999). La correlación de la densidad de estas secuencias con la frecuencia de recombinación, y por consiguiente con el contenido en genes *Rec,* apoya su papel en las funciones mediadas or RecA y la coevolución de estos genes y secuencias. Precisamente estos elementos, genes *Rec* y secuencias Chi, se encuentran en mayor densidad en organismos patógenos oportunistas y obligados, para los cuales detectamos niveles de recombinación elevados como discutió anteriormente. Otro aspecto analizado en diferentes especies (*E. coli, B.*

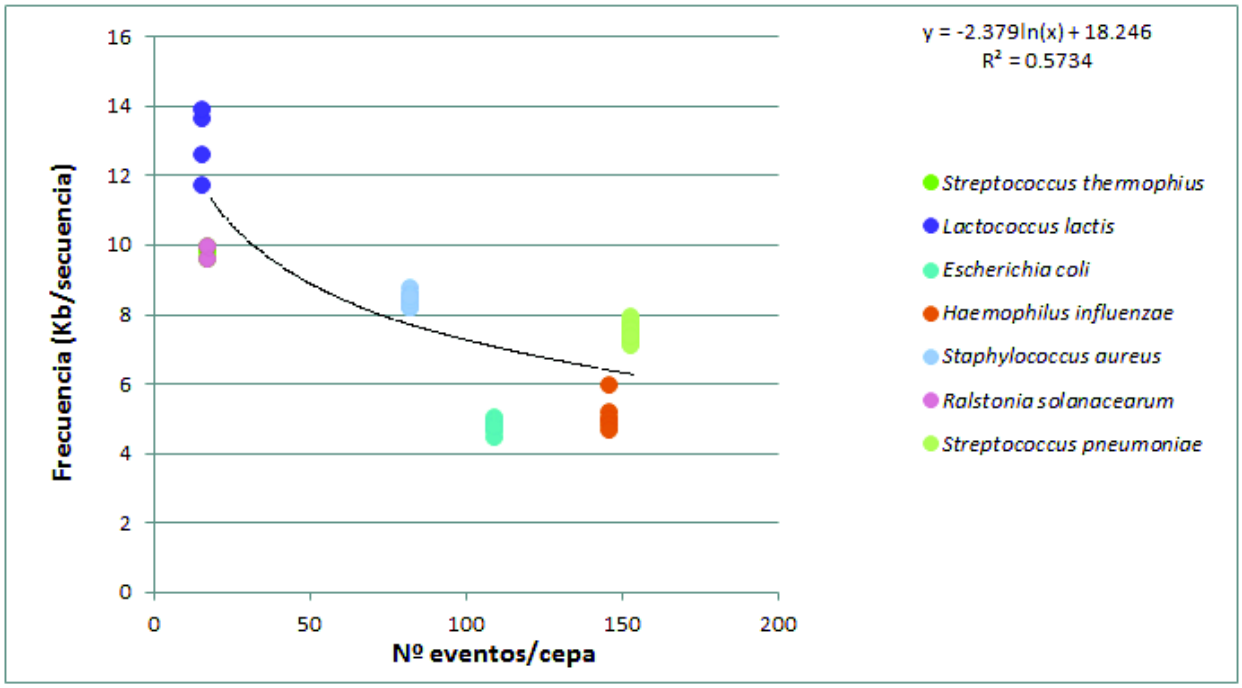


Figura C3.14. Correlación entre la densidad de secuencias Chi y la frecuencia de recombinación homóloga en los genomas de *S. thermohilus, L. lactis, E. coli, H. influenzae, S. aureus, R. solanacearum y S. pneumoniae*.

subtilis, H. influenzae y *L. lactis*) fue la distribución de estas secuencias a lo largo del cromosoma, la cual parece no ser aleatoria (El Karoui *et al*., 1999). Según estudios previos, las secuencias Chi se encontrarían en una densidad mucho mayor en regiones pertenecientes al genoma *core* contribuyendo a la estabilidad de esta fracción que contiene genes esenciales (Halpern *et al.*, 2007). En el caso de *R. solanacearum*, se encontró una mayor densidad de secuencias Chi y tasas de transferencia para regiones que contienen genes *housekeeeping*, entre ellos *mut*S y *rec*A (Fall *et al.*, 2007). La frecuencia de recombinación de estos dos genes apenas disminuyó con la distancia génica entre las cepas aisladas, limitando el efecto de la divergencia de secuencia que afecta de manera crítica a la eficiencia de los procesos de recombinación (Delmas *et al*., 2005; Gogarten y Townsend 2005) a la vez que restringe su efecto a nivel de especie o género (Halpern *et al*., 2007). Las secuencias Chi actuarían como un mecanismo que regula la recombinación homóloga y por tanto la diversidad de un subconjunto de genes. En el caso de secuencias heterólogas, el mecanismo más frecuente para su incorporación a los *core* genomas implica la integración en regiones calientes como GIs, y posteriormente su desplazamiento a zonas más estables con la ayuda de elementos transponibles, mecanismo observado en *S. ruber,* facilitando su transferencia al resto de cepas (Fall *et al.*, 2007; Polz y Hanage 2013). De este modo, y por medio de mecanismos de recombinación homóloga, las secuencia Chi contribuirían al mantenimiento y cohesión del genoma *core* reforzando la estabilidad de la maquinaria basal y permitiendo el acceso y combinación de mutaciones potencialmente beneficiosas a lo largo de toda la población a la vez que reduciendo el riesgo de deriva génica por mutaciones puntuales en linajes independientes (Fall *et al*., 2007; Torsvik *et al*., 2002). En nuestro caso además encontramos una distribución similar en el mapeo de las posiciones de los eventos de recombinación y secuencias Chi en las cepas de *B. subtilis*. En las cepas en que se había observado ya la distribución de secuencias Chi (Chédin *et al*.,1998; El Karoui *et al*., 1999), ésta coincidió perfectamente.

Conjugación y transformación. La conjugación y la transformación son los dos mecanismos predominantes en los intercambios de material genético entre cepas cercanas (Tribble *et al*., 2012). Esto es debido a que el éxito de ambos procesos depende de la recombinación homóloga mediada por RecA durante la integración del material intercambiado,

lo que limita las trasferencias potenciales a un umbral de homología entre donador y receptor (revisado en Thomas y Nielsen 2005; revisado en Didelot y Maiden 2010). En el caso de la HGT interespecífica, el papel relevante corresponde a los mecanismos de transducción mediados por enzimas víricas, tal como indican estudios previos basándose en el contenido elevado en genes víricos en los fragmentos intercambiados (Ochman *et al*., 2000; Thomas y Nielsen 2005; ; Kloesges *et al*., 2010; Popa *et al*., 2011) y procesos dependientes del sistema de reparación de rotura de doble cadena de DNA no homóloga mediados por Ku y LigD (Kloesges *et al*., 2010) (NHEJ, del inglés *nonhomologous end-joining*) (Shyuman y Glickman 2007; Lieber 2010; Davids y Chen 2013). En el caso concreto del *phylum Proteobacteria,* la conjugación actuaría como mecanismo relevante en los intercambios de DNA (Kloesges *et al*., 2010). Entre los genes intercambiados entre especies de este último *phylum* un 4% habitualmente se encuentran en plásmidos y codificaron elementos transponibles, integrasas y elementos de estabilización y un 2% proteínas relacionadas con fagos.

Con el objetivo de analizar el efecto de la capacidad competente en los niveles de recombinación detectados, distribuimos las especies estudiadas en dos clases funcionales, competentes y no competentes basándonos en el contenido en genes del regulón *com,* que codifica elementos indispensables en la maquinaria de transformación (Claverys *et al*., 2006). Aunque no se apreciaron diferencias significativas en el porcentaje del genoma total recombinado entre ambas clases ($p>0,05$, test de Jonckheere-Tepstra) si se dieron en el número de eventos de recombinación, sobre todo al comparar la distribución de aquellos de menos de10 Kb (**figura C3.15**). Estos datos fueron consistentes con la buena correlación obtenida entre el contenido en genes *com* y los eventos menores de 10 kb con un coeficiente de correlación de Pearson mayor en aquellos fragmentos menores de 10 kb ($p< 0,05$ $r^2= 0,317$, correlación lineal) e incrementándose para la fracción de menos de 2 kb ($p< 0,01$ $r^2= 0,414$, correlación lineal). En conjunto los datos anteriores sugieren que aquellos organismos competentes incorporan una mayor proporción de regiones cortas ya que, aunque el porcentaje de genoma total recombinado es similar entre organismos que difieren en su capacidad competente, la cantidad de eventos recombinantes es mayor en los competentes. Estas observaciones concuerdan con el elevado contenido en eventos de recombinación de menos de 1Kb en organismos competentes como *L.*

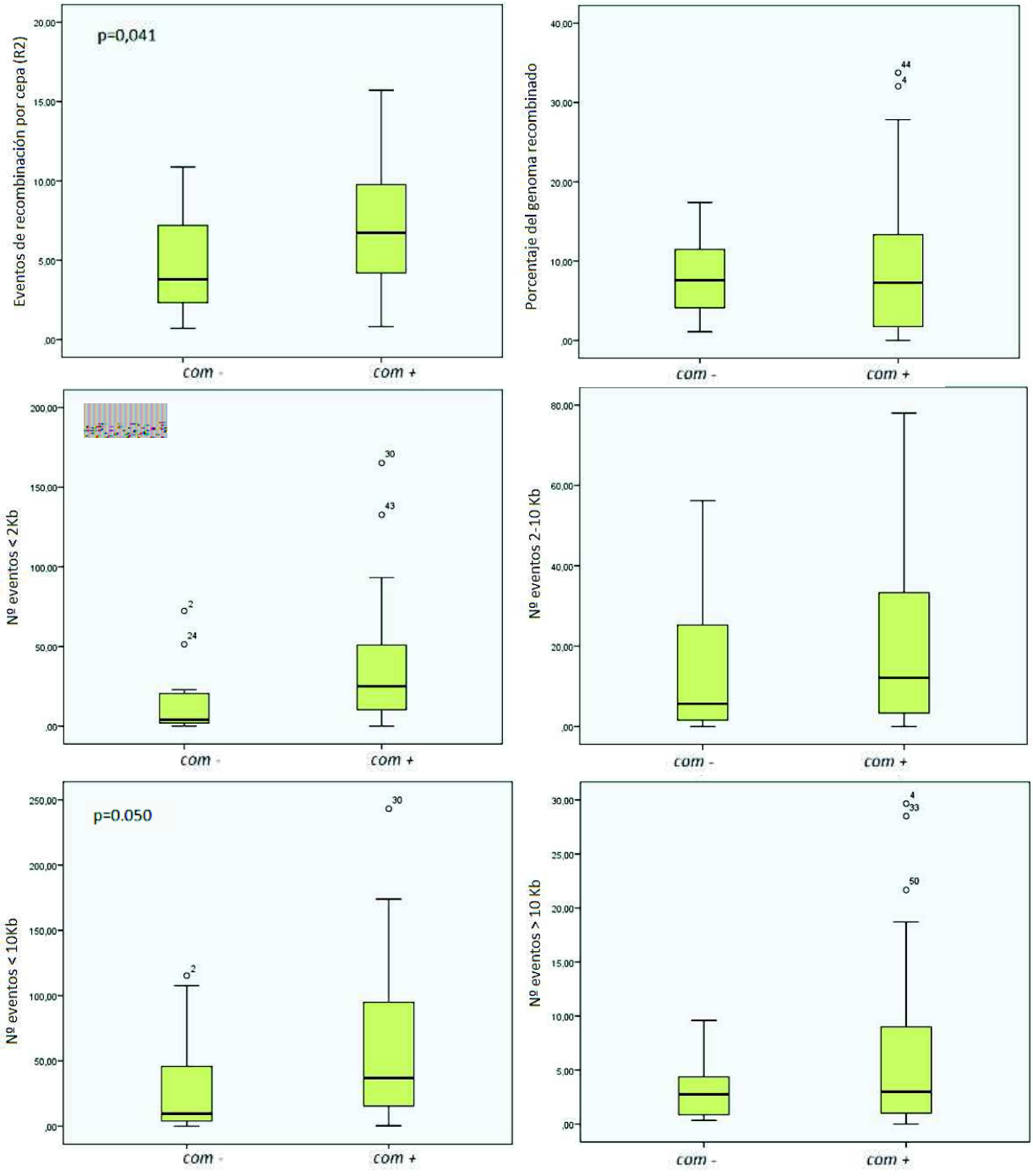


Figura C3.15. Diagramas de caja muestran las diferencias en los promedios de eventos de recombinación, porcentaje del genoma recombinado y distribución de tamaño de eventos en las 54 especies analizadas en función de su capacidad competente (0= No competente 1= Competente) En caso de ser significativas las diferencias ($p<0,05$, test de la T) se muestran los valores de p.

monocitogenes (Den Backer *et al*., 2010) y *S. aureus* (Falush *et al*., 2001). Existe además una correlación directa entre la proporción de eventos de entre 2-10 kb y el contenido en elementos transponibles ($p< 0,05$ $r^2= 0,72$, correlación lineal) y entre el contenido de genes *com* y elementos transponibles ($p< 0,05$ $r^2= 0,37$, correlación lineal) que sugiere la participación activa de estos en la incorporación de fragmentos de menor tamaño que, como se comentó anteriormente, facilitan la incorporación de regiones por recombinación homóloga. Mediante mecanismos alternativos, probablemente conjugativos, se produciría el intercambio de la mayoría de eventos de un tamaño superior a las 10 kb tal como se sugiere en estudios previos (Den Bakker et al., 2010; Didelot *et al*., 2010), lo que afectaría a la naturaleza de los fragmentos recombinados.

Los resultados obtenidos con los genes *com* y la gran presencia de elementos transponibles y genes *Xer*D en los eventos de recombinación nos lleva a pensar que quizá exista una correlación entre otros parámetros genómicos y mecanismos de intercambio e integración que, además de la estrategia ecológica y capacidad de competencia, contribuyan a la comprensión de los patrones observados. Las diferencias observadas demuestran como la transformación media la transferencia de fragmentos menores de 10 kb mientras que la conjugación, tal como se sugiere en revisiones anteriores (Didelot *et al*., 2010; Thomas y Nielsen 2005) media una importante proporción de intercambios de fragmentos de gran tamaño. En conjunto, la estrategia ecológica y la capacidad competente explicaron cerca del 32% de la distribución de eventos de recombinación observada a lo largo de las 54 especies analizadas ($r^2= 0,317$; $p < 0,05$, ANCOVA, modelo lineal).

3.3- Mecanismos barrera que limitan la recombinación homóloga interespecífica e intraespecífica.

Como se ha discutido en el apartado anterior y en la introducción (apartado 2.4.1), son varios los mecanismos que limitan la especificidad de secuencia entre donador y aceptor en procesos de recombinación homóloga, entre los que el sistema de recombinación mediado por RecBCD y RecA y el sistema de reparación de errores mediado por MutB, constituyen la

principal barrera que limita el rango de *pool* genómico ambiental (Vulic *et al*., 1997; Zawadzki *et al.*, 1995). Además de estos, en este capítulo analizamos la influencia de otros sistemas barrera como factores a tener en cuenta a la hora de interpretar la heterogeneidad de estrategias de evolución de los genomas *core* mencionadas anteriormente, ya que actúan como elementos que favorecerían la recombinación homóloga al seleccionar negativamente la entrada de fragmentos heterólogos. Entre los principales mecanismos de defensa procariota encontramos los toxina-antitoxina, sistemas abortivos de infección (Labrie *et al*., 2010), sistemas de modificación-restricción (revisado en Bailiss *et al*., 2006, Roberts *et al*., 2014) y CRISPR-Cas (Mojica *et al*., 2005; Koonin y Makarova 2009., Samson *et al*., 2013), cuyos patrones de distribución en genomas procariotas, especialmente en el caso de los dos últimos, se han estudiado ampliamente (Barrangou *et al*., 2007; Vasu y Nagaraja 2013; Makarova *et al*., 2011, 2013). Debido a la mayor disponibilidad de datos para los sistemas MR y CRISPR-Cas para los genomas incluidos en nuestro estudio, decidimos centrarnos los análisis en estos dos últimos, estudiando su distribución en función de las estrategias ecológicas y su relación con los niveles de recombinación detectados.

La distribución media del número de genes relacionados con sistemas MR mostró diferencias significativas entre las cuatro estrategias ecológicas estudiadas ($p=0,001 <0,05$, test de Kruskal-Wallis y Jonkheere-Tepstra). Estas diferencias se observaron también al comparar individualmente las medias de los genes pertenecientes a cada uno de los tipos de sistemas de MR explorados (tipo I a tipo IV) en los MR de tipo I, II ($p<0,05$, test de Kruskal-Wallis y Jonkheere-Tepstra) y III ($p<0,05$, test de Kruskal-Wallis) (**figura C3.16**). La comparación de funciones individuales entre los genes involucrados indicó que aquellos con función metilasas presentaron las diferencias más significativas. Los endosimbiontes y patógenos obligados fueron los que tuvieron un menor contenido de estos tres sistemas tal como se observó en estudios anteriores (Makarova *et al*., 2012, 2013, Vasu y Nagaraja 2013). En estos estilos de vida se asocia la restricción del nicho y su estabilidad y la menor presión adaptativa por partículas víricas con la reducción del genoma y pérdida de elementos. Por el contrario, su presencia en patógenos oportunistas y organismos de vida libre fue elevada. La abundancia de sistemas MR parece corresponder con el observado al comparar la presencia de genes *Rec* y la distribución de

eventos de recombinación entre las cuatro estrategias ecológicas definidas (discutido en los apartados 3.1 y 3,2 respectivamente). Además observamos una correlación positiva significativa entre los eventos de recombinación por cepa y el contenido en genes de los sistemas de MR de tipo I y II ($p<0,05$, $r^2= 0,3$ y $r^2= 0,43$ respectivamente, Rho de Spearman) y con el contenido en genes *com* ($p<0,05$, $r^2= 0,28$ y $r^2= 0,37$, Rho de Spearman) y porcentaje del total del genoma que representan las GIs ($p<0,05$, $r^2= 0,43$ y $r^2= 0,40$, Rho de Spearman). Estos datos concuerdan con observaciones realizadas en estudios previos, que indican la presencia de un gran número de

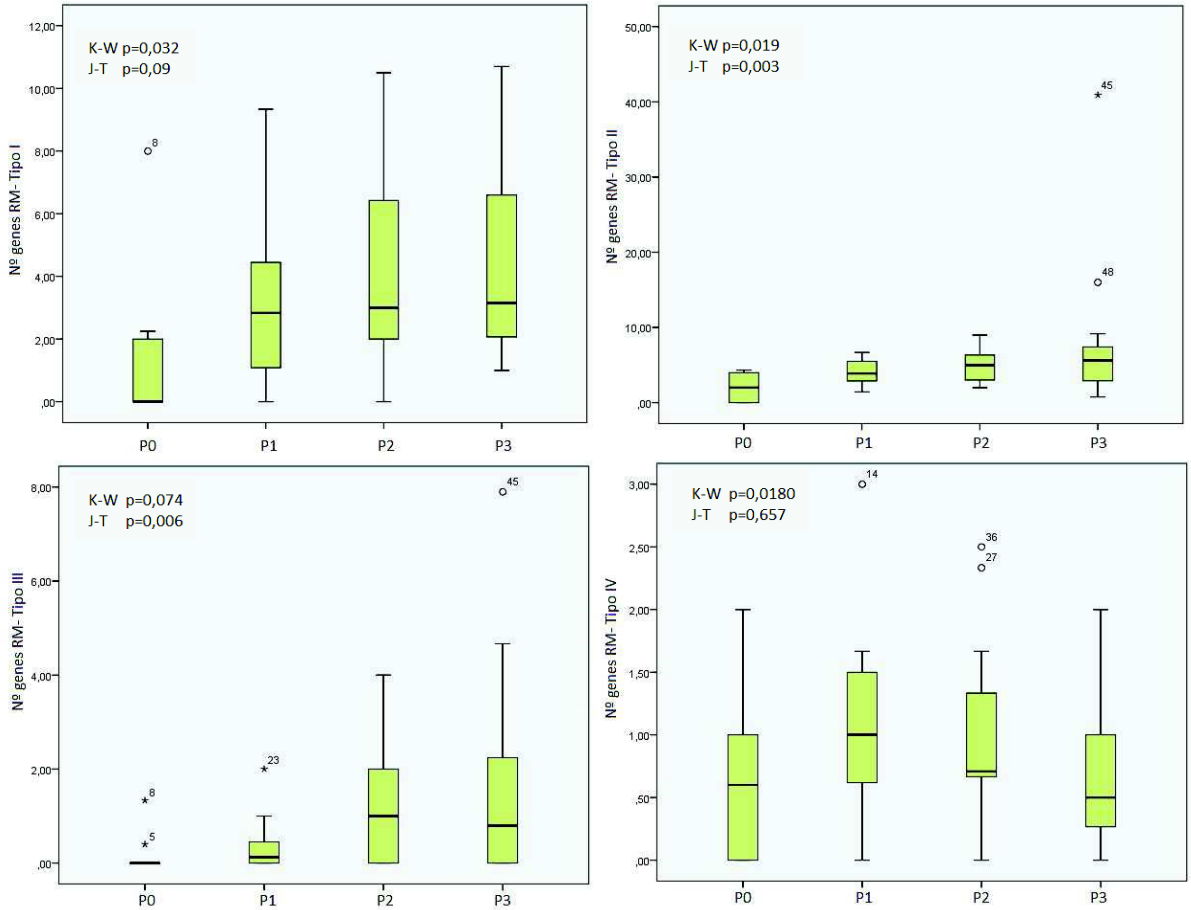


Figura C3.16. Diferencias en los contenidos de sistemas MR tipo I, II, II y IV según la estrategia ecológica. Diagramas de caja muestran las diferencias en los promedios de eventos de recombinación y porcentaje del genoma recombinado. En cada caso se muestran los p-valores significativos entre diferentes clases: Simbiontes/patógenos intracelulares (P0), no patógenos (comensales y vida libre) (P1), patógenos obligados (P2) y patógenos oportunistas (P3) ($p<0,05$, test de Kruskal-Wallis (K-W) y Jonkheere-Tepstra (J-T).

sistemas MR y tipos entre especies competentes y patógenos oportunistas, destacando *N. gonorrhoeae* (Steinn *et al*., 1999), *H. pylori* (Corvaglia *et al*., 2010), *H. influenzae* y *S. pneumoniae* (Vasu y Nagaraja 2013), en las que la entrada de material genético heterólogo justifica un mayor control y selección por mecanismos barrera.

