TESINA PARA OPTAR POR EL GRADO DE LICENCIADO EN CIENCIAS BIOLÓGICAS

¿Cómo evolucionan los ARN pequeños de *Salmonella*?

Lucía Balestrazzi

Orientador: Dr. Andrés Iriarte

Laboratorio de Biología Computacional - Dpto. de Desarrollo Biotecnológico

Instituto de Higiene - Facultad de Medicina - UdelaR

Diciembre 2017

Índice General

1 Resumen

Los ARN pequeños no codificantes (ARNp) son críticos en la regulación post-transcripcional de la expresión génica en bacterias. Sin embargo, las fuerzas que moldean su evolución no son conocidas.

Investigamos la distribución de los ARNp en el género *Salmonella* y encontramos un nivel de conservación general muy alto, señal de su importancia funcional en este género bacteriano. Encontramos también que aquellos ARNp que se integran en grandes redes de regulación son evolutivamente más antiguos que los que regulan un único gen. Además, son significativamente más conservados y evolucionan significativamente más lento que los que regulan un único gen, posiblemente por efecto de una selección purificadora más fuerte.

Rastreamos la historia evolutiva de IsrM, un ARNp importante para la invasividad de células epiteliales y para la supervivencia en macrófago en *S.* Typhimurium, y encontramos que posiblemente se haya originado por rearreglos genómicos dentro del linaje correspondiente a los serovares *S.* Typhimurium, *S.* Heidelberg y *S.* Saint Paul. Esto sugiere que además de la duplicación, la transferencia horizontal, y las pérdidas, los rearreglos genómicos pueden contribuir a las diferencias en los repertorios de ARNp entre cepas bacterianas.

2 Introducción

2.1 Salmonelosis, un problema de alto impacto en salud pública

La salmonelosis es una de las infecciones transmitidas por alimentos más frecuentes, con una estimación global de más de 90 millones de casos de gastroenteritis y 155.000 muertes por año (Majowicz et al., 2010). El género *Salmonella*, cuyo esquema de clasificación taxonómica se muestra en la Fig. 1, consta de sólo 2 especies, *S. enterica* y *S. bongori*. Dicha clasificación fue realizada en base a experimentos de hibridación de ADN, y al igual que sucede con la serotipificación, si refleja las relaciones evolutivas entre los organismos es una cuestión en discusión (Falush et al., 2006; Timme et al., 2013).

La enorme mayoría de los casos de salmonelosis se asocian a serovares de *S. enterica* subespecie *enterica* (Brenner, Villar, Angulo, Tauxe, y Swaminathan, 2000). Si bien hay más de 2500 serovares asignados a esta especie (Issenhuth-Jeanjean et al., 2014), los brotes y epidemias principales se asocian comúnmente con unos pocos serovares dominantes, que muestran variación tanto temporal como geográfica (Betancor et al., 2009).

Los serovares de *S. enterica enterica*, si bien son genéticamente muy similares entre sí, difieren tanto en su potencial epidémico como en el rango de hospederos que son capaces de infectar y en la severidad de las enfermedades que causan (Uzzau et al., 2000). Los de hospedero restringido están estrictamente adaptados a un único hospedero, tales como *Salmonella enterica* ser. Typhi y *Salmonella enterica* ser. Gallinarum, que causan enfermedades exclusivamente en seres humanos y aves,

respectivamente. Los serovares adaptados a hospedero infectan preferentemente una especie, aunque en ocasiones pueden causar enfermedades en otros mamíferos. Un ejemplo de serovar adaptado a hospedero es *Salmonella enterica* ser. Dublin, que suele infectar ganado bovino, pero ocasionalmente es aislado en humanos. Los serovares ubicuos, como *Salmonella enterica* ser. Enteritidis y *Salmonella enterica* ser. Typhimurium, pueden infectar a un amplio espectro de hospederos, incluyendo humanos.

En general, en humanos, los serovares ubicuos causan una gastroenteritis aguda auto-limitada (es decir, que se resuelve por los propios mecanismos de defensa del hospedero), mientras que los restringidos o adaptados a hospedero se asocian frecuentemente a infecciones sistémicas que pueden producir focos extra-intestinales de infección (Uzzau et al., 2000).

El estudio y caracterización de los determinantes genéticos y moleculares que puedan dar una base a estas diferencias patogénicas resulta de gran interés, dado que la salmonelosis constituye un importante problema de salud pública.

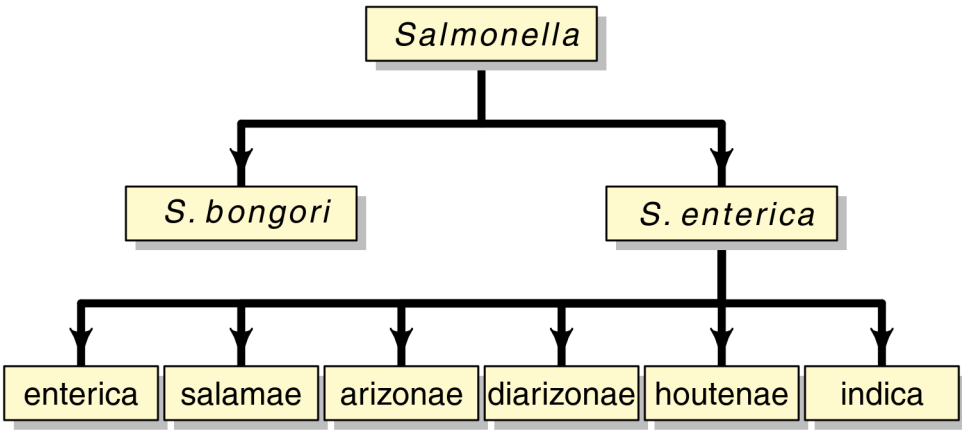

Figura 1: Clasificación de las especies y subespecies del género *Salmonella*.

2.2 ARNp como reguladores de la expresión génica

Durante el curso de la infección en tracto digestivo, *Salmonella* se va encontrando con una serie de ambientes cambiantes a los que debe ser capaz de responder para llevar a cabo una infección exitosa. Después de la ingestión de alimentos contaminados, *Salmonella* pasa por el estómago y luego al íleon distal, donde se asocia con el revestimiento epitelial, se adhiere y penetra en la membrana apical de las células M en las placas de Peyer. Las cepas de *Salmonella* que atraviesan la membrana basolateral son engullidas por macrófagos (Padalon-Brauch et al., 2008). La capacidad de *Salmonella* de adaptarse por ejemplo al ambiente ácido del estómago y al entorno intracelular del macrófago contribuye en gran medida a su virulencia. *Salmonella* puede sensar su entorno y adaptarse rápidamente a las condiciones cambiantes, un proceso que está mediado por la regulación de la expresión génica en los niveles transcripcional, post-transcripcional y traduccional. Los principales mediadores en este proceso de adaptación son los factores de transcripción y las proteínas asociadas a nucleoide, así como los ARN pequeños no codificantes (ARNp).

Si bien la regulación mediada por ARN fue descubierta recientemente (Fire et al., 1998; Tabara, Grishok, y Mello, 1998) el concepto de ARN como regulador de la expresión génica se remonta al modelo de operón (Jacob y Monod, 1961), publicado hace más de 50 años, que proponía la existencia de genes reguladores, que podían actuar como ARN o como proteínas.

Los ARNp son moléculas de entre 50 y 500 nucleótidos de longitud, que regulan la expresión génica mediante la unión por complementariedad de bases a un ARN mensajero (ARNm) expresado de forma independiente (Hébrard et al., 2012; Storz,

Vogel, y Wassarman, 2011; Updegrove, Shabalina, y Storz, 2015).

La acción de los ARNp en coordinación con factores de transcripción para reprimir o potenciar de forma simultánea o recíproca la expresión génica tanto a nivel transcripcional como post-transcripcional coloca a los ARNp en grandes redes de regulación (revisado en Storz et al., 2011). La incorporación de los ARNp en dichas redes permite una regulación mas fina con respecto a la fuerza y velocidad de la respuesta (Kacharia, Millar, y Raghavan, 2017; Updegrove et al., 2015).

En una enterobacteria con un genoma promedio de entre 4 y 5 millones de pares de bases, se estima que hay 300-400 ARNp (Gottesman y Storz, 2011; Kröger et al., 2013; Raghavan, Groisman, y Ochman, 2011), cada uno normalmente controlando múltiples ARNm blancos que actúan en procesos interrelacionados, influyendo en diferentes aspectos de la fisiología bacteriana, su virulencia y comportamiento (Kacharia et al., 2017; Storz et al., 2011). Ese número podría aumentar con el reciente descubrimiento de nuevas clases de ARNp que habían sido pasados por alto, como los producidos a partir de las regiones 3 prima no traducidas (3' UTR), sea por procesamiento del ARNm o por transcripción a partir de promotores internos (Miyakoshi, Chao, y Vogel, 2015). La mayor parte de los ARNp son considerados "huérfanos", ya que no está identificada su red de interacción (Fröhlich, Haneke, Papenfort, y Vogel, 2016).

Estudios funcionales han demostrado la importancia de los ARNp en la virulencia y patogenicidad de *Salmonella* (Gong et al., 2011; Hébrard et al., 2012; Padalon-Brauch et al., 2008). Por todo lo anterior, se ha postulado que las diferencias en el repertorio de ARNp entre bacterias estrechamente relacionadas tengan un impacto importante en su fisiología y potencial patogénico.

2.3 Formas de regulación mediadas por ARNp

La regulación post-transcripcional por ARNp ocurre de diversas maneras. Se han identificado dos clases distintas de ARNp: los codificados en *trans* que se transcriben desde regiones intergénicas del genoma, y los codificados en *cis*, también conocidos como ARN anti-sentido (ARNas), que se codifican en la hebra complementaria a las secuencias codificantes o las regiones no traducidas de los transcriptos. Mientras que los ARNas tienen una región de complementariedad perfecta con su ARNm blanco, los ARNp codificados en *trans* generalmente poseen pequeñas regiones de complementariedad limitada, y generalmente necesitan de la ayuda de la proteína chaperona Hfq para estabilizar la unión. Hfq ayuda también a aumentar la concentración local de ARNp y ARNm al tener sitios de unión para ambos elementos (Brennan y Link, 2007). El apareamiento ocurre a través de unos pocos nucleótidos expuestos en las regiones en forma de horquilla del regulador, el ARNm o ambos. Después de esta primera interacción, el apareamiento se puede extender, a menudo con reordenamientos en la estructura secundaria de ARN.