Los niveles de correlación observada entre el número de sistemas MR tipo I y II y los eventos de recombinación sugieren que los sistemas MR juegan un papel importante en la selección y especificidad de los fragmentos potencialmente recombinables. Sin embargo, cabe recordar que también se ha observado su efecto no sólo a la hora de limitar intercambios de material genético no sólo interespecíficos sino también entre cepas cercanas de una misma especie (Tock y Driden 2005; Hoskisson y Smith 2005). La heterogeneidad en el contenido de sistemas MR entre las cepas de una especie conduce a diferencias en los patrones de metilación y en ocasiones pueden favorecer fenómenos de especiación incipiente como se ha observado en especies como *S. islandicus* (Cadillo-Quiroz *et al*., 2012) y *H. pylori* (Corvaglia *et al*., 2012). Se han identificado patrones comunes y específicos en el contenido de sistemas de MR entre diferentes líneas clonales en especies patógenas oportunistas como *L. monocitogenes* (Den Backer *et al*., 2010), *S. aureus* (Waldrow y Lindsay 2006; Basic-Hammer *et al*., 2010), *E. coli* (Sibley y Raleigh 2004), *S. enterica* (Sibley y Raleigh 2004), patógenas obligadas como *B. pseudomallei* (Nandi *et al*., 2015) y organismos de vida libre, entre ellos *S. ruber* como se muestra en el capítulo 2 de esta tesis, que presentan frecuencias de recombinación distintas al resto de líneas clonales, recombinando más entre cepas del mismo clado, y valores de dN/dS intra-clado inferiores. Estos estudios muestran la incorporación de estos sistemas al genoma accesorio, a menudo en plásmidos GIs y sujetos a selección positiva (Furuta *et al*., 2010., Corvaglia *et al*., 2010; Fernández-Gómez et al., 2012., revisado en Bellanger *et al*., 2013), y desde estos a regiones más estables del genoma (Makarova et al., 2011).

Entre los sistemas de MR, los de tipo III son los menos estudiados y se ha observado que están asociados a procesos de selección positiva. En nuestro estudio encontramos mayor diversidad y número de sistemas de MR de tipo III en patógenos oportunistas, comensales y de vida libre (**figura C3.16**). Precisamente estos fueron los que presentaron genomas accesorios más dinámicos como refleja su mayor tasa de recombinación homóloga y proporción de GI

(**figuraC3.7**) y acorde con la correlación observada entre el tamaño de islas, el contenido en genes de los MR tipo I y tipo II y la presencia de elementos transponibles. Estos datos soportan la hipótesis de que los sistemas de MR contribuye a la estabilización de elementos móviles y de GIs dentro de los cromosomas (Vasu y Nagaraja., 2013) así como el papel de los elementos transponibles en la transferencia horizontal de los mismos (Furuta *et al*., 2010., Takahashi *et al*., 2011), con los que suelen aparecer formando *clusters* o **islas de defensa** (Makarova *et al*., 2011., 2013).

Por otra parte analizamos el contenido en sistemas CRISPR-Cas y su relación con los niveles de recombinación. Por medio de una representación de regresión por cuantiles se observó la relación existente entre el porcentaje de genoma recombinado por cada especie y el número de sistemas CRISPR-Cas albergados en el mismo, representada por la línea azul el percentil 90 (**figura C3.17**). Se aprecia un máximo de genoma recombinado que disminuye con el

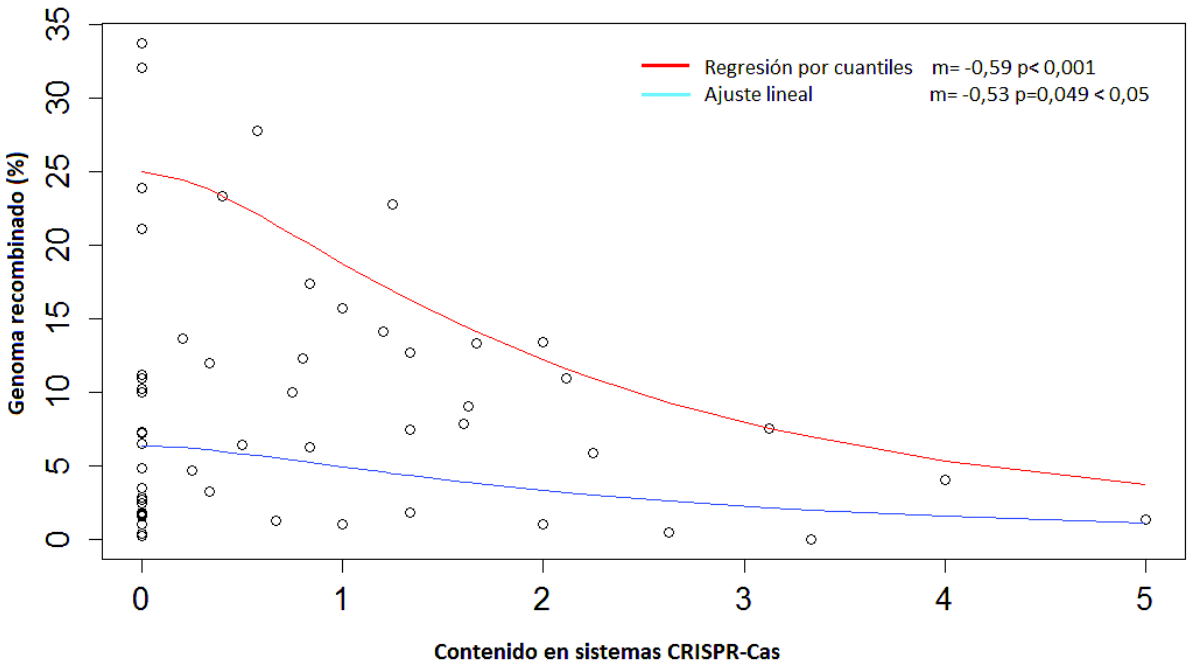


Figura C3.17. Representación de la relación entre la proporción de genoma recombinado y el contenido genómico de sistemas CRISPR-Cas. La línea roja muestra el percentil 90 para el ajuste de regresión por cuantiles realizada mediante el paquete estadístico en R Quantreg y la azul el ajuste lineal de ambas variables transformadas por logaritmo. El incremento en sistemas CRISPR-Cas limita la cantidad de genoma recombinado. La leyenda muestra las pendientes (m) y valores de p asociados a cada uno de los ajustes y ecuaciones de las curvas.

incremento de sistemas CRISPR-Cas. Aparentemente estos datos sugieren que de algún modo el efecto de los sistemas CRISPR-Cas además de actuar contra secuencias heterólogas, DNA vírico y de especies distantes, también podría afectar a secuencias homólogas captadas por conjugación o transformación. Sin embargo recientemente se ha sugerido la posibilidad de que exista una coevolución entre los genes de competencia y los sistemas CRISPR-Cas, en la cual la pérdida de la capacidad competente, por ausencia de genes *com*, se vería secundada por la pérdida de sistemas CRISPR-Cas (Jorth y Whiteley 2012). Aunque se desconocen exactamente las razones de esta coevolución, el hecho es que esta hipótesis se ha testado recientemente en genomas de la bacteria patógena oportunista *Aggregatibater actinomycetemcomitans* (Jorth y Whiteley 2012), en la que el 30% de las cepas tiene capacidad competente (Fujise *et al*., 2004). Las cepas competentes representaron linajes independientes y un mayor grado de reordenamiento génico mientras que las no competentes evolucionaron por incorporación de fragmentos derivados de plásmidos conjugativos o eventos de transducción.

Tanto la capacidad competente como sus mecanismos de regulación y sentido evolutivo y adaptativo son aspectos que se han estudiado en diferentes linajes de distintas especies procariotas (Clayvers et al. 2006; Seitz y Blokesh 2012), considerando la competencia como una estrategia evolutiva que permite acelerar la evolución y adaptación de transformantes naturales frente a cambios en el nicho alcanzando rápidamente el *fitness* por incorporación de secuencias homólogas o heterólogas (Jorth y Whiteley 2012). En conjunto, el mayor contenido en sistemas CRISPR-Cas y proporción de eventos de menos de 10 kb entre las especies competentes consideradas en este estudio apoya las observaciones de Jorth y colaboradores en 2012 validando la hipótesis planteada.

4. Relevancia de la recombinación homóloga en la estructura y evolución de las poblaciones.

El impacto de la recombinación homóloga sobre la estructura poblacional procariota y la evolución de los genomas *core* han sido ampliamente discutido (Whitaker *et al*., 2005; Cohan *et al*., 2006; Fraser *et al*., 2007, 2009; Polz *et al*., 2013; Polz y Hanage 2013). Entre los diferentes

modelos evolutivos y ecológicos existentes, el modelo neutral (Fraser *et al*., 2007, 2009) y la teoría de ecotipos (Cohan *et al*., 2006) son los más aceptados y aplicables a situaciones generales (Fraser *et al*., 2009; Vos *et al*., 2009) (véase introducción, apartado 1.2.2). Con el objetivo de identificar aquellas especies con patrones evolutivos que se ajusten a las especificaciones de estos modelos y por tanto evaluar el efecto evolutivo de la recombinación homóloga sobre los genomas *core,* analizamos tanto el impacto relativo que ejercen sobre éstos los mecanismos de mutación y recombinación como el efecto sobre los niveles de ANI entre las cepas de una misma especie. Como medidas del impacto de la recombinación empleamos los valore r/m y rho/theta,, mientras que los niveles de identidad se evaluaron con el ANI global y proporción afectada de genoma *core*. Los valores de $\log_{10}$ de las tasas r/m y rho/theta mostraron una elevada correlación (**figura 3.18**), correspondiendo los valores logarítmicos positivos con valores de r/m >1(marcados en azul). Estos valores superan con creces los que delimitarían el carácter clonal o sexual de una especie (Fraser 2009).

La comparación del porcentaje de genoma sinténico alineable de los genomas *core* frente a las medidas de ANI global 2 a 2 (**figura 3.18**), muestra una caída importante de ambos valores

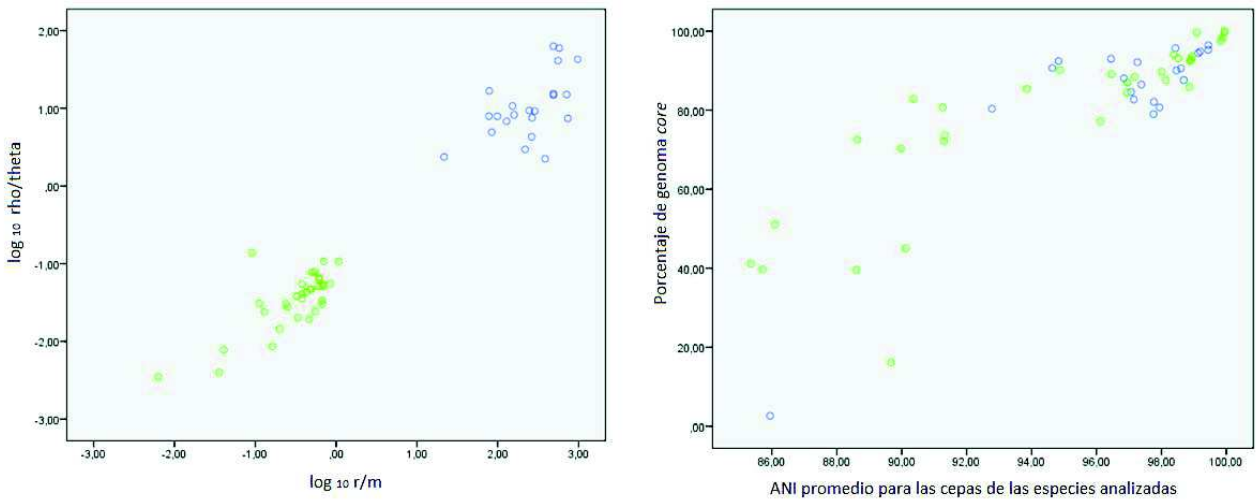


Figura C3.18. Correlación de los parámetros rho/theta y r/m para las 54 especies incluidas en este estudio (figura izquierda). Aparecen marcadas en azul los valores de r/m superiores a 1 y en verde los inferiores a este valor. La figura de la derecha representa la relación entre la proporción del genoma que forma parte del *core* y su identidad (ANI). Los colorea azul y verde muestran la distribución de especies según el nivel de r/m apreciándose un agrupamiento de aquellas que presentan un r/m superior a 1.

por debajo del 95% de identidad. Este valor corresponde con los valores umbral observados en estudios metagenómicos realizados en *P. marinus* (Coleman *et al*., 2006) o *H. walbyi* (Cuadros-Orellana *et al*., 2007; Tully *et al*., 2015) que presentaron una estructura poblacional en *clusters* con identidades de secuencia por encima de este umbral. Estudios de genómica comparativa entre cepas de una misma especie identificaron estos mismos *clusters* identificando diferencias en un rango de ANI entre el 1% y el 5% (Konstantinidis y DeLong 2008; Papke *et al*., 2004, 2007), atribuyendo un papel predominante a los procesos de recombinación homóloga en el mantenimiento de los mismos. Además esta identidad del 95% corresponde con una hibridación DNA-DNA entre cepas del 70%, por encima del cual dos cepas se consideran de la misma especie (Konstantinidis y Tidje 2005; Konstantinidis *et al*., 2006; Goris *et al*., 2007; Caro-Quintero y Konstantinidis 2011) y bajo el cual decrecen rápidamente los niveles de recombinación homóloga.

La distribución de puntos muestra valores de ANI mayores del 95% para la mayoría de especies con r/m superiores a 1 (en azul) (**figura 3.18**), por debajo del cual se aprecia una caída significativa del ANI y de la proporción de regiones sinténicas compartidas entre cepa de una misma especie. El resto de especies (en verde) con un ANI por encima de este punto de inflexión presentaron valores de r/m superiores a 0,33. Todos estos valores de r/m estarían situados por encima de 0,25, valor mínimo considerado en el modelo neutral para recombinación actúe como fuerza cohesiva sobre los *clusters* emergentes en la población, evitando su divergencia y reabsorbiéndolos en el *cluste*r parental (Fraser *et al*., 2007; 2009). Por debajo de este umbral de 0,25 tendríamos especies clonales, donde la diferenciación en distintos *clusters* genéticos se daría incluso en la ausencia de selección, y por encima de éste especies sexuales, que requerirían de periodos de diferenciación alopátrica o ecológica para reducir la tasa de recombinación entre *clusters* en procesos de especiación para promover la divergencia de los mismos (Fraser *et al*., 2009). En caso de alcanzar el **punto de especiación**, la fusión de dos *clusters* ya no sería posible. El proceso de fusión, considerado mucho más rápido que el de divergencia (Fraser *et al*., 2009), se está produciendo entre las *especies C. jejuni y C.coli* (Sheppard *et al*., 2008), entre las que se han observado niveles de recombinación homóloga relevantes (Caro-Quintero *et al*., 2009).

En conjunto los datos anteriores apoyan el efecto cohesivo de la recombinación homóloga entre *clusters* de una misma especie. La comparación de los valores de rRNA16S 2 a 2 entre las cepas de las especies con valores elevados de r/m, rho/theta, en azul en la **figura 3.18**, y la del resto de especies descartó que los resultados observados se debieran a un sesgo filogenético, por ejemplo por el muestreo de cepas de un mismo linaje o *cluster* dentro de determinada especie.

Si la recombinación homóloga actúa manteniendo la cohesión los genomas *core* y evitando la divergencia de *clusters* emergentes dentro de la población, debería de poder apreciarse su efecto no sólo en los valores de ANI intraespecíficos sino en los niveles de dN/dS de los *core* genoma. Aunque no se observó una clara correlación entre los valores de rho/theta o r/m y los promedios de dN/dS para cada una de las especies, sí se identificaron diferencias significativas en relación con la especialización ecológica, ya que las medias de los valores de dN/dS en el caso de patógenos obligados y organismos simbiontes y e intracelulares fueron casi el doble que las obtenidas en organismos de vida libre, comensales y patógenos oportunistas (**figura 3.19**). Este patrón se asemejó al obtenido al comparar porcentajes de genoma recombinados (**figura 3.6**). Como se ha discutido previamente, eventos de recombinación recientes, como sería el caso de la CR de *S. ruber*, hacen disminuir los valores de dN/dS y ejercen un efecto marcado sobre la divergencia alélica (Castillo-Ramírez et al., 2011). Estas evidencias sugieren que la recombinación homóloga podría ser uno de los factores que afecten a estos valores de dN/dS genómicos, ejerciendo un efecto importante genomas *core* y su evolución. Sin embargo han de tenerse en cuenta otros muchos factores como el tamaño poblacional ya que tamaños efectivos poblacionales mayores implican mayor presión selectiva y eficiencia de la selección natural, y por lo tanto valores menores de dN/dS (Hendrick, 2000). Especies patógenas oportunistas y comensales, con **poblaciones efectivas** grandes y elevada presencia de recombinación homóloga como *E. coli* (Mau *et al*., 2006), *N. meningitidis* o *S. aureus* muestran valores menores a los de patógenas obligadas como *C. pneumoniae* (Jordan *et al*., 2002). Incluso dentro de un mismo género como *Francisella*, la especie *F. novicida*, de vida libre con un 20% de su genoma afectado por recombinación homóloga, presenta valores de

dN/dS muy inferiores a los de *F. tularensis*, patógeno intracelular con población efectiva reducida, reflejando su efecto significativo (Larsson *et al*., 2006).

Una vez evaluado el efecto de la recombinación homóloga sobre los genomas *core*, nos planteamos si la adaptación de los genomas accesorios se da a la vez que la de los genomas *core* de la especie, es decir, si aquellos genomas con un genoma accesorio mayor y variable además presentan gran divergencia en su genoma *core*. Para responder a esta cuestión comparamos las niveles de identidad de los genomas *core* con el tamaño de los accesorios y el porcentaje de genes sujetos a intercambio interespecífico. También observamos si las relaciones dependían de algún modo de a la estrategia de vida, cuestión planteada anteriormente (Polz y Hanage 2013) y comparamos los niveles promedio de dN/dS de los genomas *core* para cada una de las especies y clases de especialización ecológicas. No se apreció ninguna correlación clara entre las variables de homología del genoma *core* (dN/dS) y la transferencia interespecífica (%HGT) (**figura 3.19**).

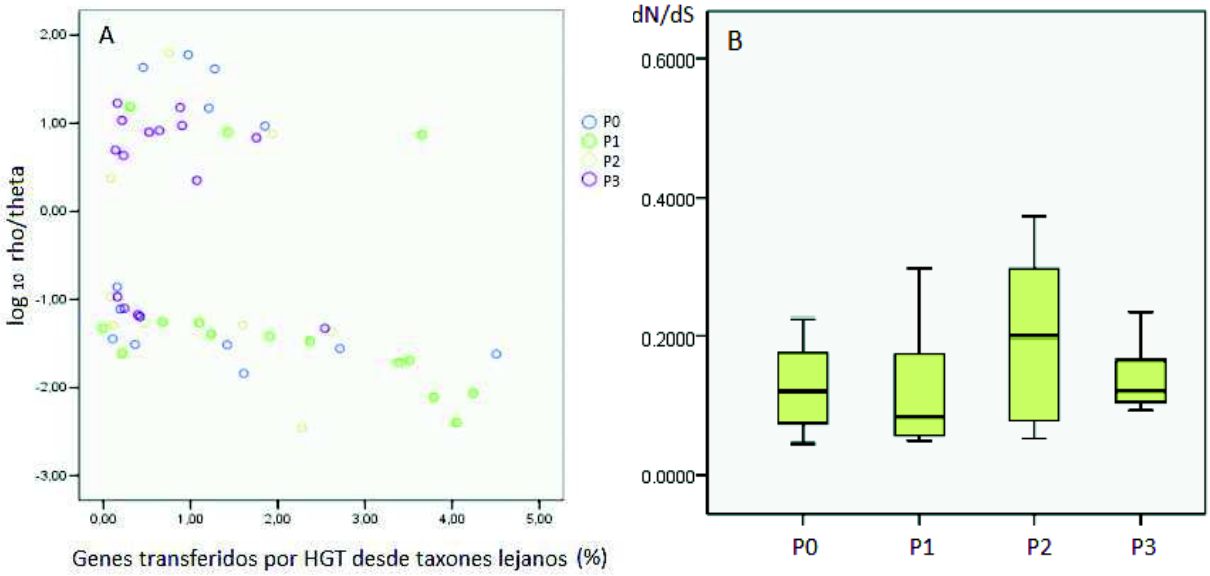


Figura C3.19. **Figura A:** Correlación de los parámetros rho/theta y porcentaje de genes incorporados por transferencia interespecífica (HGT) para las 54 especies incluidas en este estudio. **Figura B**: Diagrama de caja en el que se representan los promedios de dN/dS para las cuatro estrategias ecológicas: distribuidas en las 4 estrategias ecológicas: Simbiontes/patógenos intracelulares (P0), no patógenos (comensales y vida libre) (P1), patógenos obligados (P2) y patógenos oportunistas (P3).