Muchos de los ARNp se aparean en el sitio de unión al ribosoma o cerca del mismo, bloqueando así la traducción al evitar la unión del ribosoma. En la mayoría de los casos donde se bloquea la unión, también se observa una disminución asociada en la estabilidad del ARNm. Los ARNp también pueden activar la traducción, en muchos casos al evitar la formación de una estructura secundaria inhibitoria, en otros casos aumentando la estabilidad del ARNm, o ambos. La unión del ARNp conduce a la remodelación de la estructura del ARNm, descubriendo el sitio de unión al ribosoma y permitiendo así la traducción (Gottesman y Storz, 2011).

2.4 Elementos que requiere un ARNp funcional

El estudio de un número creciente de ARNp ha permitido notar que varios elementos deben estar presentes para que estos se puedan expresar, para que sean estables, y capaces de aparearse con un ARNm.

Un punto en común de todos los ARNp es que su transcripción está fuertemente regulada (Updegrove et al., 2015), lo que requiere de la presencia de un promotor regulado que responda a las señales del ambiente. Además, casi todos los ARNp conocidos cuentan con un terminador Rho independiente, que consiste en una horquilla estable seguida de una región poli-U. Esta característica es utilizada en muchos de los algoritmos que predicen ARNp. Otra característica esencial de los ARNp es un sitio donde se inicia el apareamiento de bases con el ARNm, frecuentemente llamado región *seed*, una región que suele implicar unos pocos nucleótidos y que es la parte más conservada del ARNp (Peer y Margalit, 2011). Los sitios de unión para Hfq son otra característica fundamental de la mayoría de los ARNp de las enterobacterias. Hfq protege a los ARNp de la degradación por ribonucleasas y facilita la interacción con el ARNm blanco al aumentar la concentración local de ambas moléculas (Brennan y Link, 2007). Por último, los ARNp poseen regiones de doble hebra que ayudan a estabilizar la molécula y facilitan la orientación de la región *seed* y los sitios de unión para Hfq (Updegrove et al., 2015). Aunque las secuencias específicas de estas regiones pueden no conservarse, están enriquecidas en nucleótidos que co-varían para retener el apareamiento de bases intramolecular y así mantener la estructura, permitiendo que las regiones de simple hebra estén accesibles para las distintas interacciones.

2.5 Islas de patogenicidad de *Salmonella*

Salmonella también debe su virulencia en gran parte a sus islas de patogenicidad (SPI, por *Salmonella Pathogenicity Island*). Análisis comparativos de genomas de cepas patógenas de *Salmonella* con el de cepas comensales como *Escherichia coli* revelaron la presencia de muchas inserciones presentes sólo en las cepas patógenas (Padalon-Brauch et al., 2008). Estas inserciones o islas varían en tamaño desde genes únicos hasta grandes islas de decenas de miles de pares de bases.

Las cepas de *Salmonella* poseen SPI-1 y SPI-2, que codifican dos sistemas de secreción de tipo III (SST3) separados. El SST3 forma un complejo en forma de aguja que inyecta proteínas efectoras en el citosol de las células del hospedador eucariota. Los efectores de SPI-1 incluyen SipA, SipB, SipC, SopA, SopB, SopD, SopE y SopE2. Estos efectores trabajan coordinadamente para reorganizar el citoesqueleto de actina del hospedador, facilitando la invasión de células epiteliales por parte de *Salmonella* (Agbor y McCormick, 2011). Por el contrario, los efectores SPI-2 son responsables de la replicación de *Salmonella* dentro de las células fagocíticas, promoviendo la supervivencia bacteriana y la infección sistémica (Schatten y Eisenstark, 2007).

Además de codificar los SST3, las SPI codifican decenas de ARNp, muchos de los cuales se expresan diferencialmente en condiciones que se asemejan a las que enfrenta la bacteria durante el curso de la infección (Padalon-Brauch et al., 2008).

La importancia de los ARNp codificados en SPI para la virulencia y patogenicidad de *Salmonella* ha sido demostrada tanto *in vitro* como *in vivo* (Gong et al., 2011; Padalon-Brauch et al., 2008).

2.6 Evolución de los ARNp

A pesar de la reconocida importancia de los ARNp para la fisiología de las bacterias, poco se conoce de los mecanismos que moldean su evolución (Gottesman y Storz, 2011; Kacharia et al., 2017; Peer y Margalit, 2014; Skippington y Ragan, 2012; Updegrove et al., 2015).

Un análisis de la conservación de los ARNp en *E. coli* y *Shigella* mostró que la variación en el repertorio entre cepas se debe principalmente a eventos de deleción (Skippington y Ragan, 2012). Otro estudio reveló que la mayoría de los ARNp de *E. coli* se originaron después de la separación de las enterobacterias de otras gammaproteobacterias (Peer y Margalit, 2014). Esto habla de un ciclo dinámico de nacimiento y pérdida de genes codificantes para ARNp. Además de este ciclo de nacimiento y pérdida, los ARNp evolucionan a tasas más rápidas que los genes codificantes para proteínas, lo que dificulta identificar ARNp homólogos en bacterias filogenéticamente distantes (Kacharia et al., 2017). Una familia de bacterias que se ha demostrado que está a la distancia evolutiva óptima para el análisis evolutivo de los ARNp es la familia Enterobacteriaceae (Lindgreen et al., 2014), que incluye entre otros a *E. coli* y *S. enterica*, que por ser organismos modelo, se han caracterizado extensamente sus repertorios de ARNp (ver por ejemplo, Kröger et al., 2013; Perkins et al., 2009; Raghavan et al., 2011). Las características patogénicas de *S. enterica* y los conocimientos acumulados sobre sus mecanismos de virulencia, evolución del genoma y muchas vías de expresión génica y metabolismo, hacen que esta especie sea un modelo especialmente atractivo para el estudio de la evolución de los ARNp.

3 Hipótesis

Existen diferencias en la forma de evolución de los ARNp que se explican al menos en parte por el tamaño de la red de regulación en la que participan.

4 Objetivo general

Estudiar las fuerzas que moldean la evolución de los ARNp dentro y entre serovares de *Salmonella*.

5 Objetivos específicos

- Generar un muestreo de la diversidad genómica de cepas de *Salmonella* existente en la base de datos Enterobase.
- Estudiar el patrón de distribución de los ARNp en la filogenia.
- Estudiar la tasa de evolución molecular de los ARNp, al menos para un subgrupo cuyas redes de regulación estén bien estudiadas.

6 Materiales y métodos

6.1 Secuencias nucleotídicas y datos de interacción

Todos los genomas fueron obtenidos de la base de datos EnteroBase, disponible en http://enterobase.warwick.ac.uk, a excepción de algunos de los genomas utilizados para el estudio de genómica comparativa de IsrM que fueron obtenidos de la base de datos Genbank del NCBI (Benson, Karsch-Mizrachi, Lipman, Ostell, y Wheeler, 2005). Para representar la mayor diversidad posible dentro del género se eligieron como máximo 5 genomas al azar de cada secuenciotipo (ST) reportado por tipificación multilocus de secuencias (MLST) (Achtman et al., 2012), donde cada secuenciotipo representa una combinación única de alelos de siete genes *housekeeping*. Se obtuvo un total de 5762 genomas.

El serovar y subespecie de cada genoma fueron predichos *in silico* con la plataforma SISTR 1.0.2 (C. E. Yoshida et al., 2016), que identifica el serovar basándose en la predicción de la fórmula antigénica. SISTR examina los genes altamente variables *wzx* y *wzy* de la región *rfb* para determinar el antígeno somático, y los genes *fliC* y *fljB* para determinar los antígenos flagelares H1 y H2, respectivamente. La fórmula antigénica predicha se consulta contra una base de datos para asignar un serovar. Además, SISTR analiza 330 genes conservados (cgMLST) para proveer un contexto filogenético, donde el serovar predominante en el *cluster* de cgMLST es usado para elegir el serovar más probable entre los potenciales (Yachison et al., 2017). Tras en análisis en SISTR se descartó uno de los genomas ya que ninguno de los alelos fue encontrado en la base de datos.

El conjunto de ARNp utilizado como referencia es la combinación de los ARNp

descriptos y validados experimentalmente en *S.* Typhimurium y *S.* Typhi por Kröger et al. (2013) y Perkins et al. (2009), respectivamente. Los ARNp idénticos fueron agrupados realizando un análisis de *clusters* con el programa CD-HIT versión 4.5.7 (Fu, Niu, Zhu, Wu, y Li, 2012), resultando un total de 572 *clusters*.

La información de las interacciones validadas experimentalmente para 30 ARNp fue obtenida de la base de datos sRNATarBase 3.0 (J. Wang et al., 2015).

6.2 Filogenia

Para la reconstrucción filogenética se utilizaron únicamente genomas con un promedio de cobertura mayor a 30 (según reportado en la base de datos EnteroBase), y se seleccionó un genoma al azar de cada ST, resultando un total de 1449 genomas. Para obtener el conjunto de genes de copia única presentes en todos los genomas se partió del pangenoma de *Salmonella* según descripto en Jacobsen et al. (2011), compuesto de 12722 genes, que fueron utilizados como *query* para realizar una búsqueda con el programa tBLASTn del paquete BLAST v2.2.31+ (Altschul, Gish, Miller, Myers, y Lipman, 1990) contra la base de datos local. Los 619 genes de copia única identificados en todos los genomas (e-value $>$ 1e-03, identidad $>$ 70% y cobertura $>$ 80%) fueron seleccionados para continuar el análisis. Se extrajo su secuencia nucleotídica y se tradujo a aminoácidos con la función seq_reformat de T-COFFEE v11.00 (Notredame, Unpublished). Se alineó la secuencia aminoacídica de cada gen con el programa MUSCLE v3.8.31 (Edgar, 2004) (3 iteraciones), y se volvió a la secuencia nucleotídica. La reconstrucción filogenética se realizó mediante dos estrategias distintas: la de concatenación y la de árboles de genes.

6.2.1 *Filogenia por concatenación de genes*

Para esta aproximación se utilizó el programa RAxML v8.2.9 (Stamatakis, 2014) ya que permite paralelizar el trabajo en múltiples procesos, haciendo uso de todos los recursos del servidor. Como RAxML sólo está optimizado para la familia de modelos GTR (el más general de los modelos reversibles en el tiempo), no se trató de elegir el modelo que mejor ajustara a los datos. La aproximación a nivel del genoma completo hace que el riesgo de sobreparametrización de los datos, que podría asociarse al modelo GTR por su gran cantidad de parámetros, sea muy bajo. Se extrajeron los bloques de alineamiento confiable de los 619 ortólogos con el programa GBLOCKS 0.91b (Castresana, 2000), utilizando los parámetros por defecto. Se concatenó su secuencia, y se extrajeron los polimorfismos de un solo nucleótido (SNPs) con el programa snp-sites 2.3.2 (Page et al., 2016), obteniéndose un alineamiento de 126412 sitios variables. Las zonas que han estado bajo influencia de eventos de recombinación fueron excluidas del alineamiento con el programa Profile del paquete PhiPack (T. Bruen y Bruen, 2005) que implementa el test Phi, que estima la probabilidad de observar el alineamiento bajo la hipótesis nula de que no hubo recombinación.

Se obtuvo un alineamiento final de 96743 sitios variables, sobre el cual se realizó la inferencia filogenética, utilizando el método de máxima verosimilitud (ML, por *Maximum Likelihood*) con el modelo GTRCAT implementado en RAxML.

Brevemente, el método de ML estima la verosimilitud de observar los datos (en este caso, el alineamiento de secuencias) para un modelo de sustitución dado y para cada topología de árbol en particular. La tarea de encontrar la topología de árbol que maximice la verosimilitud de los datos es un problema NP-complejo

debido a la gran cantidad de topologías posibles incluso para un número moderado de individuos, por lo que se resuelve heurísticamente.

El modelo GTR (Tavaré, 1986) o general tiempo-reversible permite frecuencias nucleotídicas desiguales y tasas independientes para cada una de las seis posibles rutas de sustitución. Para permitir la variación de las tasas de sustitución se puede aplicar la distribución Gamma (+G), que describe la heterogeneidad de las tasas en los distintos sitios. La distribución continua se aproxima con una distribución discreta que es computacionalmente manejable y los sitios se dividen en k categorías de tasas equiprobables. Un solo parámetro α gobierna la forma de esta distribución. El modelo GTRCAT es una aproximación al modelo GTR + G mediante un algoritmo computacionalmente más eficiente, con optimización de las tasas de sustitución individuales por sitio y clasificación de esas tasas en 25 categorías. La opción de corrección de sesgo de largo de ramas de Stamatakis fue aplicada según recomendado por los autores para SNPs y otros datos que incluyen sólo sitios variables (Leaché, Banbury, Felsenstein, Oca, y Stamatakis, 2015).

Debido al alto costo computacional que implica la realización de réplicas de bootstrap, ya que en teoría cada réplica debe analizarse de la misma manera que los datos originales, se utilizó la aproximación sugerida por Kishino y Hasegawa (1990), donde su utilizan las estimaciones de máxima verosimilitud de largo de ramas y los parámetros de sustitución de los datos originales para calcular los valores de log-verosimilitud de cada réplica. Simulaciones computacionales sugieren que esta aproximación, conocida como RELL bootstrap, proporciona una buena aproximación al bootstrap real de Felsenstein (Minh, Nguyen, y Haeseler, 2013). Se realizaron 1000 réplicas de RELL bootstrap.

6.2.2 *Filogenia a partir de árboles de genes*

Para la segunda aproximación, luego de obtener los bloques de alineamiento confiables con GBLOCKS, se detectaron los genes recombinantes mediante el test Phi implementado en PhiPack (T. Bruen y Bruen, 2005). Se descartaron los 195 genes en los que se identificaron eventos de recombinación, con un *p-value* para el test Phi < 0.05. Para los 424 ortólogos restantes, se reconstruyó un árbol para cada gen, nuevamente utilizando el método de ML con el modelo GTRCAT implementado en RAxML.

La filogenia de especies fue inferida teniendo en cuenta las discordancias entre los árboles de especies y los árboles de genes debidas al reparto incompleto de linajes (ILS, por *incomplete lineage sorting*), que pueden complicar la reconstrucción filogenética si no son debidamente consideradas. Para ello se utilizó el programa ASTRAL-III (C. Zhang, Sayyari, y Mirarab, 2017), que implementa un método para inferir árboles de especies a partir de árboles de genes mediante el modelo del coalescente para múltiples especies (MSC), que modela cómo los alelos segregan en las poblaciones. Bajo el modelo MSC, los árboles de genes pueden ser distintos del árbol de especies en presencia de ILS. Un árbol de cuartetos o *quartet tree* es un árbol de 4 hojas. Cualquier árbol binario sin raíz está definido por su conjunto de árboles de cuarteto. Se ha demostrado que para todos los árboles de cuarteto con longitudes de ramas en unidades coalescentes, el árbol de genes sin raíz más probable es topológicamente idéntico al árbol de especies de esas 4 hojas sin raíz (Allman, Degnan, y Rhodes, 2011). ASTRAL computa los árboles de especies de cuartetos que definen el conjunto de árboles de genes de entrada y trata de encontrar el árbol de especies que maximiza el *quartet score*. El *quartet score* de

un árbol de especies con respecto a un conjunto de árboles de genes es la suma del número de topologías de árboles de cuarteto inducidas que son compartidas entre el árbol de especies y cada árbol de genes. El problema de encontrar el árbol de especies con el máximo *quartet score* con respecto al conjunto de árboles de genes de entrada, es conocido como el problema del árbol de especies con máximo soporte de cuartetos (MQSST, por *Maximum Quartet Support Species Tree*). Este problema es NP-complejo, por lo que ASTRAL lo resuelve heurísticamente definiendo una versión restringida de el problema MQSST. Para este análisis se realizaron 1000 réplicas de bootstrap.

6.3 Patrón filogenético de los ARNp

Se creó una base de datos local de BLAST con los 5761 genomas muestreados. Se realizó una búsqueda con el programa BLASTn del paquete BLAST v2.2.31+ (Altschul et al., 1990) con un conjunto de 572 ARNp como *query* contra la base de datos local creada (e-value $>$ 1e-03, identidad $>$ 90% y cobertura $>$ 90%). Dado que dentro de los ARNp de referencia hay secuencias con un porcentaje de identidad mayor al umbral, es de esperar que dentro de los resultados existan distintos *hits* que apuntan a sitios solapantes. El resultado de BLAST fue filtrado con un script desarrollado localmente para que cada sitio apunte únicamente al mejor *hit*.

Se creó una matriz de datos de presencia/ausencia de ARNp en cada genoma, y una matriz de porcentaje de aparición de los ARNp dentro de cada serovar (para esta última se incluyeron únicamente aquellos serovares representados por más de 5 genomas). Se realizó un *heatmap* a partir de cada matriz con la función

heatmap.2 del paquete gplots v3.0.1 (Warnes et al., 2016) en la plataforma R v3.4.2 (R Core Team, 2017), usando la información de la correlación de Spearman. El *clustering* jerárquico fue realizado con el método completo con la función hclust implementada en R.

6.4 Distancia genética y frecuencia de aparición de los ARNp

Se extrajo de cada genoma de la base de datos local la secuencia nucleotídica de los 30 ARNp de los que existe información validada experimentalmente de su red de regulación, según reportado en la base de datos sRNATarBase (J. Wang et al., 2015). Se alinearon las secuencias con el programa MUSCLE v3.8.31 (Edgar, 2004) (2 iteraciones). Las distancias fueron obtenidas con la función dist.dna del paquete APE v4.1 (Paradis, Claude, y Strimmer, 2004) en la plataforma R (R Core Team, 2017), usando el modelo Jukes-Cantor. Las diferencias fueron evaluadas mediante el test no paramétrico de Wilcoxon no pareado, implementado en la plataforma R.

Se obtuvieron los datos de la conservación de estos 30 ARNp en 27 genomas de *E. coli* de bibliografía existente (Padalon-Brauch et al., 2008; Skippington y Ragan, 2012).

6.5 Genómica comparativa de la región del gen isrM

El genoma de *S.* Typhimurium SL1344 (número de acceso en NCBI FQ312003.1) fue comparado usando el programa BLASTn del paquete BLAST v2.2.31+ (parámetros por defecto) contra los genomas de *S. enterica* subsp. *arizonae* ser. 62:z4,z23:–

(número de acceso en NCBI CP000880.1), *S.* Typhi CT18 (número de acceso en NCBI NC_003198.1), *S. enterica* subsp. *salamae* ST2274 (código de ensamblado en EnteroBase SAL_CA3169AA_AS) y *S. enterica* subsp. *houtenae* ser. 50:z4,z23:- (código en EnteroBase SAL_LA6190AA_AS). Los resultados fueron visualizados con el paquete genoPlotR (Guy, Roat Kultima, y Andersson, 2010) en la plataforma R.

La fórmula antigénica del genoma correspondiente *S. salamae* no pudo ser determinada con precisión mediante serotipificación *in silico*, por lo que nos referimos a él por su secuenciotipo.

7 Resultados

7.1 Filogenia

1449 genomas, pertenecientes cada uno a uno a un secuenciotipo distinto, fueron utilizados para realizar una realizar una reconstrucción filogenética del género mediante dos estrategias distintas.