Sin embargo al comparar el impacto ejercido por la recombinación homóloga si se pudo observar una agrupación de 5 especies de vida libre, *Desulfovibrio vulgaris*, *Dickeya dadantii*, *Zymomonas mobilis*, *P. fluorescens* y *P. marinus*, con valores bajos derho/theta y r/m con valores de %HGT superiores al 3% (**figura 3.19**). A excepción de *P. marinus*, con un valor de r/m de 0,47, el resto de cepas presentaron valores inferiores a 0,25- 0,5 que según el modelo neutral marcan el umbral entre una especie con comportamiento clonal y sexual. En el caso de estas especies la transferencia horizontal interespecífica podría jugar un papel relevante debido al patrón clonal predominante del genoma *core*. La distribución del contenido en genes de HGT interespecífico explicó un 8% de la varianza de la fracción de genoma recombinado, incrementando la explicada sólo por la estrategia ecológica a un junto a esta hasta un 26% (r^2= 0,26; $p < 0,05$, ANCOVA, modelo lineal).

5. Modelo del efecto de los factores analizados sobre la recombinación homóloga.

A lo largo de este capítulo se han ido comentando múltiples factores que afectan a los niveles de recombinación homóloga, ya sea de manera directa, como es el caso de la estrategia ecológica y algunas de las variables clasificadas como de "movilidad" o "barrera". Los modelos lineales con dos variables mostraron que variables como la estrategia ecológica en combinación con otras como el contenido en genes relacionados con la capacidad competente y genes trasferidos entre especies lejanas explican más del 30% de la variabilidad observada en las variables relativas a la recombinación homóloga. Además de estas relaciones aparentemente más directas, existen otras relacionadas con estas últimas cuyos efectos sobre las tasas de recombinación aunque no sean tan directos, pueden afectar a la capacidad de recombinación de un organismo así como a la configuración de su genoma. Realizamos un *path analysis* con el objetivo profundizar en las interacciones directas e indirectas que ejercen todas las variables discutidas, más allá de las interacciones que puedan explorar correlaciones parciales o modelos lineales, y en última instancia sobre los niveles de recombinación de cada especie. La comparación de las variables entre si implicó su transformación de estas a sus correspondientes matrices de distancia. Se compararon estas matrices entre sí, y se calcularon correlaciones entre

todas ellas (**figura 3.20**). En una segunda etapa se generaron las matrices consenso más generales englobando las de varias variables individuales, obteniendo una matriz general para la estrategia ecológica, filogenia, variables genómicas, movilidad-barrera y recombinación. Esta última incluyó porcentaje de genoma recombinado, eventos por cepa y valores r/m y rho/theta). La comparación de todas las matrices de distancia por medio de un test de Mantel permitió establecer las correlaciones parciales directa e indirectas existentes entre casa una de estas matrices y si los niveles apreciados resultan o no significativos.

La **figura 3.20** muestra el modelo conceptual construido, donde cada recuadro representa una de las matrices generales obtenida del consenso de las matrices de las variables individuales que la componen. Además, las tablas anexas en color azul que acompañan muestran los valores de correlación de las matrices para las variables individuales. Aunque, como ya se comentó anteriormente, la filogenia no afectó de manera significativa y directa a la recombinación

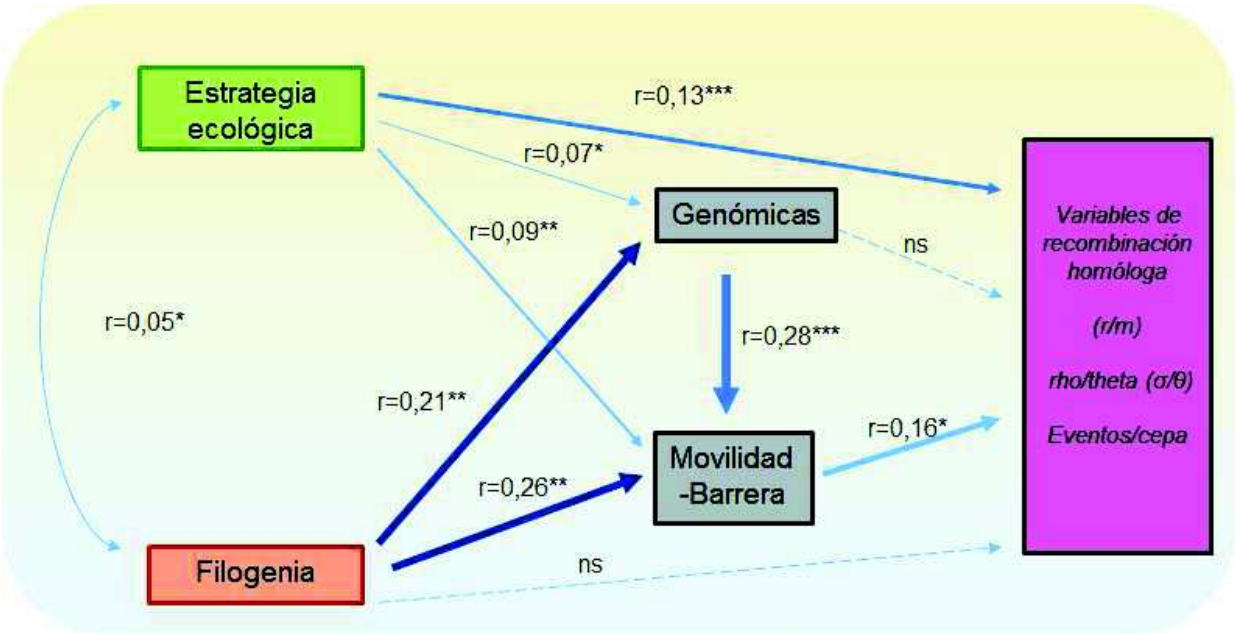
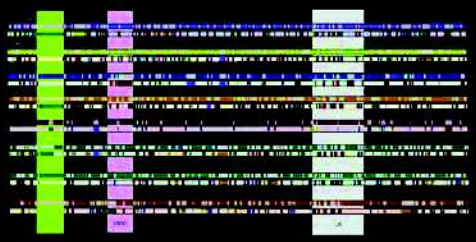


Figura C3.20. Esquema del modelo de *path analysis* propuesto para el análisis de la influencia de los factores estrategia ecológica, filogenia, variables movilidad- barrera y características de los genomas sobre la distribución de os genomas de las 54 especies analizadas. Se muestras los valores de r significativos (niveles <0,05*, 0,01**; 0,001*) para las comparaciones parciales llevadas a cabo durante es test de Mantel en el que se compararon las matrices de similitud compuestas de las variables mencionadas.

homóloga, si influye de manera decisiva en las variables de movilidad-barrera y las características genómicas de las diferentes especies. La estrategia de vida fue la variable que más influyó de manera directa junto con el conjunto de los elementos barrera y movilidad como ya apuntaban los modelos lineales.

En conjunto los resultados mostrados en este capítulo destacan el fuerte impacto de la recombinación homóloga en los procesos adaptativos de especies con estrategias de vida muy diversas y su contribución a la evolución de los genomas *core* y la cohesión de *clusters* poblacionales.



1. Pequeñas diferencias genómicas entre las cepas M8y M31 de *S. ruber* afectan notablemente a los transcriptomas individuales de cada una de ellas y a su respuesta en presencia de la otra en cultivo mixto.

2. Las cepas M8 y M31 de *S. ruber* expresaron el 98% de sus genes en cultivo puro. La mayoría de diferencias entre transcriptomas en estas condiciones se dieron a nivel de genes relacionados con respuestas a estímulos ambientales (transportadores de membrana, sistemas de dos componentes y proteínas flagelares).

3. La suma de los transcriptomas individuales de las cepas M8 y M31 de *S. ruber* no equivale al resultante en cultivo mixto, lo que demuestra que en presencia de cepas muy cercanas con las que cohexisten en la naturaleza éstas responden modificando sus actividades celulares.

4. La respuesta derivada de la interacción de M8 y M31 fue específica de cepa. El genoma *core* mostró más cambios que el accesorio y la respuesta de cada cepa fue específica al compartir sólo un 25% de los que presentaron expresión diferencial entre cultivo puro y mixto. En M31 el incremento de expresión se dio en genes relacionados con el crecimiento exponencial y en M8 con respuesta a condiciones de estrés.

5. Cambios de expresión en barteriocinas y genes implicados en la síntesis de moléculas señales, llevan a pensar que estas moléculas podrían actuar como mediadores en la interacción intraespecífica de cepas de *S. ruber*.

6. El análisis de los genomas de 8 cepas de *S. ruber* indica que genoma *core* estaría constituido por 2434 genes, representando en promedio un 73,5% de los genes de las cepas estudiadas. *S. ruber* presenta un pangenoma abierto y cada una de las cepas mostró un contenido plasmídico distinto.

7. Los genomas de las cepas de aislados de *S. ruber* presentan características genómicas

comunes, detectándose una elevada sintenia y la presencia conservada de las HRVs I y II. Las principales diferencias genómicas se concentraron a nivel de las propias HRVs, de GIs e *indels* específicos de cepa y contenido plasmídico. Estos elementos accesorios están enriquecidos en elementos transponibles, proteínas hipotéticas y genes pertenecientes a la categoría COGM. Además encontramos un contenido heterogéneo de sistemas RM y CRISPR-Cas. Estos datos confirman la gran microdiversidad presente a nivel de envolturas celulares y sistemas barrera de entrada de DNA, posiblemente en respuesta a la presión vírica.

8. La recombinación homóloga a nivel de clusters homólogos dentro de las fGI, recombinación homóloga mediada por XerD y la integración de plásmidos fueron los principales mecanismos de acción sobre los genomas accesorios.

9. La recombinación homóloga es el principal mecanismo evolutivo que actúa sobre los genomas *core* de *S. ruber* afectando a más del 30% del genoma, rompiendo la estructura clonal y evitando la divergencia de *clusters*.

10. En *S. ruber*, los genes de las categorías traducción, estructuras ribosómicas y biogénesis (COG J) y replicación, recombinación y reparación del DNA (COG L) serían los que más recombinan homólogamente en *S. ruber*, entre ellos una elevada proporción de elementos transponibles y genes codificantes de XerD, que podrían estar mediando estos intercambios. Más del 70% de los genes recombinados de la categoría COG V estuvieron relacionados con sistemas de transporte multidroga o resistencia a antibióticos. Estos últimos podrían estar involucrados en procesos de comunicación celular.

11. El análisis del efecto de la recombinación homóloga en 54 especies procariotas la sitúa como un mecanismo evolutivo relevante que actuaría prevalentemente sobre genomas *core* de organismos de vida libre, comensales y patógenos oportunistas. Aunque la estrategia ecológica es el factor que mejor explica la distribución de la recombinación en las distintas especies, factores genómicos como la distribución de genes de reparación y replicación, densidad de

secuencias Chi, distribución de sistemas RM y CRISPR-Cas o la capacidad competente explicaron una parte importante de la variabilidad observada. Los mecanismos de competencia estarían implicados principalmente en la transferencia de eventos de menor tamaño.

12. Existe una relación funcional entre el contenido de las regiones recombinadas y la presión adaptativa del ambiente asociado a las cuatro estrategias de vida estudiadas (simbiontes y patógenos intracelulares, patógenos obligados, organismos de vida libre y patógenos oportunistas), en muchos casos relacionados con mecanismos de evasión, factores de virulencia o resistencia a antibióticos.

13. Los patógenos obligados y oportunistas presentaron enriquecimientos en términos relativos a la categoría COG N implicados en la evasión de los mecanismos del sistema inmune. Tanto organismos de vida libre como patógenos obligados mostraron enriquecimientos en términos de la categoría COG V relacionados con resistencia a antibióticos.

14. El enriquecimiento en elementos transponibles y genes codificantes de XerD confirmó la implicación de estos mecanismos en los procesos de integración de DNA mediante recombinación homóloga.

15. La recombinación homóloga podría tener un efecto de cohesión sobre los *clusters* emergentes en aquellas especies con tasas de r/m superiores a 0,25 ya que en las mismas se aprecian valores de ANI superiores al 95% y valores inferiores de dN/dS que se ajustarían a las predicciones del modelo neutral de Fraser.

Introducción

Objetivos

Materiales y métodos

Resultados y discusión

Capítulo 1

Análisis de las diferencias transcripcionales e interacción de cepas cercanas de *S.ruber* mediante RNAseq.

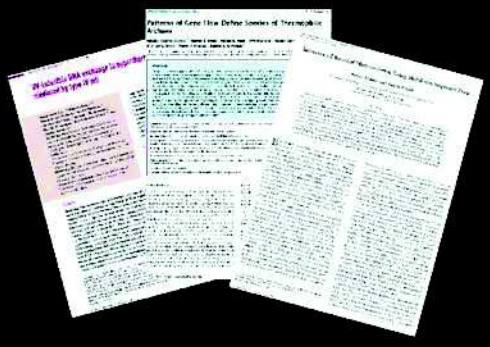
Capítulo 2

Estudio de los mecanismos y estrategias de diversificación genómica en *S. ruber*

Capítulo 3

Impacto de la recombinación homóloga sobre la evolución de genomas *core* procariotas

Conclusiones



Bibliografía

Anexos

Abby S, Daubin V. (2007). Comparative genomics and the evolution of prokaryotes. *TRENDS Microbiol*. **15:** 135–141.

Achtman M, Morelli G, Zhu P, Wirth T, Diehl I, Kusecek B, Vogler AJ, Wagner DM, Allender CJ, Easterday WR, Chenal-Francisque V, Worsham P, Thomson NR, Parkhill J, Lindler LE, Carniel E, Keim P. (2004). Microevolution and history of the plague bacillus, *Yersinia pestis*. Proc Natl. *Acad Sci U S A*. **101**:17837–17842.

Achtman M, Wagner M. (2008). Microbial diversity and the genetic nature of microbial species. Nat Rev Microbiol. **6**:431–40.

Achtman M, Wagner M. (2008). Microbial diversity and the genetic nature of microbial species. Nat *Rev Microbiol*. **6**:431–40.

Ajon M, Fröls S, van Wolferen M, Stoecker K, Teichmann D, Driessen AJM, Grogan DW, Albers SV, Schleper C. (2011). UV-inducible DNA exchange in hyperthermophilic archaea mediated by type IV pili. *Mol Microbiol*. **82**: 807–817.

Altschul, S.F, Gish, W, Miller, W, Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*. **215**: 403-410.

Ambur OH, Davidsen T, Frye S a, Balasingham S V, Lagesen K, Rognes T, Tønjum T. (2009). Genome dynamics in major bacterial pathogens. *FEMS Microbiol Rev.* **33**: 453–470.

Andam CP, Williams D, Gogarten JP. (2010). Biased gene transfer mimics patterns created through shared ancestry. Proc Natl Acad Sci U S A **107**:10679–10684.

Anders S, Huber W. (2010). Differential expression analysis for sequence count data. *Genome Biol*. **11**(10):R106.

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

Antón J, Llobet-Brossa E, Rodríguez-Valera F y Amann R. (1999). Fluorescence *in situ* hybridization analysis of the prokaryotic community inhabiting crystallizer ponds. *Environ Microbiol*. **1:** 517 – 523.

Antón J, Lucio M, Peña A, Cifuentes A, Brito-Echeverría J, Moritz F, Tziotis D, López C, Urdiain M, Schmitt-Kopplin P, Rosselló-Móra R. (2013). High Metabolomic Microdiversity within Co-Occurring Isolates of the Extremely Halophilic Bacterium *Salinibacter ruber*. *PLoS One* **8**:1–14.

Antón J, Oren A, Benlloch S, Rodríguez-Valera F, Amann R y Rosselló-Mora R. (2002). *Salinibacter ruber* gen. nov, sp, a novel, extremely halophilic member of the *Bacteria* from saltern crystallizer ponds. *Int J Syst Evol Microbiol.* **52:** 485 – 491.

Antón J, Peña A, Santos F, Martínez-García M, Schmitt-Kopplin P y Rosselló-Mora R. (2008). Distribution, abundance and diversity of the extremely halophilic bacterium *Salinibacter*

ruber. Saline Systems. **4:** 15.

Antón J, Peña A, Valens M, Santos F, Glöckner FO, Bauer M, Dopazo J, Herrero J, Rosselló-Mora R, Amann R. (2005). Salinibacter ruber: genomics and biogeography. In adaptation to Life at High Salt Concentrations in Archaea, Bacteria, and Eukarya, Gunde-Cimerman, N, Oren, A, Plemenitas, A, *eds. Springer, New York*, **pp**. 255-266.

Antón J, Rosselló-Mora R, Rodríguez-Valera F y Amann R. (2000). Extremely halophilic bacteria in crystallizer ponds from solar salterns. *Appl Environ Microbiol.* **7:** 3052 – 3057.

Arvand M, Feil EJ, Giladi M, Boulouis HJ, Viezens J. (2007). Multi-locus sequence typing of *Bartonella henselae* isolates from three continents reveals hypervirulent and feline-associated clones. *PLoS One* **2**:e1346.

Ashburner M, Ball C A, Blake J A. (2000). Gene Ontology: tool for the unification of biology. *Nat Genet.* **25**:25–29.

Aziz RK, Bartels D, Best A a, DeJongh M, Disz T, Edwards R a, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek R a, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. (2008). The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics.* **9**: 75.

Balashov SP, Imasheva ES, Wang JM, Boichenko VA, Anto J, Lanyi JK. (2007). Xanthorhodopsin : A Proton Pump with a Light-Harvesting Carotenoid Antenna **2061**.

Baldo L, Hotopp JCD, Jolley K a, Bordenstein SR, Biber S a, Choudhury RR, Hayashi C, Maiden MCJ, Tettelin H, Werren JH. (2006). Multilocus sequence typing system for the endosymbiont *Wolbachia pipientis. Appl Environ Microbiol* **72**:7098–7110.

Baldo L, Hotopp JCD, Jolley K A, Bordenstein SR, Biber S. A, Choudhury RR, Hayashi C, Maiden MCJ, Tettelin H, Werren JH. (2006). Multilocus sequence typing system for the endosymbiont Wolbachia pipientis. Appl Environ Microbiol **72**:7098–7110.

Baliga NS, Bonneau R, Facciotti MT, Pan M, Glusman G, Deutsch EW, Shannon P, Chiu Y, Weng RS, Gan RR, Hung P, Date S V, Marcotte E, Hood L, Ng WV. (2004). Genome sequence of. *Genome Res*. 2221–2234.

Balleza E, López-Bojorquez LN, Martínez-Antonio A, Resendis-Antonio O, Lozada-Chávez I, Balderas-Martínez YI, Encarnación S, Collado-Vides J. (2009). Regulation by transcription factors in bacteria: Beyond description. *FEMS Microbiol Rev.* **33**:133–151.

Bankevich A, Nurk S, Antipov D, Gurevich A a, Dvorkin M, Kulikov AS, et al.(2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol*. **19**(5):455–477.

Barkay T, Smets BF. (2005). Horizontal gene flow in mmicrobial communities. *ASM News*. **71**: 412 – 419.

Barrangou R, Fremaux C, Deveau H, Richards M, Patrick Boyaval, Moineau S, Romero D, Horvath P. (2007). CRISPRProvides Acquired Resistance Against Viruses in Prokaryotes. *Science*. **315**: 1709–1712.

Basic-Hammer N, Vogel V, Basset P, Blanc DS. (2010). Impact of recombination on genetic variability within Staphylococcus aureus clonal complexes. *Infect Genet Evol* **10**:1117–23.

Bayliss, C.D, Callaghan, M.J, and Moxon, E.R. (2006). High allelic diversity in the methyltransferase gene of a phase variable type III restriction-modification system has implications for the fitness of *Haemophilus influenzae*. *Nucleic Acids Res*. **34**: 4046–4059

Bellanger X, Payot S, Leblond-Bourget N, Guédon G. (2014). Conjugative and mobilizable genomic islands in bacteria: Evolution and diversity. *FEMS Microbiol Rev*. **38**:720–760.

Bellard E, Bertelsmeier C, Leadley P, Thuiller W, Courchamp F. (2014). *Europe PMC Funders Group. Ecol Lett.* **15**: 365–377..

Benítez-Páez A, Cárdenas-Brito S, Corredor M, Villarroya M, Armengod ME. (2014). Impairing methylations at ribosome RNA, a point mutation-dependent strategy for aminoglycoside resistance: the rsmG case. *Biomedica*.1:41-9.

Benlloch S, Martínez-Murcia AJ y Rodríguez-Valera F. (1995). Sequencing of bacterial and archaeal 16S rRNA genes directly amplified from a hypersaline environment. *Syst Appl Microbiol*. **18**: 574 – 581.

Bergqvist S, Williams M a, O'Brien R, Ladbury JE. (2003). Halophilic adaptation of protein-DNA interactions. *Biochem Soc Trans*. **31**:677–680.

Bernier SP, Surette MG. (2013). Concentration-dependent activity of antibiotics in natural environments. *Front Microbiol*. **4**:1–14.

Bernstein H, Byers GS, Michod RE. (1981). Evolution of sexual reproduction: importance of DNA repair, complementation, and variation. *Am Nat.* **117**: 537–549.

Bhaya, D, Grossman, A.R, Steunou, A.S, Khuri, N, Cohan, F.M, Hamamura, N, *et al*. (2007) Population level functional diversity in a microbial community revealed by comparative genomic and metagenomic analyses. *ISME J.* **1**: 703–713.