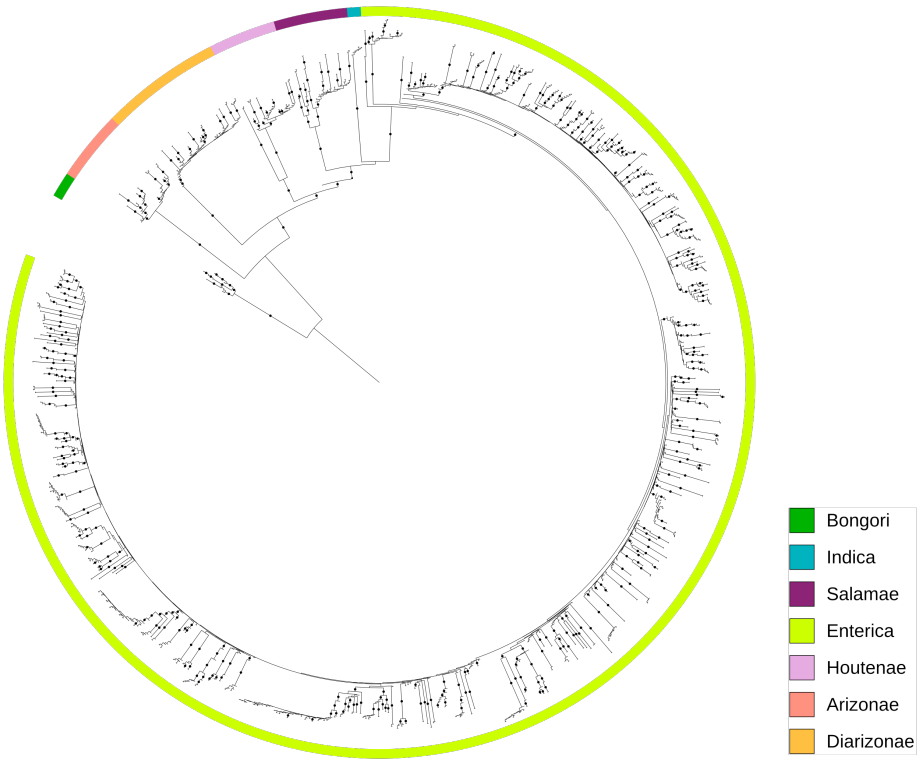

Figura 2: Filogenia reconstruída con ASTRAL-III a partir de los árboles de genes construídos con RAxML bajo el modelo GTRCAT. Los puntos representan valores de bootstrap mayores a 80.

De acuerdo al análisis de ASTRAL-III (Fig. 2), cada subespecie representa un grupo monofilético. *S. bongori* se separa primero del ancestro común de todas

las subespecies de *S. enterica*. El siguiente en separarse es *arizonae*, seguido del cercanamente emparentado *diarizonae*, luego *houtenae*, *salamae*, y por último se separan *indica* y *enterica*. El árbol resultante es asimétrico en su forma: el ancestro de *S. enterica* participa en todos los eventos de diversificación.

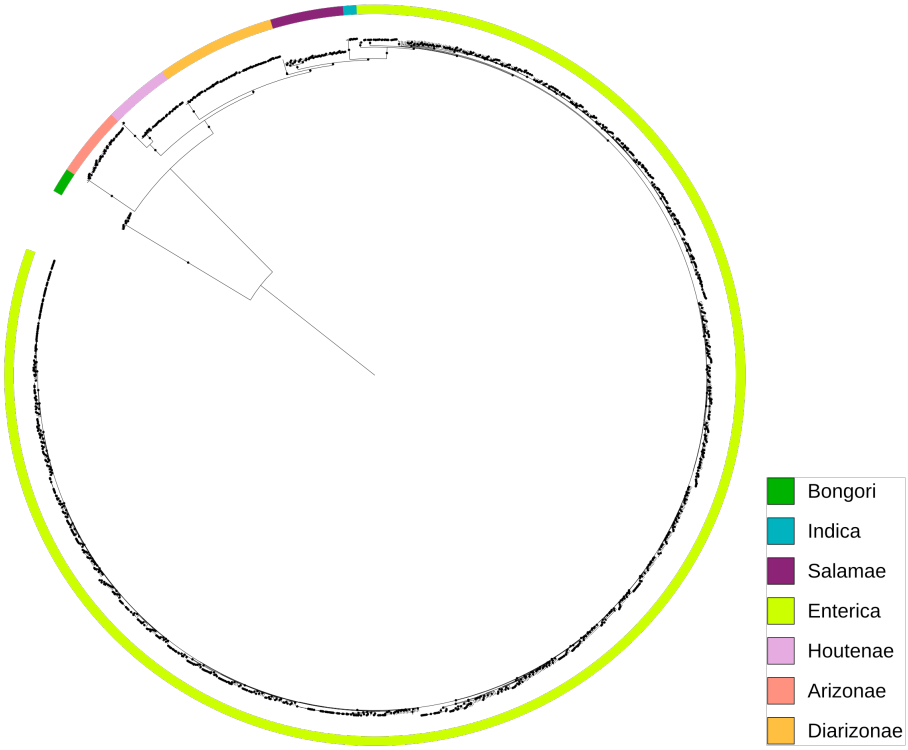

Figura 3: Filogenia reconstruída con RAxML sobre el alineamiento de genes concatenados. Los puntos representan valores de bootstrap mayores a 80.

Resultados similares se obtuvieron mediante la estrategia de concatenación de genes (Fig. 3). Si bien el patrón de ramificación no es el mismo en todos los casos, también en este árbol se verifica la monofilia de los grupos y la asimetría del árbol. En esta reconstrucción los grupos cercanamente emparentados *arizonae* y *diarizonae* no aparecen como grupos hermanos.

7.2 La subespecie *enterica* es la más diversa del género

Debido a la enorme cantidad de genomas disponibles en la base de datos al momento de comenzar el análisis, un total de 80142, el número de genomas por ST fue topeado en un máximo de 5 al realizar el muestreo. Los 5761 genomas muestreados correspondientes a 2525 ST distintos fueron asignados por serotipificación *in silico* a 554 serovares. Un 54% de los ST tienen un único genoma asignado en la base de datos, al igual que un 15% de los serovares. Además, el 57% de los serovares tiene asignados 5 genomas o menos.

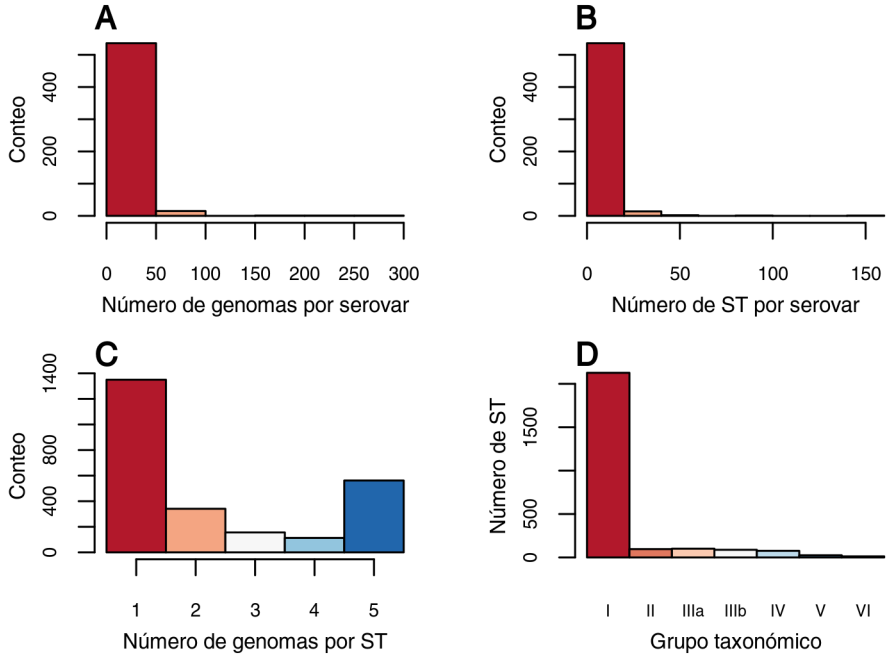

Figura 4: Diversidad del muestreo. **A** Histograma del número de genomas por serovar. **B** Histograma del número de ST por serovar. **C** Número de genomas por ST. **D** Número de ST por grupo taxonómico. Los grupos I, II, IIIa, IIIb, IV, V y VI corresponden a *enterica, salamae, arizonae, diarizonae, houtenae, bongori,* e *indica*, respectivamente.

Como ya ha sido reportado (ver por ejemplo, Falush et al., 2006; Timme et al., 2013), la subespecie *S. enterica enterica* contiene la mayor diversidad del género, con un 84% de los ST y un 73% de los serovares reportados (Fig. 4).

A nivel de serovares, *S.* Typhimurium y *S.* Enteritidis, correspondientes ambos a la subespecie *enterica*, representan entre los dos el 10% de los ST de la base de datos y el 9% de los genomas del muestreo.

7.3 Un 55% de los ARNp son comunes a todos los genomas de *S. enterica enterica*

Se identificaron los homólogos de 560 ARNp en 5761 genomas pertenecientes al género *Salmonella* usando BLASTn. Esta estrategia ha resultado efectiva para identificar ARNp homólogos dentro de la familia Enterobacteriaceae (Kacharia et al., 2017; Peer y Margalit, 2014; Skippington y Ragan, 2012). Para visualizar el patrón de distribución de los ARNp se realizó un *heatmap* a partir de los datos de presencia/ausencia en cada genoma (Fig. 5) y otro a partir del porcentaje de conservación de los ARNp dentro de cada serovar (Fig. 6).

Identificamos un conjunto de 305 ARNp presentes en al menos el 95% de los genomas de *S. enterica enterica*, 165 ARNp presente en al menos el 95% de los genomas del género *Salmonella*, 50 ARNp presentes en el 100% de los genomas analizados, y 255 ARNp cuya presencia es variable (Figs. 5 y 6). Entre los ARNp variables dentro de *S. enterica enterica*, 109 ARNp aparecen con alta frecuencia (50-95% de los genomas), 89 aparecen con baja frecuencia (5-50% de los genomas), y 57 ARNp aparecen apenas esporádicamente, en un 5% o menos de los genomas. Algunos de estos ARNp no sólo varían entre subespecies, sino que su presencia

también varía dentro de un mismo serovar, aunque es probable que en algunos casos esto refleje la no monofilia de algunos serovares (es decir, la existencia de serovares que no representan grupos naturales). El alto nivel de conservación general de los ARNp en la filogenia señala su importancia funcional en este grupo de bacterias.

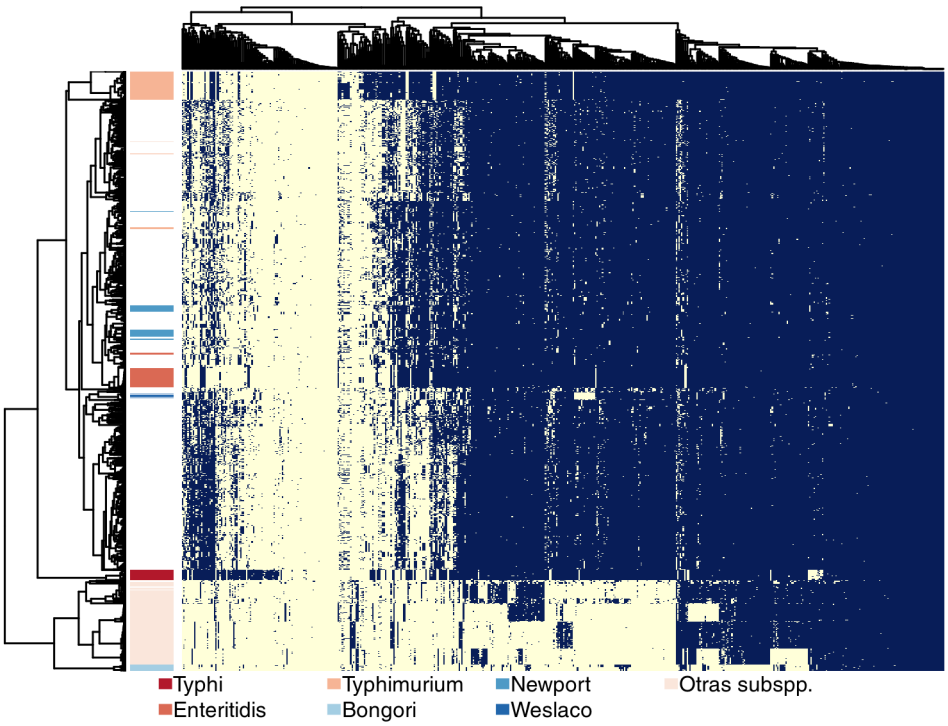

Figura 5: Heatmap basado en los datos de presencia/ausencia (azul y blanco, respectivamente) de 572 ARNp en 5761 genomas de *Salmonella*, mostrando los patrones de similitud en el repertorio de genes codificantes para ARNp.

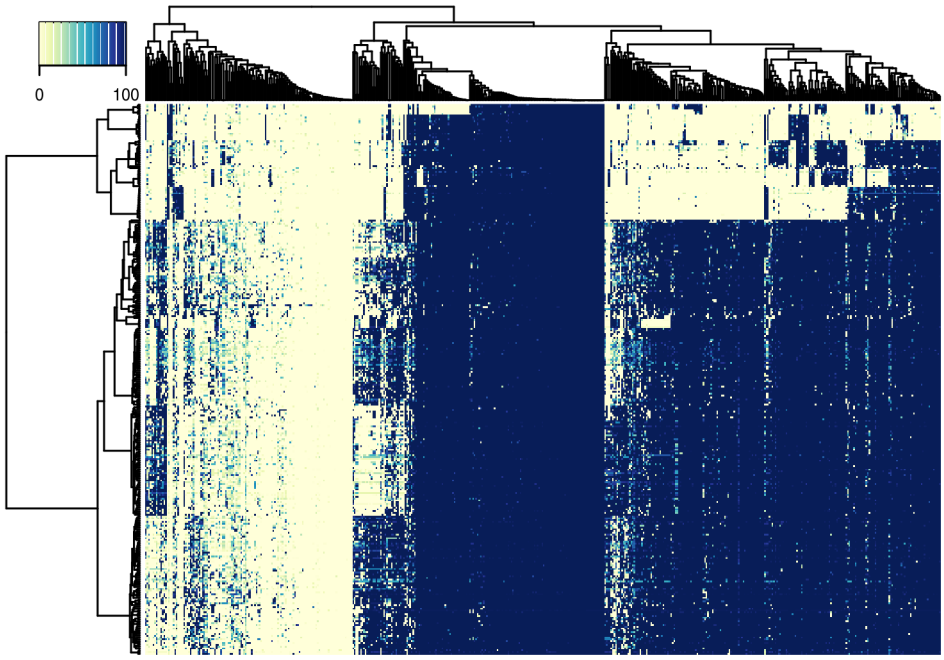

Figura 6: Heatmap basado en el porcentaje de aparición intra-serovar de 572 ARNp, en aquellos serovares representados por al menos 5 genomas.

El repertorio de genes ARNp de los serovares de *S. enterica enterica* varía en un continuo entre 363 y 471 por genoma, si bien un 97% de los genomas de la subespecie caen encima del umbral de 400 (Fig. 7). En los genomas de otras subespecies se identificaron en todos los casos menos de 300 ARNp, pero no debe perderse se vista que el conjunto de genes de referencia del que se parte corresponde a serovares de la subespecie *S. enterica enterica*. Es de esperar entonces que en el tiempo que pasó desde la divergencia se hayan perdido y obtenido ARNp en las distintas subespecies que con esta aproximación no se pueden detectar.

Es posible identificar algunos serovares que parecen haber perdido grandes grupos de ARNp, como *S.* Typhi y el linaje compuesto por *S.* Weslaco, *S.* Denver, *S.* San Juan y *S.* Troy (Fig. 5). La degradación del genoma ha sido reportada en el caso de *S.* Typhi (McClelland et al., 2004).

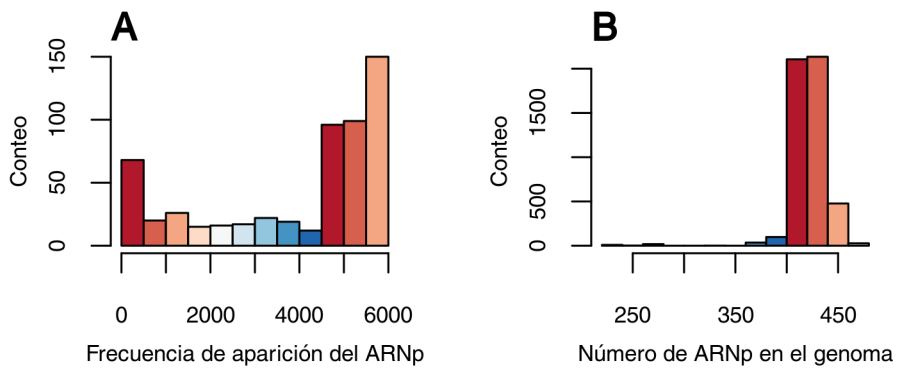

Figura 7: Algo **A** Histograma de la frecuencia de aparición de los ARNp. **B** Histograma del número de ARNp identificados por genoma.

La comparación de tres serovares muy cercanamente emparentados pero con diferencias en su rango de hospedero como lo son *S.* Enteritidis (ubicuo), *S.* Dublin (adaptado a hospedero), y *S.* Gallinarum (hospedero restringido) muestra que si bien el repertorio de ARNp es muy similar, *S.* Enteritidis tiene en promedio 425 ARNp por genoma, contra 421 de *S.* Dublin y 416 de *S.* Gallinarum.

Es notorio que la filogenia se refleja en el patrón de distribución de los ARNp, ya que los distintos grupos monofiléticos (tanto a nivel de las grandes divisiones como a nivel de serovares) quedan agrupados en los *heatmap*, demostrando que los genomas que tienen repertorios de ARNp más similares entre sí son los evolutivamente más cercanos. Aquellos serovares como *S.* Derby y *S.* Newport que se ha demostrado que son polifiléticos, es decir que no reflejan relaciones evolutivas, como es de esperar no aparecen agrupados en los *heatmaps*, mientras que los serovares filogenéticamente cercanos como *S.* Dublin y *S.* Enteritidis aparecen agrupados.

7.4 Los ARNp insertos en grandes redes de regulación evolucionan más lento y son más conservados

Se identificaron los homólogos de 30 ARNp de los que se dispone información validada experimentalmente de su red de regulación. Los ARNp fueron asignados a dos categorías: los que interactúan con un único ARNm y los que lo hacen con más de uno, en un rango que va de 2 a 44 ARNm (Tabla 1).

Los ARNp que regulan más de un gen son más conservados a nivel de presencia/ausencia. La frecuencia de aparición media de los ARNp que regulan más de un gen es 5451.15, y la de los que regulan un único gen es 3967. El test de Wilcoxon indica una diferencia significativa en la media de las dos muestras (*p-value*=0.04985) (Fig. 8 A-B).

La distancia nucleotídica media entre los ARNp para los ARNp que regulan más de un gen es 0.00683181, mientras que para los que regulan un único gen es 0.00683181 (Fig. 8 C-D). El test de Wilcoxon indica una diferencia significativa en la media de las distancias de los dos grupos (*p-value*=2.2e-16).

Con la única excepción de IsrM, los ARNp que regulan más de un gen son comunes a *Salmonella* y a *E. coli*. Entre los 10 ARNp que regulan un único gen (Tabla 1), dos de ellos, InvR e IsrN, están presentes únicamente en *Salmonella* pero ausentes en 27 genomas de *E. coli* (Skippington y Ragan, 2012). Otros dos de ellos, dicF e IsrC, están presentes en *S. enterica* pero ausentes en *S. bongori*, y si bien aparecen esporádicamente en *E. coli*, se ha postulado que su presencia se debe a transferencias horizontales, y ninguno de ellos estaría presente en el ancestro común de *E. coli* y el cercanamente emparentado género *Shigella* (Skippington y

Ragan, 2012). Otros dos ARNp de los que regulan un único gen, sibA y sibD, son verdaderos ARNas, codificados en hebra complementaria al gen que regulan, por lo que es muy improbable que puedan sumar otros genes a su red de regulación. Los otros cuatro ARNp que regulan un solo gen, glmY, glmZ, MicL y sraL, son comunes a *Salmonella* y *E. coli* y tienen un único blanco conocido, también conservado en ambos géneros.