Bi D, Xu Z, Harrison EM, Tai C, Wei Y, He X, et al. (2014). ICEberg: a web-based resource for integrative and conjugative elements found in Bacteria. *Nucleic Acids Res*. D621–626.

Bickle T a, Krüger DH. (1993). Biology of DNA restriction. *Microbiol Rev.* **57**: 434–450.

Bisharat N, Cohen DI, Maiden MC, Crook DW, Peto T, Harding RM. (2007). The evolution of genetic structure in the marine pathogen, Vibrio vulnificus. *Infect Genet Evol* **7**:685–93.

Bisharat N, Cohen DI, Maiden MC, Crook DW, Peto T, Harding RM. (2007). The evolution of genetic structure in the marine pathogen, Vibrio vulnificus. *Infect Genet Evol* **7**:685–93.

Blakely G, Sherratt D. (1996). Determinants of selectivity in Xer site-specific recombination. Genes Dev **10**: 762–773.

Bloomquist EW, Dorman KS, Suchard MA. (2009) StepBrothers: inferringpartially shared ancestries among recombinant viral sequences. *Biostatistics*. **10**: 106–120.

Bolhuis H, te Poele EM y Rodríguez-Valera F. (2004). Isolation and cultivation of Walsby´s square archaeon. *Environ Microbiol*. **6:** 1287 – 1291.

Boni MF, Posada D, Feldman MW. (2007) An exact nonparametric method for inferring mosaic structure in sequence triplets. *Genetics*. **176**:1035–1047.

Boucher Y, Douady CJ, Sharma AK, Kamekura M, Doolittle WF. 2007. Intragenomic heterogeneity and intergenomic recombination among *Vibrio parahaemolyticus* 16S rRNA genes. *Microbiology.* **153**: 2640–7.

Boujelben I, Gomariz M, Martínez-García M, Santos F, Peña A, López C, Antón J, Maalej S. (2012). Spatial and seasonal prokaryotic community dynamics in ponds of increasing salinity of Sfax solar saltern in Tunisia. *Antonie Van Leeuwenhoek*. **101**: 845-857.

Boyaval P, Moineau S, Romero D a, Horvath P. (2007). Against viruses in *Prokaryotes*. *Science*. **315**: 1709–1712.

Breitbart, M, and Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? *Trends Microbiol*. **13**: 278–284.

Breuert S, Thorsten Allers, Gabi Spohn, Jorg Soppa. (2006)**.** Regulated polyploidy in halophilic archaea. *Plos One.* **1**(1): e92.

Brigulla, M, and Wackernagel, W. (2010). Molecular aspects of gene transfer and foreign DNA acquisition in prokaryotes with regard to safety issues. Appl. Microbiol. Biotechnol. **86**: 1027–1041.

Brochet M, Couve E, Glaser P, Guedon G, Payot S. (2008). Integrative Conjugative Elements and related elements are major contributors to the genome diversity of *Streptococcus agalactiae. J Bacteriol* **190**: 6913–6917

Budroni, S, Siena, E, Dunning Hotopp, J.C, Seib, K.L, Serruto, D, Nofroni, C, Comanducci, M, Riley, D.R, Daugherty, S.C, Angiuoli, S.V, *et al*. (2011). *Neisseria meningitidis* is structured in clades associated with restriction modification systems that modulate homologous recombination. Proc. *Natl. Acad. Sci. U.S.A.* 108: 4494–4499.

Buffet J-P, Pisanu B, Brisse S, Roussel S, Félix B, Halos L, Chapuis J-L, Vayssier-Taussat M. (2013. Deciphering bartonella diversity, recombination, and host specificity in a rodent community. *PLoS One*. **8**:e68956.

Burrus V, Pavlovic G, Decaris B, Guédon G. (2002). The ICESt1 element of Streptococcus thermophilus belongs to a large family of integrative and conjugative elements that exchange modules and change their specificity of integration. *Plasmid*. **48**:77–97.

Burrus, V, Pavlovic, G, Decaris, B, and Guedon, G. (2002). Conjugative transposons: the tip of the iceberg. *Mol. Microbiol*. **46:** 601–610.

Cadillo-Quiroz H, Didelot X, Held NL, Herrera A, Darling A, Reno ML, Krause DJ, Whitaker RJ. (2012). Patterns of gene flow define species of thermophilic Archaea. PLoS Biol **10**.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T.(2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. **15**: 1972–1973.

Carbone A, Zinovyev A, Képès F. (2003). Codon adaptation index as a measure of dominating codon bias. *Bioinformatics*. **19**:2005–2015.

Caroff M, Karibian D. (2003). Structure of bacterial lipopolysaccharides. *Carbohydr Res* **338**:2431–2447.

Caro-Quintero A, Deng J, Auchtung J, Brettar I, Höfle MG, Klappenbach J, Konstantinidis KT. (2011). Unprecedented levels of horizontal gene transfer among spatially co-occurring Shewanella bacteria from the Baltic Sea. *ISME J* **5**: 131–140.

Caro-Quintero A, Konstantinidis KT. (2012). Bacterial species may exist, metagenomics reveal. *Environ Microbiol* **14**: 347–355.

Caro-Quintero A, Konstantinidis KT. (2012). Bacterial species may exist, metagenomics reveal. *Environ Microbiol* **14**:347–355.

Carver T, Harris SR, Berriman M, Parkhill J, McQuillan J A. (2012). Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*. **28**:464–469.

Casadesús J, Low D. (2006). Epigenetic gene regulation in the bacterial world. *Microbiol Mol Biol Rev* **70**:830–856.

Caspi, R, Pacek, M, Consiglieri, G, Helinski, D.R, Toukdarian, A, and Konieczny, I. (2001) A broad host range replicon with different requirements for replication initiation in three bacterial species. *EMBO J* 20: 3262–3271

Castillo J A, Greenberg JT. (2007). Evolutionary dynamics of Ralstonia solanacearum. *Appl Environ Microbiol* **73**:1225–1238.

Castillo-Ramírez S, Harris SR, Holden MTG, He M, Parkhill J, Bentley SD, Feil EJ.(2011). The impact of recombination on dN/dS within recently emerged bacterial clones. *PLoS Pathog* **7**: e1002129.

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

Chaisson MJ, Brinza D, Pevzner P A. (2009). De novo fragment assembly with short mate-paired reads: Does the read length matter? *Genome Res* **19**:336–346.

Chédin F, Noirot P, Biaudet V, Ehrlich SD. (1998). A five-nucleotide sequence protects DNA from exonucleolytic degradation by AddAB, the RecBCD analogue of Bacillus subtilis. *Mol Microbiol***. 29**:1369–77.

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Müller WEG, Wetter T, Suhai S. (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res* **14**:1147–1159.

Chu T-C, Lu C-H, Liu T, Lee GC, Li W-H, Shih AC-C. (2013). Assembler for de novo assembly of large genomes. *Proc Natl Acad Sci U S A* . 110:E3417–24.

Chugani S, Kim BS, Phattarasukol S, Brittnacher MJ, Choi SH, Harwood CS, Greenberg EP. (2012). PNAS Plus: Strain-dependent diversity in the Pseudomonas aeruginosa quorum-sensing regulon. *Proc Natl Acad Sci*. **109**:E2823–E2831.

Claverys J-P, Prudhomme M, Martin B. (2006). Induction of Competence Regulons as a General Response to Stress in Gram-Positive Bacteria. *Annu Rev Microbiol* **60**:451–475.

Cohan FM. (2002). Sexual isolation and speciation in bacteria. *Genetica*.**116** (2-3):359-70.

Cohan FM. (2006). Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Philos Trans R Soc Lond B Biol Sci* **361**: 1985–1996.

Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, Delong EF, Chisholm SW. (2006). Genomic islands and the ecology and evolution of Prochlorococcus. *Science* **311**: 1768–1770.

Conesa A, Götz S. (2008). Blast2GO: A comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics*. 619832.

Cordero OX, Wildschutte H, Kirkup B, Proehl S, Ngo L, Hussain F, Le Roux F, Mincer T, Polz MF. (2012). Ecological Populations of Bacteria Act as Socially Cohesive Units of Antibiotic Production and Resistance. *Science*. (80-) **337**:1228–1231.

Corvaglia a. R, Francois P, Hernandez D, Perron K, Linder P, Schrenzel J. (2010). A type III-like restriction endonuclease functions as a major barrier to horizontal gene transfer in clinical Staphylococcus aureus strains. *Proc Natl Acad Sci* **107**: 11954–11958.

Coscollá M, González-Candelas F. (2007). Population structure and recombination in environmental isolates of *Legionella pneumophila. Environ Microbiol*. **9**:643–56.

Croucher NJ, Thomson NR. (2010). Studying bacterial transcriptomes using RNA-seq. *Curr Opin Microbiol* **13**: 619–624.

Cuadros-Orellana S, Martin-Cuadrado A-B, Legault B, D'Auria G, Zhaxybayeva O, Papke RT, Rodriguez-Valera F. (2007). Genomic plasticity in prokaryotes: the case of the square haloarchaeon. *ISME J*. **1**: 235–245.

D'Auria G, Jiménez-Hernández N, Peris-Bondia F, Moya A, Latorre A. (2010). Legionella pneumophila pangenome reveals strain-specific virulence factors. *BMC Genomics*. **11**: 181.

Dan T, Liu W, Song Y, Xu H, Menghe B, Zhang H, Sun Z. (2015). The evolution and population structure of Lactobacillus fermentum from different naturally fermented products as determined by multilocus sequence typing (MLST*). BMC Microbiol* **15**:107.

Darling ACE, Mau B, Blattner FR, Perna NT. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res*.**14**: 1394–1403.

Darling AE, Mau B, Perna NT. (2010). progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*.5(6):e11147.

Darling AE, Tritt A, Eisen J a, Facciotti MT. (2011). Mauve assembly metrics. *Bioinformatics*. **27**(19): 2756–2757.

Darren P. Martin, Lemey P, Lott M, Moulton V., Posada D. and Lefeuvre P. (2011). Analysing recombination in nucleotide sequences. *Molecular Ecology Resources*. **11**: 943–955.

Darren P. Martin, Philippe Lemey, Martin Lott, Vincent Moulton, David Posada and Pierre Lefeuvre.(2010). RDP3: a flexible and fast computer program for analyzing recombination. **26**(19): 2462–2463.

Das B, Martínez E, Midonet C, Barre FX. (2013). Integrative mobile elements exploiting Xer recombination. *Trends Microbiol*. **21**:23–30.

Davies J. (2006). Are antibiotics naturally antibiotics? *J Ind Microbiol Biotechnol* **33**:496–499.

Davis A, Chen D. (2013). DNA double strand break repair via non-homologous end-joining. *Transl Cancer Res*. **2**:130–43.

de Vries, J, and Wackernagel, W. (2002). Integration of foreign DNA during natural transformation of *Acinetobacter* sp. by homology-facilitated illegitimate recombination. *Proc. Natl. Acad. Sci. U.S.A*. **99**: 2094–2099.

Dehal PS, Joachimiak MP, Price MN, Bates JT, Baumohl JK, Chivian D, Friedland GD, Huang KH, Keller K, Novichkov PS, Dubchak IL, Alm EJ, Arkin a. P. (2010).

MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Res* **38**: 396–400.

Delcher a L, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL. (1999). Alignment of whole genomes. *Nucleic Acids Res*. 27(11):2369–2376.

den Bakker HC, Cummings C a, Ferreira V, Vatta P, Orsi RH, Degoricija L, Barker M, Petrauskene O, Furtado MR, Wiedmann M. (2010). Comparative genomics of the bacterial genus Listeria: Genome evolution is characterized by limited gene acquisition and limited gene loss. *BMC Genomics*. **11**: 688.

den Bakker HC, Didelot X, Fortes ED, Nightingale KK, Wiedmann M. (2008). Lineage specific recombination rates and microevolution in *Listeria monocytogenes*. BMC Evol Biol **8**:277.

Denef VJ, VerBerkmoes NC, Shah MB, Abraham P, Lefsrud M, Hettich RL, Banfield JF. (2009). Proteomics-inferred genome typing (PIGT) demonstrates inter-population recombination as a strategy for environmental adaptation. *Environ Microbiol*. **11:**313–325

Diancourt L, Passet V, Chervaux C, Garault P, Smokvina T, Brisse S. (2007). Multilocus sequence typing of Lactobacillus casei reveals a clonal population structure with low levels of homologous recombination. *Appl Environ Microbiol***. 73**:6601–11.

Diancourt L, Passet V, Chervaux C, Garault P, Smokvina T, Brisse S. (2007). Multilocus sequence typing of *Lactobacillus casei* reveals a clonal population structure with low levels of homologous recombination. *Appl Environ Microbiol.* **73**:6601–11.

Diancourt L, Passet V, Verhoef J, Patrick a D, Grimont P a D, Brisse S. (2005). Multilocus Sequence Typing *of Klebsiella pneumoniae* Nosocomial Isolates Multilocus Sequence Typing of Klebsiella pneumoniae Nosocomial Isolates. *J Clin Microbiol***. 43**:4178–4182.

Didelot X, C.J. Maiden (2010). Impact of recombination on bacterial evolution (2010) *Trends in microbiol*. **18**: 315-322.

Didelot X, Falush D. (2007). Inference of bacterial microevolution using multilocus sequence data. *Genetics*. **175**(3):1251-1266.

Didelot X, Lawson D, Darling A, Falush D. (2010). Inference of homologous recombination in bacteria using whole-genome sequences. *Genetics*. **186**(4): 1435–1449.

Dieffenbach CW, Lowe TMJ, Dveksler GS. (1995). General Concepts for PCR Primer Design. In: PCR Primer, A Laboratory Manual. Dieffenbach CW, Dveksler GS. *Ed, Cold Spring Harbor Laboratory Press*. 133-155.

Dillingham MS, Kowalczykowski SC. (2008). RecBCD Enzyme and the Repair of Double-Stranded DNA Breaks. *Microbiol Mol Biol Rev* **72**:642–671.

Dmowski M, Jagura-Burdzy G. (2013). Active stable maintenance functions in low copy-number plasmids of Gram-positive bacteria I. partition systems. *Polish J Microbiol* **62**: 3–16.

Do T, Jolley K a, Maiden MCJ, Gilbert SC, Clark D, Wade WG, Beighton D. (2009). Population structure of *Streptococcus oralis*. Microbiology **155**:2593–2602.

Dobrindt U, Hochhut B, Hentschel U, Hacker J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol*. **2**: 414–24.

Doi Y, Arakawa Y. (2007). 16S Ribosomal RNA Methylation: Emerging resistance mechanism against aminoglycosides. *Clin Infect Dis* **45**:88–94.

Domínguez NM, Hackett KT, Dillard JP. (2011). XerCD-mediated site-specific recombination leads to loss of the 57-kilobase gonococcal genetic island. *J Bacteriol* **193**: 377–88.

Dominy BN, Perl D, Schmid FX, Brooks CL. (2002). The effects of ionic strength on protein stability: The cold shock protein family. *J Mol Biol*. **319**:541–554.

Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli S V, Oggioni M, Dunning Hotopp JC, Hu FZ, Riley DR, Covacci A, Mitchell TJ, Bentley SD, Kilian M, Ehrlich GD, Rappuoli R, Moxon ER, Masignani V. (2010). Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. *Genome Biol* **11**:R107.

Doolittle WF, Papke RT. (2006). Genomics and the bacterial species problem. Genome Biol **7**:116.

Dorer MS, Fero J & Salama NR. (2010) DNA damage triggers genetic exchange in *Helicobacter pylori*. PLoS Pathog 6: e1001026.

Dutta C, Pan A. (2002). Horizontal gene transfer and bacterial diversity. *J Biosci*. **27:** 27–33.

Edgar RC (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. **32**(5): 1792-1797.

El Karoui M, Biaudet V, Schbath S, Gruss a. (1999). Characteristics of Chi distribution on different bacterial genomes. *Res Microbiol* **150**:579–87.

El-Adawy H, Hotzel H, Tomaso H, Neubauer H, Taboada EN, Ehricht R, Hafez HM. (2013). Detection of Genetic Diversity in *Campylobacter jejuni* Isolated from a Commercial Turkey Flock Using flaA Typing, MLST Analysis and Microarray Assay. PLoS One **8**:1–11.

Emerson JB, Andrade K, Thomas BC, Norman A, Allen EE, Heidelberg KB, Banfield JF. (2013). Virus-host and CRISPR dynamics in *Archaea*-dominated hypersaline Lake Tyrrell, Victoria, Australia. *Archaea* **2013**: 370871.

Enersen M. (2011). *Porphyromonas gingivalis*: a clonal pathogen? *J Oral Microbiol* **3**:1–11.

Enright MC, Day NP, Davies CE, Peacock SJ, Spratt BG. (2000). Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. *J Clin Microbiol* **38**:1008–15.

Enright MC, Spratt BG, Kalia A, John H, Bessen DE, Cross JH. (2001). Multilocus Sequence Typing of *Streptococcus* pyogenes and the relationships between emm type and clone Multilocus Sequence Typing of *Streptococcus pyogenes* and the Relationships between emm Type and Clone. *Infect Immun* **69**:2416–2427.

Even S, Charlier C, Nouaille S, Ben Zakour N L, Cretenet M, Cousin F, Gautier M, Cocaign-Bousquet M, Pascal Loubière P ; Le Loir Y. (2009). *Staphylococcus aureus* virulence expression is impaired by *Lactococcus lactis* during mixed cultures. *Appl. Environ. Microbiol*.

Fall S, Mercier A, Bertolla F, Calteau A, Gueguen L, Perrière G, Vogel TM, Simonet P. (2007). Horizontal gene transfer regulation in bacteria as a "spandrel" of DNA repair mechanisms. *PLoS One* **2**:e1055.

Falush D, Wirth T, Linz B. (2003). Traces of human migrations in *Helicobacter pylori* populations. *Science*. 299: 1582–1585.

Fearnhead P, Smith NGC, Barrigas M, Fox A, French N. (2005) Analysis of recombination in*Campylobacter jejuni* from MLST population data. *J Mol Evol.*. **61**: 333–340.

Feil EJ, Enright MC, Spratt BG. (2000). Estimating the relative contributions of mutation and recombination to clonal diversification: A comparison between Neisseria meningitidis and Streptococcus pneumoniae. *Res Microbiol* **151**:465–469.

Feil EJ, Maiden MC, Achtman M, Spratt BG. (1999). The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. Mol Biol Evol **16**:1496–1502.

Fernández-Gómez B, Fernàndez-Guerra A, Casamayor EO, González JM, Pedrós-Alió C, Acinas SG. (2012). Patterns and architecture of genomic islands in marine bacteria. *BMC Genomics*. **13**: 347.

Filiatrault MJ. (2011). Progress in prokaryotic transcriptomics. *Curr Opin Microbiol*. **14:**579–586.

Fraser C, Hanage WP, Spratt BG. (2007). Recombination and the nature of bacterial speciation. *Science*. **315**: 476–80.

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, Delong EF. (2008). Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci U S A*. **105**:3805–3810.

Fricke WF, Mammel MK, McDermott PF, Tartera C, White DG, LeClerc JE, Ravel J,

Cebula T A. (2011). Comparative genomics of 28 *Salmonella enterica* isolates: Evidence for CRISPR-mediated adaptive sublineage evolution. J Bacteriol **193**:3556–3568.

Friedman J, Alm EJ, Shapiro BJ. (2013). Sympatric speciation: when is it possible in bacteria*? PLoS One*. **8**:e53539.

Fukuchi S, Yoshimune K, Wakayama M, Moriguchi M, Nishikawa K. (2003). Unique Amino Acid Composition of Proteins in Halophilic Bacteria. *J Mol Biol*. **327**: 347–357.

Furuta Y, Abe K, Kobayashi I. 2010. Genome comparison and context analysis reveals putative mobile forms of restriction-modification systems and related rearrangements. *Nucleic Acids Res* **38**:2428–2443.

Garbeva P, Silby MW, Raaijmakers JM, Levy SB, de Boer W. (2011). Transcriptional and antagonistic responses of *Pseudomonas fluorescens* Pf0-1 to phylogenetically different bacterial competitors. *ISME J*. **5:**973– 985.

Garcia Martin, H, Ivanova, N, Kunin, V, Warnecke, F, Barry, K.W, McHardy, A.C. (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat Biotechnol.* **24**: 1263–1269.

Garcia-Gonzalez A, Vicens L, Alicea M, Massey SE. (2013). The distribution of recombination repair genes is linked to information content in bacteria. *Gene*. **528**:295–303.

Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Peer Y Van De, Vandamme P, Thompson FL, Swings J. (2005). Defining prokaryotic species Re-evaluating prokaryotic species. *Focus on horizontal transfer*. **3**:733–739.

Ghai R, Hernandez CM, Picazo A, Mizuno CM, Ininbergs K, Díez B, Valas R, DuPont CL, McMahon KD, Camacho A, Rodriguez-Valera F. (2012). Metagenomes of Mediterranean Coastal Lagoons. *Sci Rep.* **2**: 1–13.

Ghai R, Pašić L, Fernández AB, Martin-Cuadrado A-B, Mizuno CM, McMahon KD, Papke RT, Stepanauskas R, Rodriguez-Brito B, Rohwer F, Sánchez-Porro C, Ventosa A, Rodríguez-Valera F. (2011). New Abundant Microbial Groups in Aquatic Hypersaline Environments. *Sci Rep.* **1**: 1–10.

Gibbs MJ, Armstrong JS, Gibbs a J. (2000). Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics*. 16(7): 573–582.