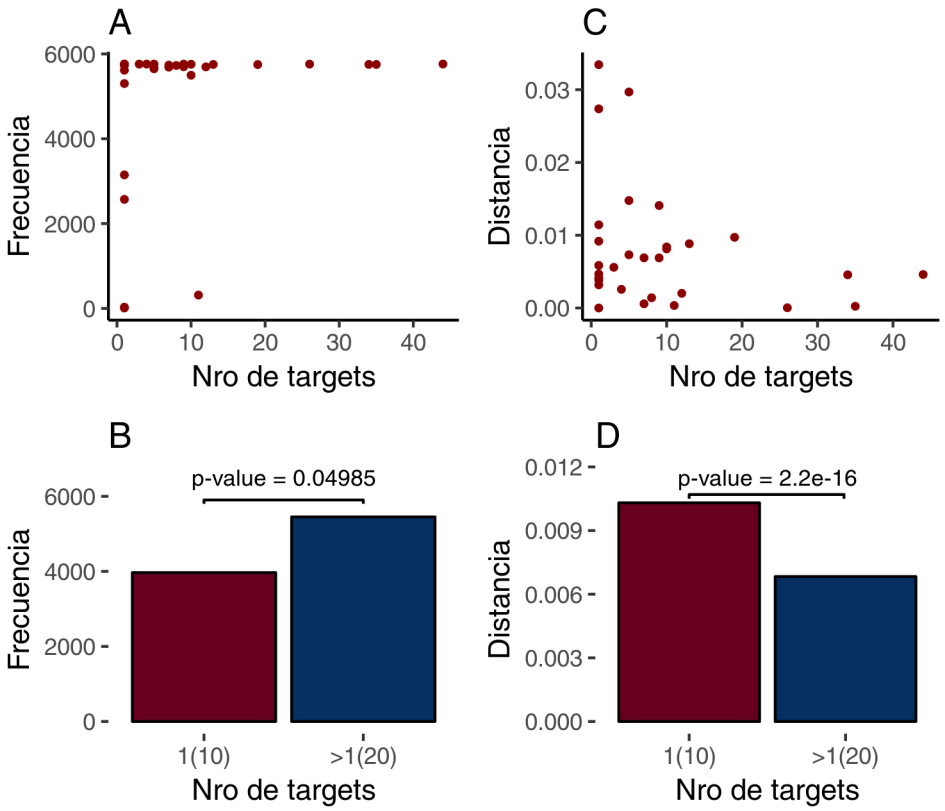

Figura 8: Los ARNp insertos en grandes redes de regulación evolucionan más lento y son más conservados. Se estudió la frecuencia de aparición y distancia evolutiva de 30 ARNp en 5761 genomas de *Salmonella*. **A** Frecuencia de aparición en función del número de genes que regulan. **B** Frecuencia media de aparición de los ARNp agrupados según el número de genes que regulan. **C** Distancia nucleotídica en función del número de genes que regulan. **D** Media de distancias de los ARNp agrupados según el número de genes que regulan. Los valores p fueron calculados usando el test de Wilcoxon no pareado.

Tabla 1: Frecuencia de aparición y distancia nucleotídica de 30 ARNp en 576 genomas. Las distancias nucleotídicas fueron estimadas con el modelo de Jukes-Cantor. Los datos de conservación en *E. coli* fueron tomados de Skippington y Ragan, 2012. El número de genes regulados es el reportado en la base de datos sRNATarBase.

ARNp	Genes regulados	Frecuencia de aparición	Distancia nucleotídica	Conservación en *E. coli*
ArcZ	5	5761	0.0073156127051007	Conservado
ChiX	4	5761	0.00256635871402444	Conservado
CyaR	8	5728	0.00140781559117047	Conservado
dicF	1	10	0.00384662808277961	Variable
dsrA	7	5689	0.00057828995557817	Conservado
fnrS	13	5753	0.00883307285916143	Conservado
GcvB	44	5761	0.0046025342166121	Conservado
glmY	1	5761	0.00585795866217692	Conservado
glmZ	1	5737	0.00585795866217692	Conservado
InvR	1	5302	0.00468134267508611	Ausente
IsrC	1	3148	0.0114326068269932	Variable
IsrM	11	314	0.000345168789168582	Ausente
IsrN	1	2571	0.0334406750645982	Ausente
MicA	19	5750	0.00971270827565274	Conservado
micF	7	5736	0.00689572829872645	Conservado
MicL	1	5760	0.0031825861122934	Conservado
mntS	10	5498	0.00812209254232027	Conservado
omrA	12	5695	0.00201606074163936	Conservado
omrb	9	5698	0.00688970761420285	Conservado
oxyS	10	5757	0.00838898652041041	Conservado
rprA	3	5760	0.00559306892622231	Conservado
RybB	34	5753	0.00456217117491003	Conservado
RyhB1	5	5685	0.0296841318280069	Conservado
RyhB2	5	5650	0.0147673538256639	Conservado
SdsR	35	5752	0.000236815871354251	Conservado
SgrS	9	5761	0.014082644367212	Conservado
sibA	1	5617	0.0273668840987419	Conservado
sibD	1	28	0	Conservado
Spot42	26	5760	3.58842532941117e-05	Conservado
SraL	1	5736	0.00917657827725035	Conservado

Es claro que unos pocos ARNp se alejan del patrón observado. Identificamos un ARNp en particular que regula 11 genes y está presente en solo 314 de los 5761 genomas. Este ARNp, denominado IsrM, fue seleccionado para tratar de rastrear su historia evolutiva.

7.5 IsrM, un ARNp quimera

IsrM es un ARNp de 329 nucleótidos codificado en una isla genómica presente en *S.* Typhimurium pero ausente en otros serovares como *S.* Typhi y otras subespecies como *arizonae* (Fig. 9). De acuerdo al análisis de conservación IsrM se aleja del patrón ya que a pesar de regular 11 genes está presente en sólo 314 genomas.

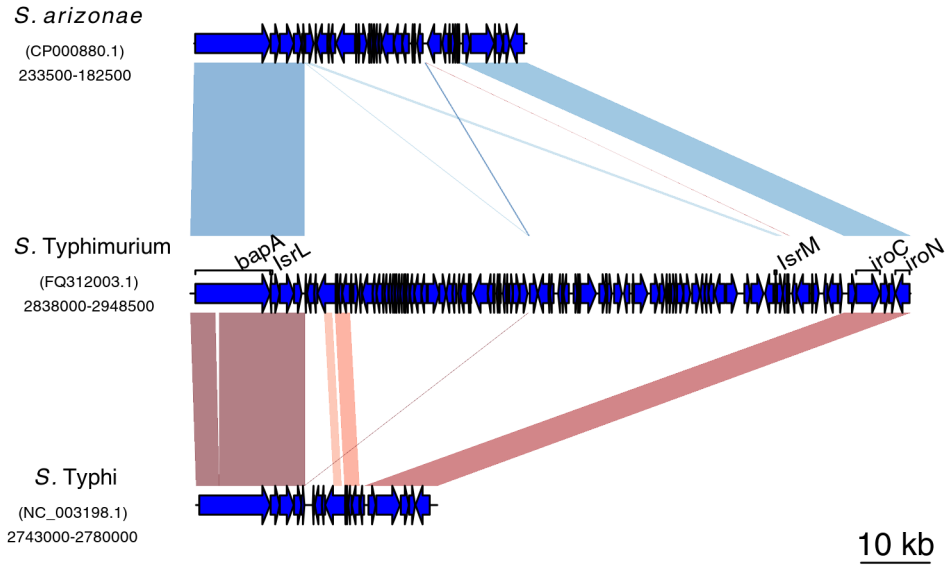

Figura 9: Análisis comparativo de la región del genoma de *S.* Typhimurium SL1344 correspondiente a la isla genómica en que se localiza el gen isrM contra los genomas de *S. arizonae* ser. 62:z4,z23:– y *S.* Typhi CT18. Se muestran las regiones de homología, visualizadas con genoPlotR. Las líneas rojas que conectan los genomas representan coincidencias de BLAST directas, las líneas azules coincidencias invertidas. Colores más oscuros corresponden a *bit scores* más altos. Los códigos entre paréntesis corresponden a códigos de acceso de NCBI. Los rangos corresponden a las coordenadas en el genoma.

La importancia de IsrM para la virulencia y capacidad de colonización en ratón de *S.* Typhimurium ha sido probada *in vivo* (Gong et al., 2011). El mismo estudio probó que este ARNp es esencial para la capacidad de Typhimurium de invadir células epiteliales del intestino y para la supervivencia y reproducción en macrófago.

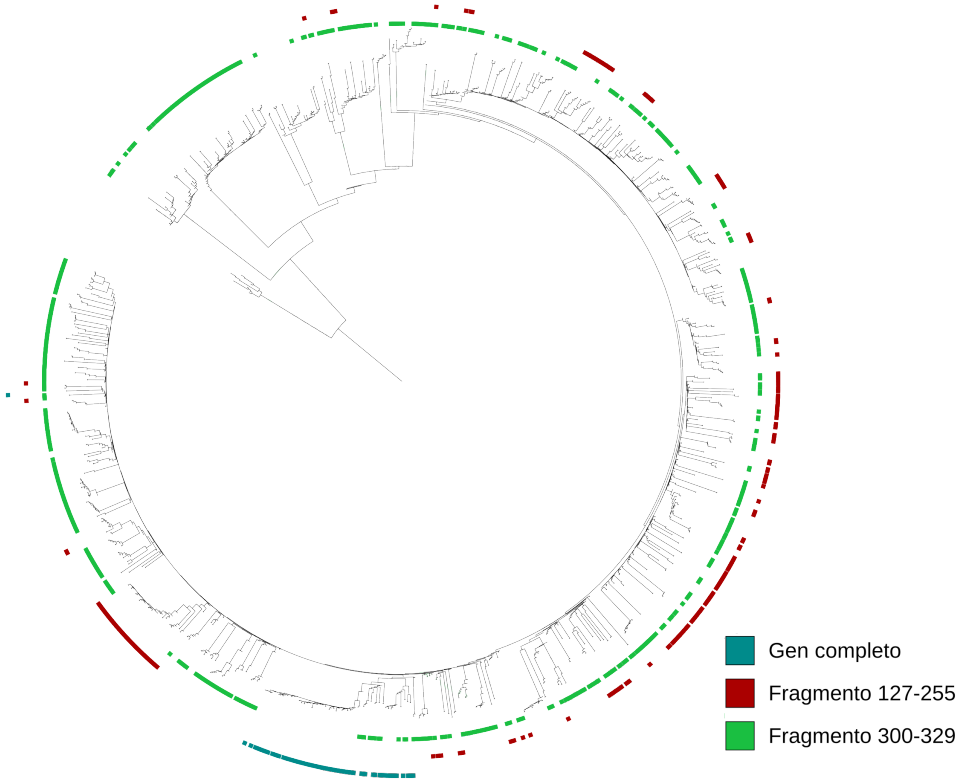

Figura 10: Distribución de IsrM en *Salmonella*. En azul se muestran las coincidencias de BLAST con el ARNp completo con un umbral del 90% de cobertura. En rojo se muestran las coincidencias con el fragmento 127-255 del gen. En verde se muestran las coincidencias con el fragmento 300-329.

IsrM se une directamente e inhibe la expresión tanto del ARNm que codifica a HilE, el represor global de las proteínas de la SPI-1, como del ARNm que codifica a SopA, una proteína efectora codificada en la misma isla que se ha demostrado que modula la respuesta inflamatoria del hospedador. Mutaciones en IsrM llevan a una desregulación de la expresión de *hilE* y *sopA*, así como de otros genes SPI-1

regulados por HilE (Gong et al., 2011).

El mapeo de la distribución de IsrM en la filogenia muestra que el gen está presente únicamente en el linaje correspondiente a los serovares *S.* Typhimurium, *S.* Heidelberg y *S.* Saint Paul, y mientras que es altamente conservado en *S.* Typhimurium (84% de los genomas), *S.* Heidelberg (100% de los genomas), no lo es en *S.* Saint Paul (20% de los genomas) (Fig. 10). El contexto genómico en el que se encuentra el gen es 100% conservado en los 314 genomas.

Encontramos que distintos fragmentos del gen tienen homología con regiones específicas de genomas esparcidos en toda la filogenia, incluso en otras subespecies de *S. enterica* como *salamae* y *houtenae* (Fig. 11). El fragmento de 30 nt que va de la posición 300 a la 329 del gen isrM fue identificado en 2801 genomas, mientras que el fragmento de 129 nt que va de la posición 127 a la 255 fue identificado en 964 genomas, en los que no forman parte del gen completo. Además, ambos fragmentos coexisten en 77 genomas sin formar parte del gen.

El sitio de unión a HilE, en la posición 257-266 del gen isrM (Gong et al., 2011), no está presente en ninguno de estos fragmentos, mientras que la región homóloga el sitio de unión a SopA, en las posiciones 159-167 (Gong et al., 2011), no es conservado en los fragmentos que no forman parte del gen, ya que distintas variantes fueron identificadas. Ambos sitios son 100% conservados a nivel de secuencia en los 314 genomas en que IsrM fue identificado.

En cuanto a sus blancos de regulación directa, HilE es 100% conservado en *Salmonella*, mientras que SopA está presente en el 85% de los genomas estudiados, y no es totalmente conservado en ninguna de las dos especies de *Salmonella*.

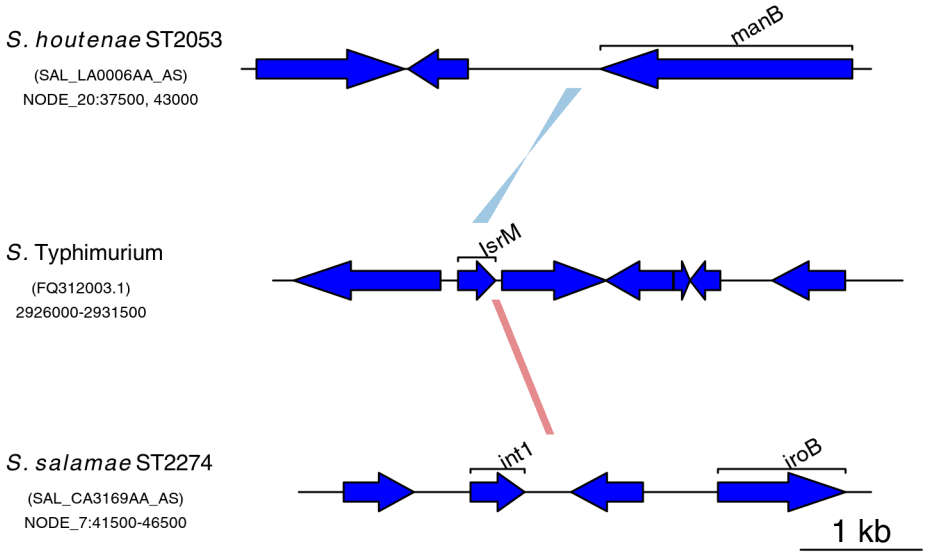

Figura 11: Comparación de la región del genoma de *S.* Typhimurium SL1344 que codifica el ARNp IsrM contra genomas correspondientes a *S. salamae* ST2274 y *S. houtenae* ser. 50:z4,z23:-. Se muestran las regiones de homología, visualizadas con genoPlotR. Las líneas rojas que conectan los genomas representan coincidencias de BLAST directas. Las líneas azules representan coincidencias invertidas. Colores más oscuros corresponden a *bit scores* más altos. Sólo los genes caracterizados fueron anotados. Los códigos entre paréntesis corresponden al código de acceso de NCBI para S. Typhimurium y a los de EnteroBase para *S. salamae* y *S. houtenae*. Los valores corresponden a las coordenadas en el genoma.

8 Discusión

8.1 Filogenia

Comenzamos este trabajo realizando una reconstrucción filogenética a partir de múltiples genes con el fin de utilizarla como referencia para mapear la distribución de los ARNp. Existen dos grandes estrategias para reconstruir filogenias a partir de múltiples genes: la aproximación clásica es la de concatenar la información de distintos genes en una super-matriz. En la otra estrategia las filogenias son inferidas para cada gen por separado, y se obtiene un resumen que representa el árbol de especies.

La estrategia de la concatenación de múltiples genes aumenta el número absoluto de cambios en cada rama permitiendo resolver politomías que no podrían ser resueltas en árboles de genes individuales por tener una señal filogenética débil, ya sea por una baja tasa evolutiva del gen, corto tiempo transcurrido tras la divergencia, secuencias cortas, etc. Sin embargo, resolver con precisión algunas relaciones evolutivas no es tan simple como aplicar más datos al problema. Aunque es claro que una señal débil puede ser reforzada cuando las secuencias se concatenan, y la mayoría de las veces esta amplificación de señal debería ocurrir para las particiones correctas, debe tenerse en cuenta que en algunos casos también puede aumentar el apoyo para las inferencias erróneas (Gadagkar, Rosenberg, y Kumar, 2005). En particular, la fuerza del enfoque de concatenación disminuye cuando muchos de los genes están influenciados por eventos de recombinación, paralogías ocultas, o coalescencias profundas, ya que el enfoque asume que todos los datos han evolucionado de acuerdo con una única historia evolutiva, posiblemente con

diferentes tasas de mutación para diferentes sitios. La estrategia de concatenación ignora la ocurrencia de diferentes historias evolutivas en diferentes genes, lo que puede conducir a un alto apoyo estadístico para inferencias incorrectas. En este estudio se excluyeron los sitios que probablemente han estado bajo la influencia de recombinación, pero las coalescencias profundas no son consideradas en esta estrategia. Otro aspecto de esta reconstrucción que debe ser tenida en cuenta es que analizar sitios únicamente variables puede llevar a sobreestimar el largo de ramas (Lewis, 2001). En este estudio se usó una corrección para la omisión de sitios invariables que ayuda a reducir el problema, pero que aún tiene errores de una magnitud tal que no permite utilizar la información de longitud de ramas en análisis posteriores (Leaché et al., 2015).

La estrategia de obtener el árbol de especies a partir de los árboles de genes tampoco está libre de problemas ni puede garantizar una filogenia de especies correcta ante los mismos eventos. Debido a la forma estocástica en que los linajes se reparten durante la especiación, los árboles de genes pueden diferir en topología entre sí y con el árbol de especies. Degnan y Rosenberg (2006) encontraron que para cualquier topología de árbol con cinco o más especies, existen longitudes de rama para las cuales la discordancia del árbol de genes es tan común que el procedimiento de usar la topología más observada entre los árboles de genes como estimación de la topología del árbol de especies probablemente estime el árbol incorrecto. Para tener en cuenta estos eventos, se utiliza el modelo del coalescente para múltiples especies. Sin embargo, este método asume que todos los árboles de genes de entrada son correctos, y errores en la estimación de los árboles de genes pueden tener un impacto grande en la estimación del árbol de especies (Warnow, 2017).

Vale la pena mencionar que la transferencia horizontal extensa desafía la noción misma del árbol de la especie. Cuando se cree que los datos con los que se quiere reconstruir una historia evolutiva se pueden subdividir en regiones que han tenido diferentes historias, posiblemente elegir un único árbol no sea la mejor manera de representar las relaciones filogenéticas, y los métodos de redes pueden ser una mejor elección.

En nuestro caso, la filogenia no intenta representar la verdadera historia evolutiva de *Salmonella* sino servir de referencia para interpretar el resto de los análisis en un contexto evolutivo, y en principio cualquier estrategia habría sido válida siempre y cuando se tengan en cuenta las limitaciones del método a la hora de interpretar los resultados.

Tanto la filogenia realizada en base al genoma *core* de todos los secuenciotipos reportados como el análisis comparativo realizado en base al contenido de ARNp sugieren que algunos serovares de *Salmonella* no forman grupos monofiléticos. Algunos de los que en ambos estudios aparentan ser monofiléticos son *S.* Heidelberg, *S.* Paratyphi A, *S.* Paratyphi B, *S.* Paratyphi C, y *S.* Typhi. Otros serovares aparecen como grupos polifiléticos, como ya ha sido reportado en algunos casos, como *S.* Newport, *S.* Oranienburg y *S.* Paratyphi B, *S.* Enteritidis y *S.* Typhimurium (ver por ejemplo, Achtman et al., 2012). La explicación mas evidente para estas incongruencias es la transferencia horizontal de los genes que codifican para los antígenos de superficie que definen a un serovar. Esto puede ser relevante para los estudios epidemiológicos, ya que es importante identificar estas incongruencias entre la fórmula antigénica y la filogenia para poder caracterizar y comunicar correctamente el patógeno involucrado en un brote.