Gifford SM, Sharma S, Booth M, Moran MA. 2012. Expression patterns reveal niche diversification in a marine microbial assemblage. *ISME J.* **7**:281–298.

Gomariz M, Martínez-García M, Santos F, Rodriguez F, Capella-Gutiérrez S, Gabaldón T, Rosselló-Móra R, Meseguer I, Antón J. (2015) From community approaches to single-cell genomics: the discovery of ubiquitous hyperhalophilic *Bacteroidetes* generalists. *ISME J*. (9) 16-31.

Gomez-Valero L, Rusniok C, Jarraud S, Vacherie B, Rouy Z, Barbe V, Medigue C, Etienne J, Buchrieser C. (2011). Extensive recombination events and horizontal gene transfer shaped the *Legionella pneumophila* genomes. *BMC Genomics*. **12**:536.

Gonzaga A, Martin-Cuadrado AB, López-Pérez M, Mizuno CM, García-Heredia I, Kimes NE, Lopez-García P, Moreira D, Ussery D, Zaballos M, Ghai R, Rodriguez-Valera F.(2012). Polyclonality of concurrent natural populations of *Alteromonas macleodii*. *Genome Biol Evol*. **4**: 1360–1374.

González-Candelas F, M. Francino P. (2012) Barriers to Horizontal Gene Transfer: Fuzzy and Evolvable Boundaries. *In* Francino P. (ed), *Caister Academic Press*, Norfolk, United Kingdom.

González-Escalona N, Martinez-Urtaza J, Romero J, Espejo RT, Jaykus LA, DePaola A. (2008). Determination of molecular phylogenetics of *Vibrio parahaemolyticus* strains by multilocus sequence typing. *J Bacteriol* **190**:2831–2840.

Goss EM. (2004). Genetic Diversity, Recombination and Cryptic Clades in *Pseudomonas viridiflava* Infecting Natural Populations of *Arabidopsis thaliana*. *Genetics* **169**:21–35.

Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res*. **36**:3420–3435.

Goudenège D, Labreuche Y, Krin E, Ansquer D, Mangenot S, Calteau A, Médigue C, Mazel D, Polz MF, Le Roux F. (2013). Comparative genomics of pathogenic lineages of *Vibrio nigripulchritudo* identifies virulence-associated traits. *ISME J.* **7**:1985–96.

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

Grissa I, Vergnaud G, Pourcel C. (2008). CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res*. **36**(19): 52–57.

Güell M, van Noort V, Yus E, Chen W-H, Leigh-Bell J, Michalodimitrakis K, Yamada T, Arumugam M, Doerks T, Kühner S, Rode M, Suyama M, Schmidt S, Gavin A-C, Bork P, Serrano L. (2009). Transcriptome complexity in a genome-reduced bacterium. *Science*. **326**: 1268–1271.

Guglielmini J, De La Cruz F, Rocha EPC. (2013). Evolution of conjugation and type IV secretion systems. *Mol Biol Evol*. **30**: 315–331.

Guglielmini J, Quintais L, Garcillán-Barcia MP, de la Cruz F, Rocha EPC. (2011). The Repertoire of ICE in Prokaryotes Underscores the Unity, Diversity, and Ubiquity of Conjugation. *PLoS Genet.* **7**: e1002222.

Guixa-Boixareu N, Calderón-Paz JI, Heldal M, Bratbak G y Pedrós-Alió C. (1996). Viral

lysis and bacterivory as prokaryotic loss factors along a salinity gradient. *Aquat Microbiol Ecol.* **11:** 215 – 227.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. **29**: 1072–1075.

Guy L, Nystedt B, Sun Y, Näslund K, Berglund EC, Andersson SGE. (2012). A genome-wide study of recombination rate variation in *Bartonella henselae*. *BMC Evol Biol* **12**: 65.

Hacker J, B. Kaper. (2010). Pathogenecity islands and the evlution on microbes. *Annu. Rev. Microbiol*. **54**: 641–79.

Hacker J, Carniel E. (2001). Ecological fitness, genomic islands and bacterial pathogenicity. A Darwinian view of the evolution of microbes. *EMBO Rep.* **2**:376–81.

Haft DH, Selengut J, Mongodin EF, Nelson KE. (2005). A Guild of 45 CRISPR-Associated (Cas) Protein Families and Multiple CRISPR/Cas Subtypes Exist in Prokaryotic Genomes. *PLoS Comput Biol*. **1**: e60.

Halpern D, Chiapello H, Schbath S, Robin S, Hennequet-Antier C, Gruss A, El Karoui M. (2007). Identification of DNA motifs implicated in maintenance of bacterial core genomes by predictive modeling. *PLoS Genet.* **3**:1614–21.

Hanage WP, Fraser C, Spratt BG. (2006). The impact of homologous recombination on the generation of diversity in bacteria. *J Theor Biol* **239**:210–9.

Hanage WP, Kaijalainen T, Herva E, Saukkoriipi A. (2005). Using Multilocus Sequence Data To Define the *Pneumococcus* Using Multilocus Sequence Data To Define the *Pneumococcus*. *Society.* **187**:6223–6230.

Handa N, Ichige A, Kobayashi I. (2009). Contribution of RecFOR machinery of homologous recombination to cell survival after loss of a restriction-modification gene complex. *Microbiology*. **155**: 2320–2332.

Hansen M T. (1978). Multiplicity of genome equivalents in the radiation-resistant bacterium *Micrococcus radiodurans*. *J. Bacteriol*. **134** (1): 71-75.

Hans-Jürgen Bandelt, Andreas W.M. Dress. (1992). Split decomposition: A new and useful approach to phylogenetic analysis of distance data. *Mol. phyl. evol.* **3(1):** 242–252.

Harth E, Romero J, Torres R, Espejo RT. (2007). Intragenomic heterogeneity and intergenomic recombination among Vibrio parahaemolyticus 16S rRNA genes. *Microbiology*. **153**: 2640–7.

Hartman AL, Norais C, Badger JH, Delmas S, Haldenby S, Madupu R, Robinson J, Khouri H, Ren Q, Lowe TM, Maupin-Furlow J, Pohlschroder M, Daniels C, Pfeiffer F, Allers T, Eisen J a. (2010. The Complete Genome Sequence of *Haloferax volcanii* DS2, a

Model Archaeon. *PLoS One.* **5**:e9605.

Held NL, Herrera A, Cadillo-Quiroz H, Whitaker RJ. (2010). CRISPR Associated Diversity within a Population of Sulfolobus islandicus. *PLoS One.* **5**: e12988.

Hernández-López A, Chabrol O, Royer-Carenzi M, Merhej V, Pontarotti P, Raoult D. 2013. To tree or not to tree? Genome-wide quantification of recombination and reticulate evolution during the diversification of strict intracellular bacteria. *Genome Biol Evol* **5**:2305–17.

Hiller NL, Ahmed A, Powell E, Martin DP, Eutsey R, Earl J, Janto B, Boissy RJ, Hogg J, Barbadora K, Sampath R, Lonergan S, Post JC, Hu FZ, Ehrlich GD. (2010). Generation of genic diversity among *Streptococcus pneumoniae* strains via horizontal gene transfer during a chronic polyclonal pediatric infection. *PLoS Pathog*. **6**: e1001108.

Hiller NL, Janto B, Hogg JS, Boissy R, Yu S, Powell E, Keefe R, Ehrlich NE, Shen K, Hayes J, Barbadora K, Klimke W, Dernovoy D, Tatusova T, Parkhill J, Bentley SD, Post JC, Ehrlich GD, Hu FZ. (2007). Comparative Genomic Analyses of Seventeen *Streptococcus pneumoniae* Strains: Insights into the *Pneumococcal* Supragenome. *J Bacteriol*. **189**: 8186–8195.

Ho Sui SJ, Fedynak A, Hsiao WWL, Langille MGI, Brinkman FSL. (2009). The Association of Virulence Factors with Genomic Islands. *PLoS One*. **4**: e8094.

Hogg JS, Hu FZ, Janto B, Boissy R, Hayes J, Keefe R, Post JC, Ehrlich GD.(2007). Characterization and modeling of the Haemophilus influenzae core and supragenomes based on the complete genomic sequences of Rd and 12 clinical nontypeable strains. *Genome Biol.* **8**: R103.

Holden MTG, Heather Z, Paillot R, Steward KF, Webb K, Ainslie F, Jourdan T, Bason NC, Holroyd NE, Mungall K, Quail M a, Sanders M, Simmonds M, Willey D, Brooks K, Aanensen DM, Spratt BG, Jolley K a, Maiden MCJ, Kehoe M, Chanter N, Bentley SD, Robinson C, Maskell DJ, Parkhill J, Waller AS. (2009). Genomic evidence for the evolution of *Streptococcus equi*: host restriction, increased virulence, and genetic exchange with human pathogens. *PLoS Pathog* **5**:e1000346.

Holmes E.C, Worobey, M. & Rambaut,A.(1999). Phylogenetic evidence for recombination in dengue virus. *Mol Biol and Evol*. **16**: 405-409.

Homan WL, Tribe D, Poznanski S, Li M, Hogg G, Spalburg E, Embden JD A, Van, Willems JL, Willems RJL. 2002. Multilocus Sequence Typing Scheme *for Enterococcus faecium*. *J Clin Microbiol* **40**:1963–71.

Hoskisson, P.A, and Smith, M.C. (2007). Hypervariation and phase variation in the bacteriophage 'resistome'. *Curr. Opin. Microbiol*. **10**: 396–400.

Hsiao WWL, Ung K, Aeschliman D, Bryan J, Finlay BB, Brinkman FSL. (2005). Evidence of a Large Novel Gene Pool Associated with Prokaryotic *Genomic Islands*. *PLoS Genet.* **1**: e62.

Huang C-L, Pu P-H, Huang H-J, Sung H-M, Liaw H-J, Chen Y-M, Chen C-M, Huang M-B, Osada N, Gojobori T, Pai T-W, Chen Y-T, Hwang C-C, Chiang T-Y. (2015). Ecological genomics in Xanthomonas: the nature of genetic adaptation with homologous recombination and host shifts. *BMC Genomics*. **16**:1369.

Huang Q, Cheng X, Cheung MK, Kiselev SS, Ozoline ON, Kwan HS. (2012). High-density transcriptional initiation signals underline genomic islands in bacteria. *PLoS One*. **7**.

Ietswaart R, Szardenings F, Gerdes K, Howard M. (2014). Competing ParA Structures Space Bacterial Plasmids Equally over the Nucleoid. *PLoS Comput Biol*. **10**: e1004009.

Ikeda, H. Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nature Biotechnol*. 21, 526–531 (2003).

Imam S, Chen Z, Roos DS, Pohlschröder M. (2011). Identification of surprisingly diverse type IV pili, across a broad range of gram-positive bacteria. *PLoS One*. **6**: e28919.

Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat Genet* **44**:226–232.

Iqbal Z, *et al*. (2012). De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet*. **199:**133–154

Ivars-Martinez E, Martin-Cuadrado A-B, D'Auria G, Mira A, Ferriera S, Johnson J, Friedman R, Rodriguez-Valera F. (2008). Comparative genomics of two ecotypes of the marine planktonic copiotroph *Alteromonas macleodii* suggests alternative lifestyles associated with different kinds of particulate organic matter. *ISME J*. **2**: 1194–212.

Jackman SD, Birol I (2010). Assembling genomes using shortread sequencing technology. *Genome Biol.* **11**: 202.

Jain R, Rivera MC, Lake J a. (1999). Horizontal gene transfer among genomes: the complexity hypothesis. *Proc Natl Acad Sci U S A*. **96**: 3801–3806.

Jeltsch A. (2003). Maintenance of species identity and controlling speciation of bacteria: A new function for restriction/modification systems? *Gene* **317**: 13–16.

Jiang SC, Paul JH. (1998). Gene transfer by transduction in the marine environment. *Appl Environ Microbiol* **64**:2780–2787.

Johnson PL, Slatkin M. (2009) Inference of microbial recombination ratesfrom metagenomic data. *PLoS Genetics*. **5**: e1000674.

Jolley K A. The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. (2004). *Mol Biol Evol*. **22**: 562–569.

Joo W a, Kim CW. (2005). Proteomics of Halophilic archaea. *J Chromatogr B Anal Technol Biomed Life Sci*. **815**: 237–250.

Jordan IK, Rogozin IB, Wolf YI, Koonin E V. (2002). Microevolutionary genomics of bacteria. *Theor Popul Biol*. **61**: 435–447.

Joseph SJ, Didelot X, Gandhi K, Dean D, Read TD. (2011). Interplay of recombination and selection in the genomes of *Chlamydia trachomatis*. *Biol Direct.* **6**: 28.

Juhas M, van der Meer JR, Gaillard M, Harding RM, Hood DW, Crook DW. 2009. Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiol Rev.* **33**: 376–93.

Kado CI. (2009). Horizontal gene transfer: Sustaining pathogenicity and optimizing host-pathogen interactions. *Mol Plant Pathol*. **10**: 143–150.

Kamita M, Kimura Y, Ino Y, Kamp RM, Polevoda B, Sherman F, Hirano H. (2011). N - Acetylation of yeast ribosomal proteins and its effect on protein synthesis. *J Proteomics* **74**:431–441.

Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P, Malmstrom RR, Stocker R, Follows MJ, Stepanauskas R, Chisholm SW. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science*. **344**:416–20.

Kennemann L, Didelot X, Aebischer T, Kuhn S, Drescher B, Droege M, Reinhardt R, Correa P, Meyer TF, Josenhans C, Falush D, Suerbaum S. (2011). *Helicobacter pylori* genome evolution during human infection. *Proc Natl Acad Sci U S A***. 108**:5033–8.

Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, Muetzel B, Wirkner U, Ansorge W, Pääbo S. (2004). A neutral model of transcriptome evolution. *PLoS Biol*. **2**.

Kimes NE, López-Pérez M, Ausó E, Ghai R, Rodriguez-Valera F. (2014). RNA sequencing provides evidence for functional variability between naturally co-existing *Alteromonas macleodii* lineages. *BMC Genomics.* **15**: 938.

Klappenbach, J.A, P.R. Saxman, J.R. Cole and T.M. Schmidt. (2001) rrnDB: the Ribosomal RNA Operon Copy Number Database. *Nucleic Acids Res*. **29**(1): 181-184.

Kleter, G.A, Peijnenburg, A.A, and Aarts, H.J. (2005). Health considerations regarding horizontal transfer of microbial transgenes present in genetically modified crops. *J. Biomed. Biotechnol***. 2005**, 326–352.

Kloesges T, Popa O, Martin W, Dagan T. (2011). Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different Phylogenetic Depths. *Mol Biol Evol.* **28**:1057–1074.

Koc H, Cimen H, Stallard A, Zeynep C, Koc C.. (2012). A Comprehensive analysis of lysine acetylation in ribosomal proteins.*The FASEB Journal*. 26:958.2.

Konstantinidis K.T, Tiedje J.M. (2005). Genomic insights that advance the species definition for prokaryotes.Proc. *Natl Acad. Sci. USA*. 102:2567–2572.

Konstantinidis KT, DeLong EF. (2008). Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. *ISME J* **2**: 1052–1065.

Konstantinidis KT, Ramette A, Tiedje JM. (2006). The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci*. **361**:1929–40.

Konstantinidis KT, Serres MH, Romine MF, Rodrigues JLM, Auchtung J, McCue L-A, Lipton MS, Obraztsova A, Giometti CS, Nealson KH, Fredrickson JK, Tiedje JM. (2009). Comparative systems biology across an evolutionary gradient within the *Shewanella* genus. *Proc Natl Acad Sci U S A*. **106**:15909–14.

Korotkov K V, Sandkvist M, Hol WGJ. (2012). The type II secretion system: biogenesis, molecular architecture and mechanism. *Nat Rev Microbiol*. **10**: 336–351.

Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD. (2006) Automated phylogenetic detection of recombination using a genetic algorithm. *Molecular Biology and Evolution*. **23**: 1891–1901.

Kotra LP, Haddad J, Mobashery S. (2000). Aminoglycosides: Perspectives on mechanisms of action and resistance and strategies to counter resistance. *Antimicrob Agents Chemother.* **44**:3249–3256.

Kumar N, Lad G, Giuntini E, Kaye ME, Udomwong P, Shamsani NJ, Young JPW, Bailly X. (2015). Bacterial genospecies that are not ecologically coherent: population genomics of *Rhizobium leguminosarum*. *Open Biol* **5**:140133.

Kumar R, Lawrence ML, Watt J, Cooksey AM, Burgess SC, Nanduri B. (2012). RNA-Seq based transcriptional map of bovine respiratory disease pathogen *Histophilus somni* 2336. *PLoS One*. **7**:1–12.

Kurtz S, Phillippy A, Delcher A.L, Smoot M, Shumway M, Antonescu C, and Salzberg S.L, (2004). Versatile and open software for comparing large genomes. *Genome Biology*. **5**: 12.

Kuwahara T, Yamashita A, Hirakawa H, Nakayama H, Toh H, Okada N, Kuhara S, Hattori M, Hayashi T, Ohnishi Y. (2004). Genomic analysis of *Bacteroides fragilis* reveals extensive DNA inversions regulating cell surface adaptation. *Proc Natl Acad Sci U S A*. **101**: 14919–14924.

Lan R, Reeves PR. (2001). When does a clone deserve a name? A perspective on bacterial species based on population genetics. *Trends Microbiol***. 9**: 419–424

Landsman D. (1992). RNP-1, an RNA-binding motif is conserved in the DNA-binding cold shock domain. *Nucleic Acids Res*. **20**:2861–2864.

Lapierre P, Gogarten JP. (2009). Estimating the size of the bacterial pan-genome. *Trends*

Genet. **25**: 107–110.

Larsson, P, D. Elfsmark, K. Svensson, P. Wikstrm, M. Forsman et al, (2009). Molecular evolutionary consequences of niche restriction in *Francisella tularensis*, a facultative intracellular pathogen. *PLoS Pathog*. **5**(6).

Leplae R, Hebrant A, Wodak SJ, Toussaint A. (2004). ACLAME: a Classification of mobile genetic Elements. *Nucleic Acids Res*. **32**: 45–59.

Li B, Ibrahim M, Ge M, Cui Z, Sun G, Xu F, Kube M. (2014). Transcriptome analysis of Acidovorax avenae subsp. avenae cultivated in vivo and co-culture with *Burkholderia seminalis*. *Nat Sci reports*. **4**:5698.

Li H, Durbin R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*. **26**(5): 589–595.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*.20: 265–272.

Liolios K, Chen IMA, Mavromatis K, Tavernarakis N, Hugenholtz P, Markowitz VM y Kyrpides NC. 2010. The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucl Acids Res.* **38:** 346 – 354.

Liu L, Li Y, Li S, Hu N, He Y, Pong R *et al.* (2012). Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology* : 1–11.

Liu X, Gutacker MM, Musser JM, Fu Y-X. (2006). Evidence for recombination in *Mycobacterium tuberculosis*. *J Bacteriol.* **188**:8169–77.

Lodders N, Stackebrandt E, Nübel U. (2005). Frequent genetic recombination in natural populations of the marine cyanobacterium Microcoleus chthonoplastes. *Environ Microbiol* **7**:434–442.

Longo LM, Lee J, Blaber M. (2013). Simplified protein design biased for prebiotic amino acids yields a foldable, halophilic protein. *Proc Natl Acad Sci*. **110**: 2135–2139.

López-Pérez M, Gonzaga A, Ivanova EP, Rodriguez-Valera F. (2014). Genomes of *Alteromonas australica*, a world apart. *BMC Genomics*. **15**: 483.

López-Pérez M, Gonzaga A, Rodriguez-Valera F. (2013). Genomic diversity of "deep ecotype" *Alteromonas macleodii* isolates: Evidence for pan-mediterranean clonal frames. *Genome Biol Evol*. **5**: 1220–1232.

Luo C, Konstantinidis KT. (2011). Phosphorus-related gene content is similar in *Prochlorococcus* populations from the North Pacific and North Atlantic Oceans. *Proc Natl Acad Sci U S A*. **108**: E62–E63; author reply E64–E66.

Luo C, Tsementzi D, Kyrpides NC, Konstantinidis KT. (2012). Individual genome assembly from complex community short-read metagenomic datasets. *ISME J*. **6**:898–901.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*. **1**: 18.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu S-M, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam T-W, Wang J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**:18.

Lythgoe K a, Chao L. (2003). Mechanisms of coexistence of a bacteria and a bacteriophage in a spatially homogeneous environment. *Ecol Let*t. **6**: 326–334.

Lythgoe K A, Chao L. (2003). Mechanisms of coexistence of a bacteria and a bacteriophage in a spatially homogeneous environment. *Ecol Lett*. **6**:326–334.

Magoc T, Pabinger S, Canzar S, Liu X, Su Q, Puiu D, Tallon LJ, Salzberg SL. (2013). GAGE-B: An evaluation of genome assemblers for bacterial organisms. *Bioinformatics*. **29**:1718–1725.

Maiden, M.C.J, (2006). Multilocus sequence typing of bacteria. *Annual review of microbiology*. **60**: 561–588.

Majewski J. (2001). Sexual isolation in bacteria. *FEMS Microbiol Lett* **199**:161–169.

Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, Moineau S, Mojica FJM, Wolf YI, Yakunin AF, van der Oost J, Koonin E V. (2011). Evolution and classification of the CRISPR–Cas systems. *Nat Rev Microbiol.* **9**: 467–477.

Makarova KS, Wolf YI, Koonin E V. (2013). Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res.* **41**:4360–4377.

Makarova KS, Wolf YI, Snir S, Koonin E V. 2011. Defense Islands in Bacterial and Archaeal Genomes and Prediction of Novel Defense Systems. *J Bacteriol* **193**:6039–6056.

Mandel MJ, Wollenberg MS, Stabb E V, Visick KL, Ruby EG. (2009). A single regulatory gene is sufficient to alter bacterial host range. *Nature*. **458**:215–218.

Marenda M, Barbe V, Gourgues G, Mangenot S, Sagne E, Citti C. (2006). A new integrative conjugative element occurs in *Mycoplasma agalactiae* as chromosomal and free circular forms. *J Bacteriol*. **188**: 4137–41.

Markowitz VM, Chen IM a, Palaniappan K, Chu K, Szeto E, Grechkin Y, et al. (2012). IMG: The integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res*. **40**: 115–

Markowitz VM, Chen I-M a, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner a, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC. (2012). IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Res.* **40**: D115–D122.

Martin D, Rybicki E. (2000). RDP: detection of recombination amongst aligned sequences. Bioinformatics. **16**(6): 562–563.

Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. (2010). RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics*. **26**(19): 2462–2463.

Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. (2010). RDP3: a flexible and fast computer program for analyzing recombination. *Bioinformatics.* **26**: 2462–2463.

Martin DP, Lemey P, Posada D. (2011). Analysing recombination in nucleotide sequences. *Mol Ecol Resour*. **11**(6): 943–955.

Martin, D. P, Posada, D, Crandall, K. A. & Williamson, C. (2005). A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. *AIDS Res Hum Retroviruses* **21**: 98-102.

Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. (2012). PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* **13**:31.

Materna AC, Friedman J, Bauer C, David C, Chen S, Huang IB, Gillens A, Clarke S a, Polz MF, Alm EJ. (2012). Shape and evolution of the fundamental niche in marine *Vibrio*. *ISME J*. **6**:2168–2177.

Maturrano L, Santos F, Rosselló-Mora R y Antón J. (2006a). Microbial diversity in Maras salterns, a hypersaline environment in the Peruvian Andes. *Appl Environ Microbiol.* **72**: 3887 – 3895.

Maturrano L, Valens-Vadell M, Rosselló-Mora R y Antón J. (2006b). *Salicola marasensis* gen. nov, sp. nov, an extremely halphilic bacterium isolated from the Maras solar salterns in Peru. *Int J Syst Evol Microbiol.* **56:** 1685 – 1691.

Mau B, Glasner JD, Darling AE, Perna NT. (2006). Genome-wide detection and analysis of homologous recombination among sequenced strains of *Escherichia coli*. *Genome Biol.* **7**:R44.

Maynard Smith J. (1992). Analyzing the mosaic structure of genes. *J Mol Evol*. **34**: 126-129.

Mayor D, Zeeh F, Frey J, Kuhnert P. (2007). Diversity of *Mycoplasma hyopneumoniae* in pig farms revealed by direct molecular typing of clinical material. *Vet Res* **38**:391–8.

McCready S, Müller J a, Boubriak I, Berquist BR, Ng WL, DasSarma S. (2005). UV irradiation induces homologous recombination genes in the model archaeon, *Halobacterium* sp. NRC-1. *Saline Systems.* **1**: 3.

McMillan DJ, Bessen DE, Pinho M, Ford C, Hall GS, Melo-Cristino J, Ramirez M. (2010. Population genetics of *Streptococcus dysgalactiae* subspecies equisimilis reveals widely dispersed clones and extensive recombination. *PLoS One* **5**:e11741.

McVean, G, Awadalla, P. & Fearnhead, P. (2002). A Coalescent-Based Method for Detecting and Estimating Recombination From Gene Sequences. *Genetics*. **160**: 1231-1241.

Meats E, Feil EJ, Stringer S, Cody AJ, Goldstein R, Kroll JS, Popovic T, Spratt BG. (2003). Characterization of encapsulated and noncapsulated *Haemophilus influenzae* and determination of phylogenetic relationships by multilocus sequence typing. *J Clin Microbiol* **41**:1623–36.

Meier, P, and Wackernagel, W. (2005). Impact of mutS inactivation on foreign DNA acquisition by natural transformation in *Pseudomonas stutzeri*. *J. Bacteriol*. **187**: 143–154.

Melville S, Craig L. (2013). Type IV Pili in Gram-Positive Bacteria. *Microbiol Mol Biol Rev*. **77**: 323–341.

Mes THM. (2008). Microbial diversity - Insights from population genetics. *Environ Microbiol*. **10**:251–264.

Metzker ML (2010). Sequencing technologies - the next generation. *Nat Rev Genet***. 11**: 31–46.

Michod RE, Bernstein H, Nedelcu AM. (2008). Adaptive value of sex in microbial pathogens. *Infect Genet Evol.* **8**: 267–285.

Michod RE, Bernstein H, Nedelcu AM. (2008). Adaptive value of sex in microbial pathogens. *Infect Genet Evol* **8**:267–285.

Mierzejewska J, Jagura-Burdzy G. (2012). Prokaryotic ParA–ParB–parS system links bacterial chromosome segregation with the cell cycle. *Plasmid.* **67**: 1–14.

Miller SR, Castenholz RW, Pedersen D. (2007. Phylogeography of the thermophilic cyanobacterium *Mastigocladus laminosus*. *Appl Environ Microbiol.* **73**:4751–9.

Minin VN, Dorman KS, Fang F, Suchard M a. (2005). Dual multiple change-point model leads to more accurate recombination detection. *Bioinformatics.* **21**: 3034–3042.

Mira A, Martin-Cuadrado a B, D'Auria G, Rodriguez-Valera F. (2010). The bacterial pan-genome:a new paradigm in microbiology. *Int Microbiol.* **13**: 45–57.

Mitsuhashi M. (1996). Technical Report: Part2. Basic requirements for designing optimal PCR primers. *Journal of Clinical Laboratory Analysis*. **10**: 285-293.

Mojica FJM, Díez-Villaseñor C, Soria E, Juez G. (2000). Biological significance of a family of regularly spaced repeats in the genomes of *Archaea, Bacteria* and mitochondria. *Mol Microbiol*. **36**: 244–246.

Mojica, F.J, Diez-Villasenor, C, Garcia-Martinez, J, and Soria, E. (2005). Intervening

sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J. Mol. Evol*. **60**: 174–182.

Mongodin EF, Nelson KE, Daugherty S, Deboy RT, Wister J, Khouri H, Weidman J, Walsh D a, Papke RT, Sanchez Perez G, Sharma a K, Nesbø CL, MacLeod D, Bapteste E, Doolittle WF, Charlebois RL, Legault B, Rodriguez-Valera F. (2005). The genome of *Salinibacter ruber*: convergence and gene exchange among hyperhalophilic bacteria and archaea. *Proc Natl Acad Sci U S A*. **102**: 18147–18152.

Morelli G, Didelot X, Kusecek B, Schwarz S, Bahlawane C, Falush D, Suerbaum S, Achtman M. (2010). Microevolution of Helicobacter pylori during prolonged infection of single hosts and within families. *PLoS Genet.* **6**: e1001036.

Muller EEL, Glaab E, May P, Vlassis N, Wilmes P. (2013). Condensing the omics fog of microbial communities. *Trends Microbiol*. **21**:325–33.

Murray NE. (2002). Immigration control of DNA in bacteria: Self versus non-self. *Microbiology.* **148**: 3–20.

Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, Kariuki S, Croucher NJ, Choi SY, Harris SR, Lebens M, Niyogi SK, Kim EJ, Ramamurthy T, Chun J, Wood JLN, Clemens JD, Czerkinsky C, Nair GB, Holmgren J, Parkhill J, Dougan G. (2011). Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature*. **477**: 462–465.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*. **320**:1344–1349.

Nandi T, Holden MTG, Didelot X, Mehershahi K, Boddey J a, Beacham I, Peak I, Harting J, Baybayan P, Guo Y, Wang S, How LC, Sim B, Essex-Lopresti A, Sarkar-Tyson M, Nelson M, Smither S, Ong C, Aw LT, Hoon CH, Michell S, Studholme DJ, Titball R, Chen SL, Parkhill J, Tan P. (2015). *Burkholderia pseudomallei* sequencing identifies genomic clades with distinct recombination, accessory, and epigenetic profiles. *Genome Res*. **25**:129–41.

Naor A, Lapierre P, Mevarech M, Papke RT, Gophna U. (2012). Low Species Barriers in Halophilic Archaea and the Formation of Recombinant Hybrids. *Curr Biol.* **22**: 1444–1448.

Naor A, Lapierre P, Mevarech M, Papke RT, Gophna U. (2012). Low species barriers in halophilic *Archaea* and the formation of recombinant hybrids. *Curr Biol*. **22**:1444–1448.

Narasingarao P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, Brocks JJ, Heidelberg KB, Banfield JF, Allen EE.(2011). De novo metagenomic assembly reveals abundant novel major lineage of *Archaea* in hypersaline microbial communities. *The ISME Journal*.

Narra HP, Ochman H. (2006). Of what use is sex to bacteria? Curr Biol **16**:R705–10.

Navarre, W.W, McClelland, M, Libby, S.J, and Fang, F.C. (2007). Silencing of xenogeneic DNA by H-NSfacilitation of lateral gene transfer in bacteria by a defense system that recognizes foreign DNA. *Genes Dev.* **21**: 1456–1471.

Nicolas P, Mondot S, Achaz G, Bouchenot C, Bernardet JF, Duchaud E. (2008). Population structure of the fish-pathogenic bacterium *Flavobacterium psychrophilum*. *Appl Environ Microbiol.* **74**:3702–3709.

Nikel PI, Silva-Rocha R, Benedetti I, De Lorenzo V. (2014). The private life of environmental bacteria: Pollutant biodegradation at the single cell level. *Environ Microbiol*. **16**:628–642.

Nolan M, Tindall BJ, Pomrenke H, Lapidus A, Copeland A, Glavina Del Rio T, Lucas S, Chen F, Tice H, Cheng J-F, Saunders E, Han C, Bruce D, Goodwin L, Chain P, Pitluck S, Ovchinikova G, Pati A, Ivanova N, Mavromatis K, Chen A, Palaniappan K, Land M, Hauser L, Chang Y-J, Jeffries CD, Brettin T, Göker M, Bristow J, Eisen J a, Markowitz V, Hugenholtz P, Kyrpides NC, Klenk H-P, Detter JC. (2009). Complete genome sequence of *Rhodothermus marinus* type strain (R-10). *Stand Genomic Sci.* **1**: 283–290.

Norman, A, Hansen, L.H, and Sorensen, S.J. (2009). Conjugative plasmids: vessels of the communal gene pool. *Philos. Trans. R Soc. Lond. B Biol. Sci*. **364**: 2275–2289.

Nunney L, Hopkins DL, Morano LD, Russell SE, Stouthamer R. (2014. Intersubspecific recombination in *Xylella fastidiosa* strains native to the united states: Infection of novel hosts associated with an unsuccessful invasion. *Appl Environ Microbiol* **80**:1159–1169.

Ochman H, Lawrence JG, Groisman EA. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature.* **405**: 299–304.

of two enhanced biological phosphorus removal (EBPR)

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. **27**(1): 29–34.

Oh, S, Caro-Quintero, A, Tsementzi, D, Deleon-Rodriguez, N, Luo, C, Poretsky, R, and Konstantinidis, K.T. (2011) Metagenomic insights into the evolution, function and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Appl Environ Microbiol.* doi:10.1038/ismej.2011.147.

Olvera A, Cerdà-Cuéllar M, Aragon V. (2006). Study of the population structure of Haemophilus parasuis by multilocus sequence typing. *Microbiology*. **152**:3683–90.

Oren A, Mana L. (2003). Sugar metabolism in the extremely halophilic bacterium *Salinibacter ruber*. FEMS *Microbiol Lett*. **223**:83–87.

Oren A. (2002a). Diversity of halophilic microorganisms: environments, phylogeny, physiology, and applications. *J Ind Microbiol & Biotech*. **28:** 56 – 63.

Oren A. (2002b). Solar salterns. En cellular origin and life in extreme habitats. Halophilic microorganisms and their environments. *Kluwer academic Publisher*. **pp:** 441 – 462.

Oren A. (2002c). Molecular ecology of extremely halophilic *Archaea* and *Bacteria*. *FEMS Microbiol Ecol*. **39**: 1–7.

Oren A. (2005). A hundred years of *Dunaliella* research: *Saline Systems*. **1**: 1905 – 2005.

Oren A. (2008). Microbial life at high salt concentrations: phylogenetic and metabolic diversity. *Saline Systems*. **4:** 2.

Oren A. (2013). *Salinibacter* : an extremely halophilic bacterium with archaeal properties. *FEMS Microbiol Lett*. **342**: 1–9.

Orsi RH, Sun Q, Wiedmann M. (2008). Genome-wide analyses reveal lineage specific contributions of positive selection and recombination to the evolution of *Listeria monocytogenes*. *BMC Evol Biol* **8**:233.

Ottesen E a, Young CR, Eppley JM, Ryan JP, Chavez FP, Scholin C a, DeLong EF. (2013). Pattern and synchrony of gene expression among sympatric marine microbial populations. *Proc Natl Acad Sci U S A*. **110**:E488–97.

Padidam, M, Sawyer, S. & Fauquet, C. M. (1999). Possible emergence of new geminiviruses by frequent recombination. *Virology.* **265**: 218-225.

Pagani I, Liolios K, Jansson J, Chen IM, Smirnova T, *et al*. (2012). The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res*. **40**: 571–579.

Pallen, M.J, and Wren, B.W. (2007). Bacterial pathogenomics. *Nature*. **449**: 835–842.

Pannekoek Y, Morelli G, Kusecek B, Morré S a, Ossewaarde JM, Langerak A a, van der Ende A. (2008. Multi locus sequence typing of Chlamydiales: clonal groupings within the obligate intracellular bacteria *Chlamydia trachomatis*. *BMC Microbiol*. **8**:42.

Papke RT, Koenig JE, Rodríguez-Valera F, Doolittle WF. (2004). Frequent recombination in a saltern population of Halorubrum. *Science*. **306**: 1928–1929.

Papke RT, Zhaxybayeva O, Feil EJ, Sommerfeld K, Muise D, Doolittle WF. (2007). Searching for species in haloarchaea. *Proc Natl Acad Sci U S A*. **104**: 14092–14097.

Pašić L, Bartualb S. G, Ulrihc N. P, Miklavž Grabnara M, Velikonjaa B.H. (2005). Diversity of halophilic archaea in the crystallizers of an Adriatic solar saltern. *FEMS. Microbiol. Ecol*. **54**: 491-498.

Pasić L, Rodriguez-Mueller B, Martin-Cuadrado A-B, Mira A, Rohwer F, Rodriguez-Valera F. (2009). Metagenomic islands of hyperhalophiles: the case of *Salinibacter ruber*. *BMC Genomics.* **10**: 570.

Paul S, Bag SK, Das S, Harvill ET, Dutta C. (2008). Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. Genome Biol **9**:R70.

Pedrós-Alió C. (2006). Marine microbial diversity: can it be determined?. *Trends Microbiol.*

Peña A, Gomariz M, Lucio M, González-Torres P, Huertas-Cepa J, Martínez-García M, Santos F, Schmitt-Kopplin P, Gabaldón T, Rosselló- Mora R, Antón J. (2014). *Salinibacter ruber*: the never ending microdiversity. p 37–53. *In* Papke T, Oren A, Ventosa A (ed), Genetics and genomics of halophiles. Caister Academic Press, Norfolk, United Kingdom.

Peña A, Teeling H, Huerta-Cepas J, Santos F, Meseguer I, Lucio M, Schmitt-Kopplin P, Dopazo J, Rossello´-Mora R, Schuler M, Oliver F, Amann R, Toni Gabaldón T, and Antón J.

Peña A, Teeling H, Huerta-Cepas J, Santos F, Yarza P, Brito-Echeverría J, Lucio M, Schmitt-Kopplin P, Meseguer I, Schenowitzs C, Dossat C, Barbe V, Dopazo J, Rosselló-Mora R, Schüler M, Glöckner FO, Amman R, Gabaldón T y Antón J. (2010). Fine-scale evolution: genomic, phenotypic and ecological differentiation in two coexisting *Salinibacter ruber* strains. *ISME J*. **4**: 882-895.

Peña A, Valens M, Santos F, Buczolits S, Antón J, Kämpfer P, Busse HJ, Amann R y Rosselló-Mora R. (2005). Intraspecific comparative analysis of the species *Salinibacter ruber*. *Extremophiles*. **9:** 151 – 161.

Peng Y, Leung HCM, Yiu SM, Chin FYL. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics.* **28**:1420–8.

Persson OP, Pinhassi J, Riemann L, Marklund B-I, Rhen M, Normark S, González JM, Hagström Å. (2009). High abundance of virulence gene homologues in marine bacteria. *Environ Microbiol.* **11**: 1348–1357.

Petersen T, Brunak S, Heijne G. and Nielsen H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*. **8**: 785-786.

Pevzner P a, Tang H, Waterman MS. (2001). An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci U S A* **98**:9748–9753.

Pierlé SA, Dark MJ, Dahmen D, Palmer GH, Brayton K a. (2012). Comparative genomics and transcriptomics of trait-gene association. *BMC Genomics*. **13**:669.

Pinto a. C, Melo-Barbosa HP, Miyoshi A, Silva A, Azevedo V. (2011). Application of RNA-seq to reveal the transcript profile in bacteria. *Genet Mol Res*. **10**:1707–1718.

Pinto UM, Pappas KM, Winans SC. (2012). The ABCs of plasmid replication and segregation. *Nat Rev Microbiol*. **10**: 755–765.

Polz MF, Alm EJ, Hanage WP. (2013). Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet*. **29**:170–5.

Popa O, Hazkani-covo E, Landan G, Martin W, Dagan T. (2011). Directed networks reveal genomic barriers and DNA repair bypasses to lateral gene transfer among prokaryotes 599–609.

Posada D, Crandall K A. (2001). Evaluation of methods for detecting recombination from DNA sequences: computer simulations. *Proc Natl Acad Sci U S A*. **98**:13757–62.

Posada D. (2001). Evaluation of methods for detecting recombination from DNA sequences: empirical data. *Mol Biol*. **3**: 708–17.

Preheim SP, Timberlake S, Polz MF. (2011). Merging taxonomy with ecological population prediction in a case study of Vibrionaceae. *Appl Environ Microbiol*. **77**:7195–206.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glockner FO. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res*. **35**:7188–7196.

Punita J.S.J, Reddy MA, Das HK. (1989). Multiple chromosomes of *Azotobacter vinelandii*. *J. Bacteriol*. **171**(6): 3133-3138.

Qin, J, Li, R, Raes, J, Arumugam, M, Burgdorf, K.S, Manichanh, C, *et al*. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature. **464**: 59–65. Rossello-Mora, R

Qiu W-G, Schutzer SE, Bruno JF, Attie O, Xu Y, Dunn JJ, Fraser CM, Casjens SR, Luft BJ. (2004). Genetic exchange and plasmid transfers in *Borrelia burgdorferi* sensu stricto revealed by three-way genome comparisons and multilocus sequence typing. *Proc Natl Acad Sci U S A*. **101**: 14150–14155.

Quail M, Smith M, Coupland P, Otto T, Harris S, Connor T *et al*. (2012). A tale of three next generation sequencing platforms: comparison of Ion torrent, pacific biosciences and illumina MiSeq sequencers″. *BMC Genomics* 13 (1).

Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*. **26**(6): 841–842.

Rasmussen, R P. (2001). Quantification on the LightCycler. In S. Meuer, C.T. Wittwer, and K. Nakagawara (Eds.), Rapid cycle real-time PCR, methods and applications. *Springer Press*. **pp**. 21-34.

Redfield RJ, Cameron ADS, Qian Q, Hinds J, Ali TR, Kroll JS, Langford PR. (2005). A novel CRP-dependent regulon controls expression of competence genes in *Haemophilus influenzae*. *J Mol Biol*. **347**: 735–747.

Reed CCJ, Lewis H, Trejo E, Winston V, Evilia C. (2013). Protein adaptations in archaeal extremophiles. *Archaea*. **2013**: 373275.

Reeves PR, Liu B, Zhou Z, Li D, Guo D, Ren Y, Clabots C, Lan R, Johnson JR, Wang L. (2011). Rates of Mutation and Host Transmission for an *Escherichia coli* Clone over 3 Years. *PLoS One*. **6**: e26907.

Reno ML, Held NL, Fields CJ, Burke P V, Whitaker RJ. (2009). Biogeography of the *Sulfolobus islandicus* pan-genome. *Proc Natl Acad Sci U S A*. **106**: 8605–8610.

Reva O, Tümmler B. (2008). Think big - Giant genes in bacteria. *Environ Microbiol*. **10**: 768–777.

Riley MA, Lizotte-waniewski M. (2009). Population genomics and the bacterial species concept. *Methods Mol Biol*. 532: 367–377.

Roberts AP, Kreth J. (2014). The impact of horizontal gene transfer on the adaptive ability of the human oral microbiome. *Front Cell Infect Microbiol*. **4**: 1–9.

Roberts, R.J, Vincze, T, Posfai, J, Macelis, D. (2010). REBASE--a database for DNA restriction and modification: enzymes, genes and genomes. *Nucl. Acids Res*. **38**: 234-236.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. (2011). Integrative genomics viewer. *Nat Biot*. **29**:24–26.

Rocha EPC, Cornet E, Michel B. (2005). Comparative and evolutionary analysis of the bacterial homologous recombination systems. *PLoS Genet* **1**:e15.

Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, Breitbart M, Buchanan J, Desnues C, Dinsdale E, Edwards R, Felts B, Haynes M, Liu H, Lipson D, Mahaffy J, Martin-Cuadrado a B, Mira a, Nulton J, Pasic L, Rayhawk S, Rodriguez-Mueller J, Rodriguez-Valera F, Salamon P, Srinagesh S, Thingstad TF, Tran T, Thurber R V, Willner D, Youle M, Rohwer F. (2010). Viral and microbial community dynamics in four aquatic environments. *ISME J*. **4**: 739–751.

Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pašić L, Thingstad TF, Rohwer F, Mira A. (2009). Explaining microbial population genomics through phage predation. *Nat Rev Microbiol.* **7**: 828–836.

Rodríguez-Valera F. (1988). Characteristics and microbial ecology of hypersaline environments. En *Halophilic Bacteria*. **1**: 3 – 30. *Editado por F. Rodríguez-Valera. Boca-Raton, FL: CRC Press.*

Rose, R.W, T. Brüser,. J. C. Kissinger, and M. Pohlschröder. (2002). Adaptation of protein secretion to extremely high salt concentrations by extensive use of the twin arginine translocation pathway. *Mol. Microbiol*. **5**: 943-950.

Rosselló-Mora R, Lucio M, Peña A, Brito-Echeverría J, López-López A, Valens-Vadell M, Frommberger M, Antón J y Schmitt-Kopplin P. (2008). Metabolic evidence for biogeographic isolation of the extremophilic bacterium *Salinibacter ruber*. *The ISME J.* **2:** 242 – 253.

Rosselló-Mora R, Lucio M, Peña A, Brito-Echeverría J, López-López A, Valens-Vadell M, Frommberger M, Antón J, Schmitt-Kopplin P. (2008). Metabolic evidence for biogeographic isolation of the extremophilic bacterium *Salinibacter ruber. ISME J*. **2**:242–253.

Rychlik W. (1995). Selection of primers for polimerase chain reaction. *Molecular Biotechnology*. volume 3.

Salcedo C, Arreaza L, Alcalá B, Fuente L De, Alcala B. (2003. Development of a Multilocus Sequence Typing Method for Analysis of *Listeria monocytogenes* clones development of a multilocus. *J Clin Microbiol*. **41**:757–762.

Salminen MO1, Carr JK, Burke DS, McCutchan FE. (1995). Identification of breakpoints in intergenotypic recombinants of HIV type 1 by bootscanning. AIDS Res Hum Retroviruses. **11**(11): 1423-1425.

Salter I, Galand PE, Fagervold SK, Lebaron P, Obernosterer I, Oliver MJ, Suzuki MT, Tricoire C. (2014). Seasonal dynamics of active SAR11 ecotypes in the oligotrophic Northwest Mediterranean Sea. *ISME J*. **9**:347–360.

Sambrook J, Fritsch EF y Maniatis T. (1989). Molecular cloning: a laboratory manual. 2nd edition. *Cold Spring Harbor Laboratory Press, New York*.

Sampson TR, Saroj SD, Llewellyn AC, Tzeng Y, Weiss DS. (2013). NIH Public Access **497**:254–257.

Samuel G, Reeves P. (2003). Biosynthesis of O-antigens: Genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly. *Carbohydr Res.* **338**:2503–2519.

Sandeep J. Joseph. (2011). Interplay of recombination and selection in the genomes of *Chlamydia trachomatis. Biology Direct* 2011. **6**: 28.

Santos F, Yarza P, Parro V, Briones C, Anton J. (2010). The metavirome of a hypersaline environment. *Environ Microbiol*. **12**: 2965–2976.

Santos F, Yarza P, Parro V, Meseguer.I.Inmaculada, Rosselló-Móra R, Antón J. 2012). Culture-independent approaches for studying viruses from hypersaline environments. *Appl Environ Microbiol.* **78**: 1635–1643.

Sarkar SF, Guttman DS. (2012. Evolution of the core genome of *Pseudomonas syringae* , a highly clonal, endemic plant pathogen **70**:1999–2012.

Scally M, Schuenzel EL, Stouthamer R, Nunney L. (2005). Multilocus Sequence Type System for the plant pathogen *Xylella fastidiosa* and relative contributions of recombination and point mutation. *Appl Environ Microbiol* **71**:8491–8499.

Scaria J, Mao C, Chen JW, McDonough SP, Sobral B, Chang YF. (2013). Differential stress transcriptome landscape of historic and recently emerged hypervirulent strains of *Clostridium*

difficile strains determined using RNA-seq. *PLoS One*. **8**:1–12.

Schmidt MA, Riley LW, Benz I. (2003). Sweet new world: Glycoproteins in bacterial pathogens. *Trends Microbiol* **11**:554–561.

Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M, Gassmann M, Lightfoot S, Menzel W, Granzow M, Ragg T. (2006). The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol Biol*. **7**:3.

Schultz AK, Zhang M, Leitner T et al. (2006) A jumping profile Hidden Markov Model and applications to recombination sites in HIV andHCV genomes. *BMC Bioinformatics*. 7: 265.

Schumacher M A. (2007). Structural biology of plasmid segregation proteins. *Curr Opin Struct Biol*. **17**: 103–109.

Schuster SC. (2008). Next-generation sequencing transforms today's biology. Nat Methods 5: 16–18.

Paszkiewicz K, Studholme DJ. (2010). De novo assembly of short sequence reads. *Brief Bioinform.* 11: 457–472.

Schwibbert K, Marin-Sanguino A, Bagyan I, Heidrich G, Lentzen G, Seitz H, Rampp M, Schuster SC, Klenk HP, Pfeiffer F, Oesterhelt D, Kunte HJ. (2011). A blueprint of ectoine metabolism from the genome of the industrial producer *Halomonas elongata* DSM 2581 T. *Environ Microbiol.* **13**: 1973–1994.

Scott D, Ely B. (2015).Comparison of genome sequencing technology and assembly methods for the analysis of a GC-Rich bacterial. *Genome.Curr Microbiol*. **70**: 338–344.

Seitz P, Blokesch M. 2013. Cues and regulatory pathways involved in natural competence and transformation in pathogenic and environmental Gram-negative bacteria. *FEMS Microbiol Rev* **37**:336–363.

Sevin EW, Barloy-Hubler F. (2007). RASTA-Bacteria: a web-based tool for identifying toxin-antitoxin loci in prokaryotes. *Genome Biol*. **8**(8): 155.

Shao Y, Harrison E.M, Bi D, Tai C, He X, Ou H,Y, Rajakumar K. and Deng Z. (2011) TADB: a web-based resource for Type 2 toxin-antitoxin loci in *Bacteria* and *Archaea*. *Nucleic Acids Res*. **39**: 606–11.

Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabo G, Polz MF, Alm EJ. (2012. Population genomics of early vents in the ecological differentiation of *Bacteria*. *Science*. **336**: 48–51.

Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G, Polz MF, Alm EJ. (2012). Population genomics of early events in the ecological differentiation of bacteria. *Science*. **336**:48–51.

Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiß S, Sittka A, Chabas S, Reiche K, Hackermüller J, Reinhardt R, Stadler PF, Vogel J. (2010). The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature*. **464**: 250–255.

Sharma PK, Fu J, Zhang X, Fristensky B, Sparling R, Levin DB. (2014). Genome features of *Pseudomonas putida* LS46, a novel polyhydroxyalkanoate producer and its comparison with *other P. putida strains. AMB Express.* **4**: 37.

Sheppard SK, Maiden MC. (2015). The Evolution of *Campylobacter jejuni and Campylobacter coli. Cold Spring Harb Perspect Biol*. 7(8): a018119.

Sheppard SK, McCarthy ND, Falush D, Maiden MCJ. (2008). Convergence of *Campylobacter* species: implications for bacterial evolution. *Science.* **320**:237–239.

Shuman S, Glickman MS. (2007). Bacterial DNA repair by non-homologous end joining. *Nat Rev Microbiol*. **5**:852–861.

Sieuwerts S, Molenaar D, van Hijum S A F T, Beerthuyzen M, Stevens M J A, Janssen P W M, Colin J, de Bok Frank A M, de Vos W M, and van Hylckama Vlieg J E T. (2010). Mixed-culture transcriptome analysis reveals the molecular basis of mixed-culture growth in *Streptococcus thermophilus* and *Lactobacillus bulgaricus. Applied and Environmental Microbiology*, p. 7775–7784.

Sigurgeirsson B, Emanuelsson O, Lundeberg J. (2014). Sequencing degraded RNA addressed by 3' tag counting. *PLoS One*. **9**:e91851.

Silva C, Vinuesa P, Eguiarte LE, Souza V, Martínez-Romero E. (2005). Evolutionary genetics and biogeographic structure of Rhizobium gallicum sensu lato, a widely distributed bacterial symbiont of diverse legumes. Mol Ecol **14**:4033–4050.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM. (2009). ABySS: A parallel assembler for short read sequence data. 1117–1123.

Slater GSC, Birney E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. **6**: 31.

Smillie C, Garcillan-Barcia MP, Francia M V, Rocha EPC, de la Cruz F. (2010). Mobility of Plasmids. *Microbiol Mol Biol Rev***. 74**: 434–452.

Smillie CS, Smith MB, Friedman J, Cordero OX, David L a, Alm EJ. (2011). Ecology drives a global network of gene exchange connecting the human microbiome. *Nature*. **480**:241–4.

Song L, Pan Y, Chen S, Zhang X. (2012). Structural characteristics of genomic islands associated with GMP synthases as integration hotspot among sequenced microbial genomes. *Comput Biol Chem.* **36**:62–70.

Sorek R, Kunin V, Hugenholtz P. (2008). CRISPR--a widespread system that provides

acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol*. **6**: 181–186.

Soria-Carrasco V, Valens-Vadell M, Peña A, Antón J, Amann R, Castresana J, Rosselló-Mora R. (2007). Phylogenetic position of Salinibacter ruber based on concatenated protein alignments. *Syst Appl Microbiol* **30**:171–179.

Sorokin A, Candelon B, Guilloux K, Galleron N, Wackerow-Kouzova N, Ehrlich SD, Bourguet D, Sanchis V. (2006). Multiple-locus sequence typing analysis of *Bacillus cereus* and *Bacillus thuringiensis* reveals separate clustering and a distinct population structure of psychrotrophic strains. *Appl Environ Microbiol.* **72**:1569–1578.

Sourice S, Biaudet V, El Karoui M, Ehrlich SD, Gruss A. (1998). Identification of the Chi site of Haemophilus influenzae as several sequences related to the *Escherichia coli* Chi site. *Mol Microbiol* **27**:1021–9.

Stedman, K.M, She, Q, Phan, H, Holz, I, Singh, H, Prangishvili, D, Garrett, R, and Zillig, W. (2000). pING family of conjugative plasmids from the extremely thermophilic archaeon Sulfolobus islandicus: insights into recombination and conjugation in *Crenarchaeota. J. Bacteriol*. 182, 7014–7020.

Stoddard R A, Miller WG, Foley JE, Lawrence J, Gulland FMD, Conrad P A, Byrne B A. (2007. Campylobacter insulaenigrae isolates from northern elephant seals (*Mirounga angustirostris*) in California. *Appl Environ Microbiol*. **73**:1729–1735.

Szabo G, Preheim SP, Kauffman KM, David L a, Shapiro J, Alm EJ, Polz MF. (2013). Reproducibility of *Vibrionaceae* population structure in coastal bacterioplankton. *ISME J.* **7**:509–19.

Takemura AF, Chien DM, Polz MF. (2014). Associations and dynamics of *Vibrionaceae* in the environment, from the genus to the population level. *Front Microbiol*. **5**:1–26.

Tanabe Y, Kasai F, Watanabe MM. (2007. Multilocus sequence typing (MLST) reveals high genetic diversity and clonal population structure of the toxic cyanobacterium *Microcystis aeruginosa*. *Microbiology.* **153**:3695–3703.

Magoc T, Pabinger S., Canzar S, Liu X, Qi Su Q, Puiu D, Luke J. Tallon and L. Salzberg (2013). GAGE-B: an evaluation of genome assemblers for bacterial Organisms. *Bioinformatics*. **29**(14): 1718–1725.

Tatusova T A, Madden TL. (1999). BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett*. **174**:247–250.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli S V, Crabtree J, Jones AL, Durkin a S, Deboy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan S a, Daugherty SC, Haft DH, Selengut J,

Gwinn ML, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A*. **102**: 13950–13955.

Thaipadungpanit J, Wuthiekanun V, Chierakul W, Smythe LD, Petkanchanapong W, Limpaiboon R, Apiwatanaporn A, Slack AT, Suputtamongkol Y, White NJ, Feil EJ, Day NPJ, Peacock SJ. (2007. A dominant clone of *Leptospira interrogans* associated with an outbreak of human leptospirosis in Thailand. PLoS Negl Trop Dis **1**:1–6.

Thingstad TF, Lignell R. (1997). Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *Aquat Microb Ecol*. **13**: 19–27.

Thomas, C.M. Nielsen, K.M. (2005). Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol*. **3**: 711-721.

Tobiason D M and H Steven Seifert. (2010). Genomic content of *Neisseria* Species. *Journal of Bacteriology*. p. 2160–2168.

Tock, M.R, and Dryden, D.T. (2005). The biology of restriction and anti-restriction. *Curr. Opin. Microbiol*. **8**: 466–472.

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, Karoui M El, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguénec C, Lescat M, Mangenot S, Martinez-Jéhanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z, Ruf C Saint, Schneider D, Tourret J, Vacherie B, Vallenet D, Médigue C, Rocha EPC, Denamur E. (2009. Organised Genome Dynamics in the Escherichia coli Species Results in Highly Diverse Adaptive Paths. *PLoS Genet*. **5**: e1000344.

Traglia GM, Chua K, Centron D, Tolmasky ME, Ramirez MS. (2014). Whole-Genome sequence analysis of the naturally competent *Acinetobacter baumannii* clinical isolate A118. *Genome Biol Evol.* **6**: 2235–2239

Trapnell C, Williams B a, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol. Nature Publishing Group*. **28**(5): 511–515.

Traxler MF, Kolter R. (2012). A massively spectacular view of the chemical lives of microbes. *Proc Natl Acad Sci*. **109**:10128–10129.

Tully BJ, Emerson JB, Andrade K, Brocks JJ, Allen EE, Banfield JF, Heidelberg KB. (2015). De Novo sequences of *Haloquadratum walsbyi* from Lake Tyrrell, Australia, Reveal a Variable Genomic Landscape. *Archaea.* **2015**:1–12.

Tyson GW, Banfield JF. (2007). Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ Microbiol* 1**0**:200-207.

Tyson, G.W, Chapman, J, Hugenholtz, P, Allen, E.E, Ram, R.J, Richardson, P.M, *et al*. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37–43.

Vaisman N and Oren A.. (2009). *Salisaeta longa* gen. nov, sp. nov, a red, halophilic member of the Bacteroidetes. *International Journal of Systematic and Evolutionary Microbiology*, **59**: 2571–2574.

van Wolferen M, Ajon M, Driessen AJM, Albers SV. (2013). How hyperthermophiles adapt to change their lives: DNA exchange in extreme conditions. *Extremophiles*. **17**: 545–563.

Vasu K, Nagaraja V. (2013. Diverse functions of Restriction-Modification systems in addition to cellular defense. *Microbiol Mol Biol Rev.* **77**:53–72.

Ventosa, A, Oren, A, and Ma, Y. (2011). From Genomics to Microevolution and Ecology: The Case of *Salinibacter ruber*. (2011). In Halophiles and Hypersaline Enviroments, *eds Springer-Velag, Berlin*. **pp**. 109-122.

Vergin KL, Tripp HJ, Wilhelm LJ, Denver DR, Rappé MS, Giovannoni SJ. (2007). High intraspecific recombination rate in a native population of *Candidatus pelagibacter ubique* (SAR11). *Environ Microbiol*. **9**:2430–2440.

Vicente M, Mingorance J. (2008). Microbial evolution: The genome, the regulome and beyond. *Environ Microbiol*. **10**:1663–1667.

Voigt K, Sharma CM, Mitschke J, Lambrecht J, Voß B, Hess WR, Steglich C. (2014). Comparative transcriptomics of two environmentally relevant cyanobacteria reveals unexpected transcriptome diversity. *ISME J***. 8:**2056–2068.

Vos M, Didelot X. (2009). A comparison of homologous recombination rates in bacteria and archaea. . *ISME J*. 3(2):199-208.

Vos M, Velicer GJ. (2008). Isolation by distance in the spore-forming soil bacterium *Myxococcus xanthus*. *Curr Biol*. **18**:386–391.

Wang X, Quinn PJ. 2010. Lipopolysaccharide: Biosynthetic pathway and structure modification. *Prog Lipid Res*. **49**: 97–107.

Weinbauer MG. (2004). Ecology of prokaryotic viruses. *FEMS Microbiol Rev*. **28**:127–181.

Westesson O, Holmes I (2009) Accurate detection of recombinant break-points in whole-genome alignments. *PLoS Computional Biology*, 5.

Whitaker RJ. (2005). Recombination shapes the natural population structure of the

hyperthermophilic archaeon *Sulfolobus islandicus*. *Mol Biol Evol*. **22**: 2354–2361.

Whitaker RJ1, Grogan DW, Taylor JW. (2003). Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science*. **301**: 976-978.

Whitehead A, Whitehead A, Crawford DL, Crawford DL. (2006). Neutral and adaptive variation in gene expression. *Proc Natl Acad Sci U S A*. **103**:5425–30.

Wilhelm LJ, Tripp HJ, Givan S a, Smith DP, Giovannoni SJ. (2007). Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data. *Biol Direct.* **2**: 27.

Williams D, Gogarten JP, Papke RT. (2012). Quantifying homologous replacement of loci between haloarchaeal species. *Genome Biol Evol*. **4**: 1223–1244.

Wilmes P, Andersson AF, Lefsrud MG, Wexler M, Shah M, Zhang B, Hettich RL, Bond PL, VerBerkmoes NC, Banfield JF. (2008). Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. *ISME J*. **2**:853–864.

Wilmes P, Simmons SL, Denef VJ, Banfield JF. (2009). The dynamic genetic repertoire of microbial communities. *FEMS Microbiol Rev.* **33**:109–32.

Wilson DJ, Gabriel E, Leatherbarrow AJH, Cheesbrough J, Gee S, Bolton E, Fox A, Hart CA, Diggle PJ, Fearnhead P. (2009). Rapid evolution and the importance of recombination to the gastroenteric pathogen *Campylobacter jejuni*. *Mol Biol Evol* **26**:385–397.

Winter C, Bouvier T, Weinbauer MG y Thingstad TF. (2010). Trade-offs between competition and defense specialist among unicellular planktonic organisms: the "Killing the Winner" hypothesis revisited. *Microbiology and Molecular Biology Reviews.* **74**: 42-57.

Winter J A, Patoli B, Bunting K A. (2012). DNA binding in high salt: analysing the salt dependence of replication protein A3 from the halophile *Haloferax volcanii*. *Archaea*. **2012**:719092.

Woodhams KL, Benet ZL, Blonsky SE, Hackett KT, Dillard JP. (2012). Prevalence and detailed mapping of the gonococcal genetic island in *Neisseria meningitidis*. *J Bacteriol.* **194**: 2275–2285.

Wu Q, Zhang Y, Han L, Sun J, Ni Y. (2009). Plasmid-mediated 16S rRNA methylases in aminoglycoside-resistant *Enterobacteriaceae* isolates in Shanghai, China. *Antimicrob Agents Chemother*. **53**:271–272.

Xu L. (2006. Average gene length is highly conserved in Prokaryotes and Eukaryotes and diverges only between the two kingdoms. *Mol Biol Evol*. **23**: 1107–1108.

Yan Y, Cui Y, Han H, Xiao X, Wong HC, Tan Y, Guo Z, Liu X, Yang R, Zhou D. (2011. Extended MLST-based population genetics and phylogeny of *Vibrio parahaemolyticus* with high levels of recombination. *Int J Food Microbiol* **145**:106–112.

Yang ZH (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*. **24**(8): 1586-1591.

Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL. (2012). Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*. **13**: 134.

Yen-Chun Che1, Tsunglin Liu, Chun-Hui Yu1, Tzen-Yuh Chiang, Chi-Chuan Hwang (2012). Effects of GC Bias in Next-Generation-Sequencing Data on De Novo Genome Assembly. *PLoSOne*. **29**; 8(4).

Yoder-Himes D, Chain P, Zhu Y, Rubin E, Wurtzel O, Tiedje J, Sorek R. (2009). Mapping the *Burkholderia cenocepacia* niche response via high-throughput sequencing. *Proc Natl Acad Sci USA*. **106**:3976–81.

Yu S, Fearnhead P, Holland BR, Biggs P, Maiden M, French N. (2012). Estimating the relative roles of recombination and point mutation in the generation of single locus variants in *Campylobacter jejuni* and *Campylobacter coli. J Mol Evol*. **74**:273–80.

Zerbino DR, Birney E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **18**:821–9.

Zhaxybayeva O, Stepanauskas R, Mohan, N.R. and Papke R.T. (2013) Cell sorting analysis of geographically separated hypersaline environments. *Extremophiles*. **17**: 265-275.

Zimin AV, *et al*. The MaSuRCA genome assembler. (2013). *Bioinformatics*. **29**: 2669–2677.

Introducción

Objetivos

Materiales y métodos

Resultados y discusión

Capítulo 1

Análisis de las diferencias transcripcionales e interacción de cepas cercanas de *S.ruber* mediante RNAseq.

Capítulo 2

Estudio de los mecanismos y estrategias de diversificación genómica en *S. ruber*

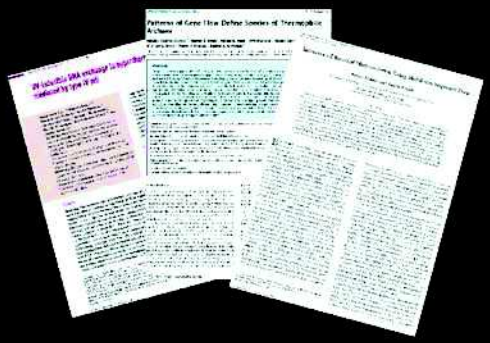
Capítulo 3

Impacto de la recombinación homóloga sobre la evolución de genomas *core* procariotas

Conclusiones

Bibliografía

Anexos



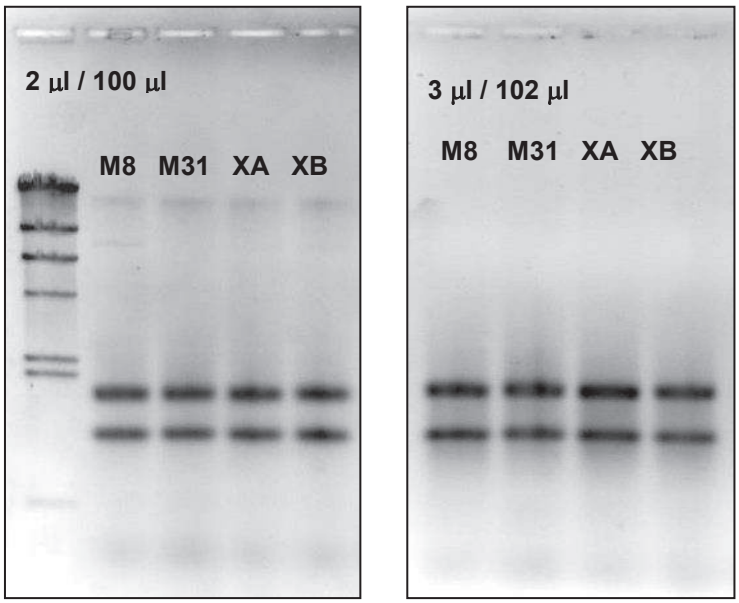


Figura S1.1. Gel de electroforesis en agarosa al 0,8% en el que se muestra la eliminación del DNA genómico tras el tratamiento con la TURBODNAase de Ambión. En los geles se cargaron alícuotas del RNA total extraído antes (izquierda) y después (derecha) de la digestión para las muestras de los cultivos puros de M8, M31 y las réplicas de los cultivos mixtos (XA, XB). El gel de la izquierda muestra una banda de DNA genómico a la altura de 23Kb correspondiente con el DNA genómico.

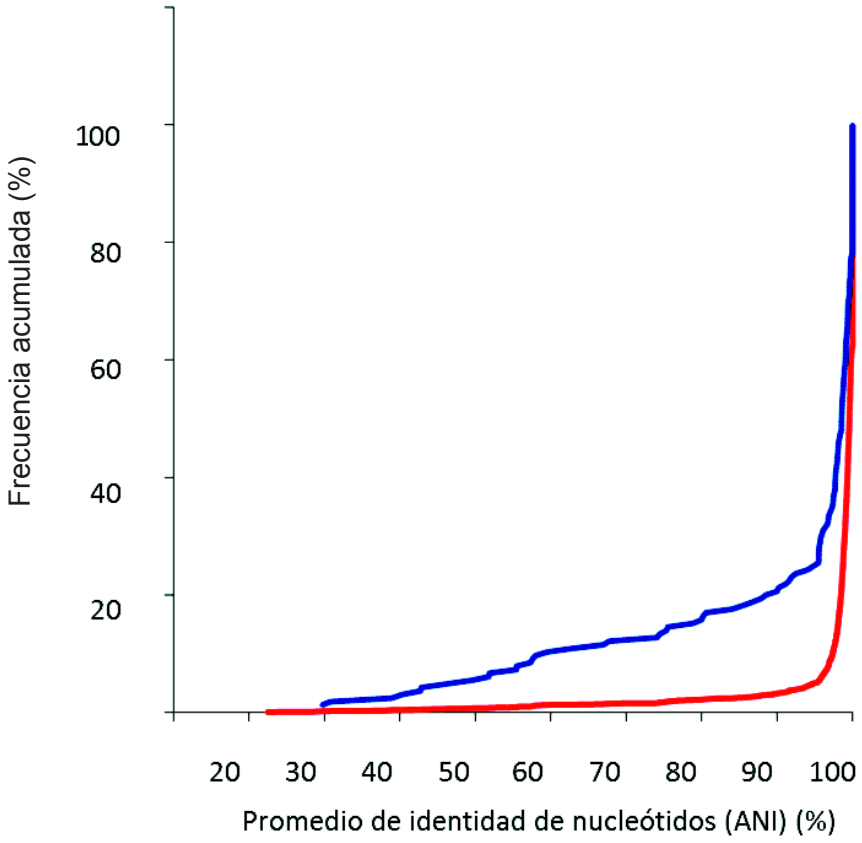


Figura S1.2 Distribución porcentual de genes en base a sus valores de GC y expresión (figura superior) y CAI y expresión (figura inferior). El tamaño de los intervalos se incrementa con los valores de expresión (RPKM) a que niveles superiores contienen un menor número de genes.

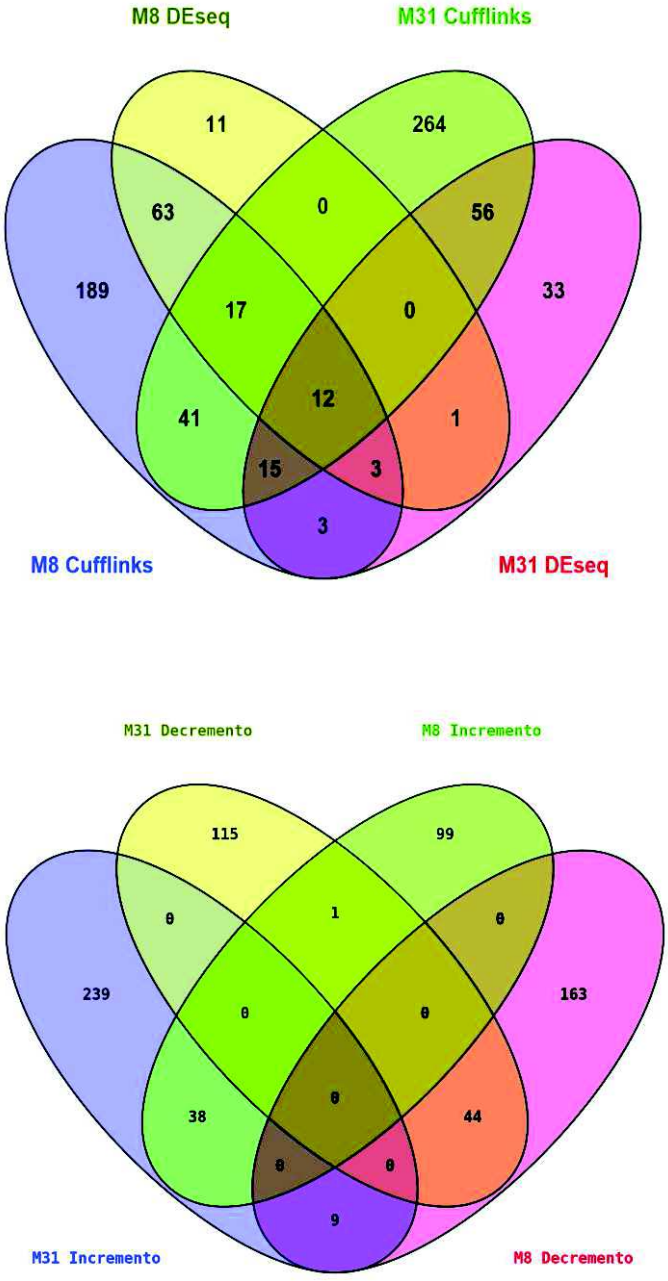


Figura S1.3 Diagramas de Venn que representan:
Figura superior: Número de genes detectados las cepas M8 y M31 de *S.ruber* como expresados diferencialmente al comparar cultivos puros y mixtos por Los programas Cufflinks y DEseq.
Figura inferior: Distribución de genes expresados diferencialmente entre cultivos puros y mixtos según su incremento o disminución en los niveles de expresión para las cepas M8 y M31 de *S.ruber*.

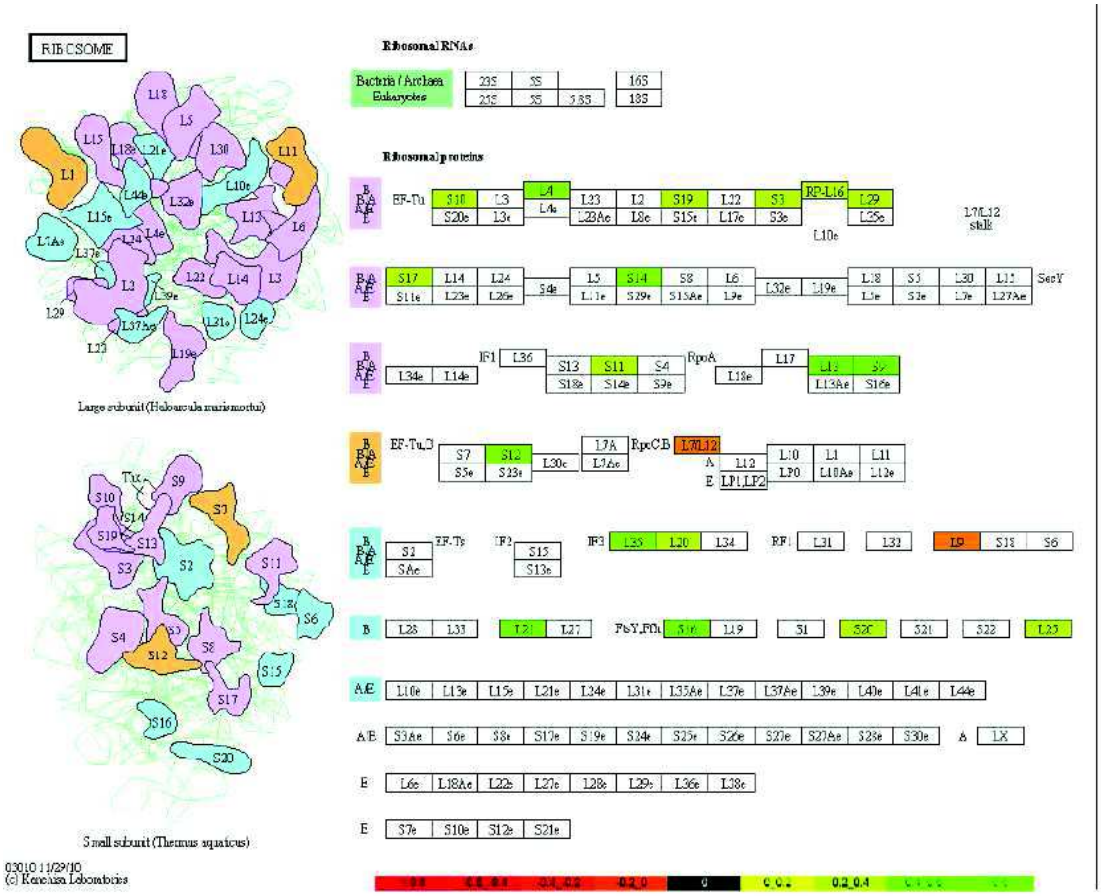


Figura S1.4. Mapa del KEGG en el que se muestran los cambios de expresión diferencial (log_2 fold change) para los genes codificantes de proteínas ribosómicas (CAT) (sru:00020) de la cepa M31 mediante un gradiente porcentual. Las cajas verdes representan los genes que incrementan sus niveles de expresión y las rojas las que los disminuyen.

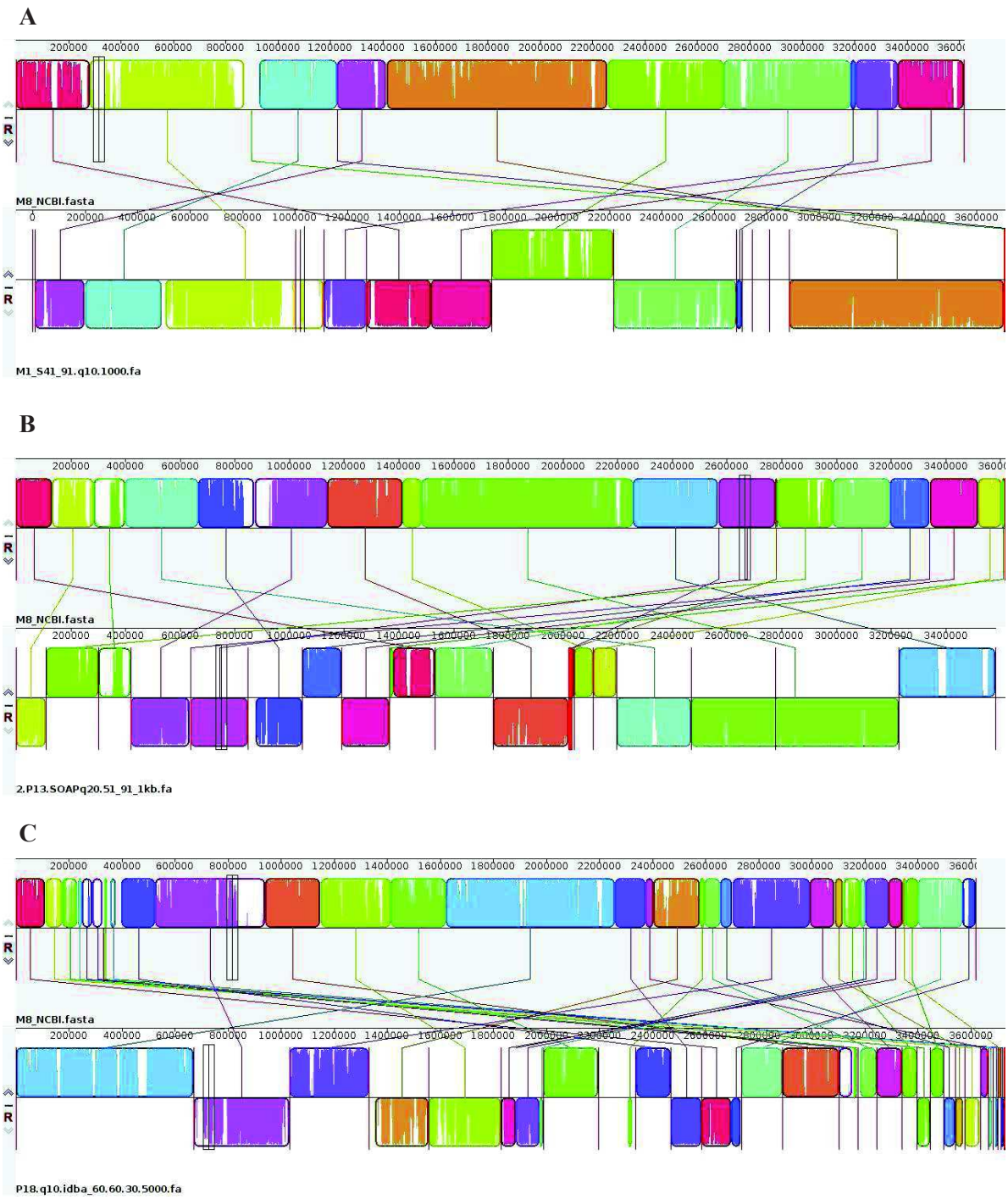


Figura S2.1.A. Detalle del trabajo de reordenamiento de *contigs* del mejor ensamblaje de las cepas *S. ruber* M1 (figura A), P13 (figura B) y P18 (figura C) empleando el programa Mauve. En la figura se muestra el mapeo de los *contigs* de cada una de las cepas (bloques inferiores) contra el cromosoma de referencia de *S. ruber* M8 (bloques superiores en los tres casos). Aquellos *contigs* representados bajo la horizontal se invirtieron, y los que no presentaron sintenia se seleccionaron como candidatos a formar parte de plásmidos.

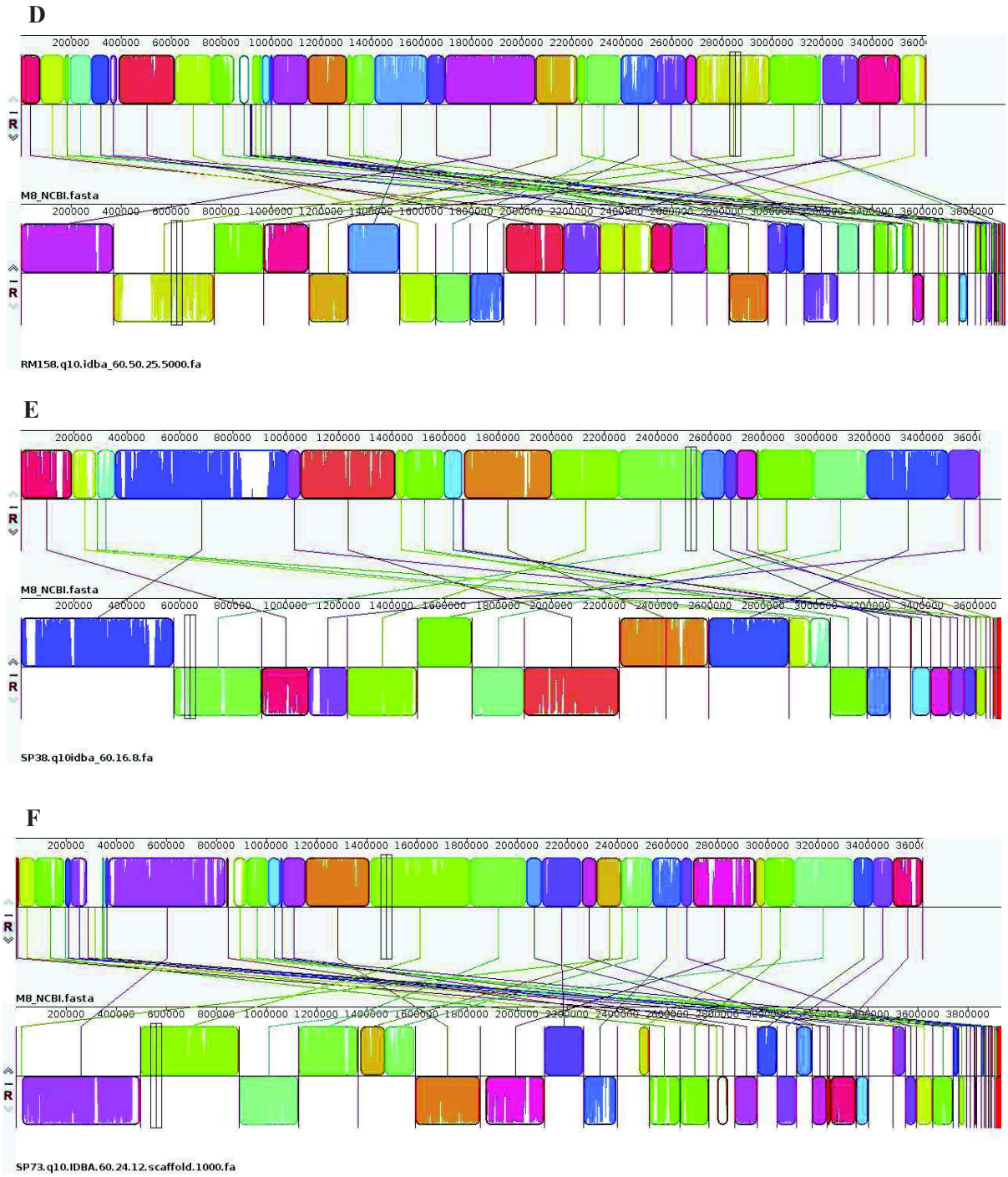
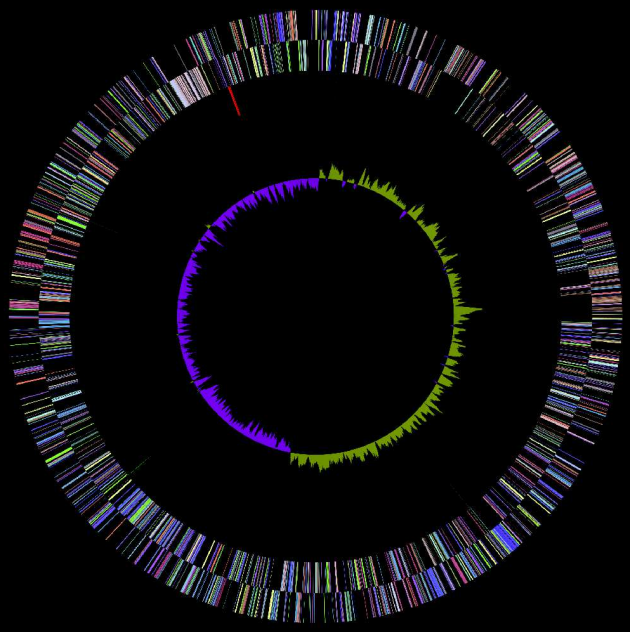


Figura S2.1.B. Detalle del trabajo de reordenamiento de *contigs* del mejor ensamblaje de las cepas *S. ruber* M1 (figura D), P13 (figura E) y P18 (figura F) empleando el programa Mauve. En la figura se muestra el mapeo de los *contigs* de cada una de las cepas (bloques inferiores) contra el cromosoma de referencia de *S. ruber* M8 (bloques superiores en los tres casos). Aquellos *contigs* representados bajo la horizontal se invirtieron, y los que no presentaron sintenia se seleccionaron como candidatos a formar parte de plásmidos.



Universitat d'Alacant
Universidad de Alicante