8.2 Evolución de los ARNp

Aunque los ARNp son críticos en la regulación de la expresión génica, la mayor parte de los estudios realizados hasta el momento se han enfocado en su función, y no está claramente entendido cómo se originan ni cuáles son las fuerzas que moldean su evolución en las bacterias. Con una mejor comprensión de los roles fisiológicos y mecanismos de acción, y las gran cantidad de ARNp detectados o predichos, es cada vez más factible abordar preguntas sobre su evolución.

Los resultados de este trabajo muestran un alto nivel de conservación general de los ARNp en *Salmonella*, lo cual señala su importancia funcional y una alta presión selectiva por ser mantenidos. Algunos de los ARNp están notablemente conservados tanto en *S. enterica* como en *S. bongori*, indicativo de una función celular que vas más allá de las necesidades de una única especie. Ejemplos de estos son GcvB, Spot_42, ChiX, ArcZ, y MicL. Además, muchos de los ARNp son específicos de *S. enterica*, y algunos son también notablemente conservados. Entre los no conservados, algunos muestran un patrón de presencia/ausencia que parece ser resultado principalmente de pérdidas, ya que las variaciones son esporádicas, como es el caso de IsrN, mientras que otros son linaje-específicos, como es el caso de IsrM. Las diferencias en los repertorios de ARNp pueden producirse por pérdidas linaje-específicas, por la emergencia de nuevos ARNp, o por eventos de transferencia horizontal. Este estudio está sesgado a detectar principalmente eventos de deleción linaje-específicos, ya que no puede detectar ARNp únicos en genomas de otros serovares más que Typhi y Typhimurium.

Típicamente cada ARNp regula la expresión de varios genes. Posiblemente el establecimiento de una primera interacción entre un nuevo ARNp y un blanco pone

al ARNp bajo presión selectiva para ser mantenido (Peer y Margalit, 2014), lo que permite además la adquisición de otros blancos de regulación. En consonancia con esto, se ha demostrado que los ARNp sufren una acumulación gradual de sitios de unión adicionales desde el establecimiento de la interacción con el primer ARNp (Peer y Margalit, 2014). Debido a que ya están reunidos los elementos que requiere un ARNp productivo, como un terminador fuerte, un promotor regulado, o el sitio de unión a Hfq, sólo debe evolucionar un sitio de unión con nuevos blancos. Sin embargo, existen algunos ARNp regulan un único gen. Buscando aportar información sobre sus diferencias, estudiamos por separado ambos grupos.

Nuestros resultados sugieren que los ARNp más conservados a nivel de presencia/ausencia están integrados en redes regulatorias más grandes que los ARNp variables. También encontramos evidencia que sugiere que los ARNp insertos en grandes redes son en promedio más antiguos que los que regulan un único gen, ya que con la única excepción de IsrM, serían ancestrales a la separación de *Salmonella* y *E. coli*, y son notablemente conservados también en esta última especie, mientras que la evidencia sugiere que varios de los que regulan un único gen son evolutivamente jóvenes. Sería interesante ampliar este estudio a otros géneros del orden de los Enterobacteriales. Además de ser más conservados, los ARNp que controlan la expresión de varios genes tienen una tasa de divergencia mas baja que los ARNp que sólo regulan un gen, posiblemente por el efecto de una selección purificadora más fuerte. Esto sugiere una restricción funcional mayor para los ARNp que son centro de regulación, posiblemente debido a que la mayoría de las mutaciones en estos ARNp afecten el *fitness* global del organismo de forma mucho más severa.

En cuanto a los ARNp que regulan un único gen, es posible que su rápida evolución molecular pueda facilitar el surgimiento de nuevas interacciones. En concordancia con esto, un estudio reciente en *S. enterica* y *E. coli* aportó evidencia de que los ARNp evolutivamente jóvenes evolucionan más rápido que los ARNp más antiguos (Kacharia et al., 2017). El mismo estudio concluye que el nivel de expresión de los primeros es más bajo que el estos últimos. La co-ocurrencia de una baja expresión y una rápida evolución podría facilitar aún más el establecimiento de nuevas interacciones, ya que los posibles efectos negativos de las interacciones nacientes serían mitigados por la baja expresión.

Debido a la falta de estudios profundos del interactoma de muchos ARNp, es plausible que algunos de los ARNp estudiados tengan una red regulatoria mayor a la actualmente descripta, lo que afectaría los resultados de este trabajo. Además, esta red de interacciones está lejos de ser completa, con sólo 30 ARNp de los más de 400 que identificamos en cada genoma de *S. enterica enterica*, por lo que nuestras conclusiones pueden ser sesgadas. El desarrollo de nuevos métodos que permiten estudiar las interacciones de los ARN *in vivo* a escala global (ver por ejemplo, Z. Lu et al., 2016) sin duda permitirá un estudio mucho más completo de la evolución de las redes de regulación.

Encontramos evidencia de grandes pérdidas de ARNp específicas del serovar *S.* Typhi, en linea con la evidencia de degradación del genoma ya reportada para este serovar (McClelland et al., 2004), un proceso conocido como *streamlining* del genoma que implica la pérdida de funciones que no son importantes para el organismo y que en el caso de *S.* Typhi se cree que está relacionado con la adaptación a hospedero. La evidencia de que *S.* Gallinarum, serovar de hospedero

restringido, ha perdido más ARNp que *S.* Dublin y *S.* Enteritidis, serovares adaptado a hospedero y ubicuo, respectivamente, apoya la idea de que el proceso de degradación del genoma relacionado con la adaptación a hospedero se ve acompañado de pérdidas de genes codificantes para ARNp. Debe notarse que sólo estamos estudiando los casos de deleción, y no la formación de pseudogenes, que también contribuye en gran medida al proceso de degradación del genoma.

8.3 Origen de IsrM

IsrM es un ARNp codificado en una isla genómica originalmente identificado en *S.* Typhimurium cuya importancia en la invasión de células epiteliales, replicación intracelular dentro de macrófagos, y virulencia y colonización en ratón fue firmemente establecida (Gong et al., 2011).

A pesar de participar en una red de regulación grande y ser fundamental para la patogenicidad y virulencia de *S.* Typhimurium, su distribución filogenética es muy limitada. Si bien los datos de expresión indican que IsrM se expresa como un ARN de 329 nucleótidos, encontramos que distintos fragmentos pequeños del gen tienen homología con regiones específicas de genomas de varios linajes, e incluso son anteriores a la separación de las subespecies de *S. enterica.* Nuestros resultados sugieren que este ARNp podría haberse originado por rearreglos genómicos en el ancestro común del linaje compuesto por los serovares Typhimurium, Saint Paul y Heidelberg. No es claro por qué el ARNp es poco conservado en el serovar Saint Paul, pero es posible que su importancia radique en su contribución a la interacción específica con algunos hospederos.

Aunque los mecanismos que moldean los repertorios de ARNp en las bacterias

no son del todo entendidos, las diferencias en los repertorios de ARNp entre especies bacterianas se han atribuido previamente a la duplicación, eliminación o transferencia horizontal (Skippington y Ragan, 2012, Gottesman y Storz (2011)). Un estudio reciente entre genomas de *Salmonella* y *E. coli* destaca la contribución de los rearreglos genómicos a este proceso, al presentar evidencia de un ARNp que se pierde en *Salmonella* cuando la región intergénica en que se encontraba localizado fue dividida en dos, y de otro ARNp que se originó en una región intergénica formada por un rearreglo genómico, probablemente por mutaciones puntuales que crearon un promotor (Raghavan, Kacharia, Millar, Sislak, y Ochman, 2015). Los resultados de nuestro trabajo apuntan en el mismo sentido.

El contexto genómico en el que se encuentra el gen isrM en los 314 genomas en que fue identificado es muy conservado, lo cual es esperable dado que las cepas que tienen el gen están cercanamente emparentadas, siendo el grado de sintenia general esperado muy alto. Los sitios de unión a blanco son 100% conservados en estos 314 genomas. El sitio de unión a HilE no tiene homología con genomas que no tienen el gen completo. El sitio de unión a SopA en cambio está contenido en el fragmento 127-255 del gen que tiene homología con genomas esparcidos en toda la filogenia. Sin embargo, la evidencia sugiere que el fragmento no coevoluciona con SopA, estando este último ausente en muchos de los genomas que presentan el fragmento, y viceversa, por lo que puede especularse que estos fragmentos no cumplen un rol fisiológico. La región homóloga al sitio de unión a SopA en estos fragmentos no es conservada, encontrándose múltiples variantes. Si bien esto es esperable ya que en ausencia de una función fisiológica no habría restricción funcional, tampoco puede descartarse que esta variabilidad se deba únicamente a la divergencia entre estos genomas.

9 Conclusiones

Con respecto al análisis del patrón filogenético de los ARNp, los resultados muestran que existen ARNp notablemente conservados en todos los genomas de *Salmonella*, otros muy conservados en la subespecie *S. enterica enterica* pero ausentes en otras subespecies, y un grupo menor de ARNp cuya presencia es variable.

Nuestros resultados sugieren que estas diferencias se explican al menos en parte por el tamaño de la red que regulan. Aquellos ARNp que se integran en grandes redes son más conservados y evolucionan más lento que los que regulan pequeñas redes, efecto de una selección purificadora más fuerte. Además, el tamaño de la red de regulación podría relacionarse con la antigüedad del ARNp, siendo más antiguos aquellos ARNp que participan en grandes redes de regulación.

Con respecto a los mecanismos que generan diferencias en el repertorio de ARNp en los distintos linajes, nuestros resultados sugieren que los rearreglos genómicos pueden contribuir a este proceso.

10 Referencias

Achtman, M., Wain, J., Weill, F.-X., Nair, S., Zhou, Z., Sangal, V., ... others. (2012). Multilocus sequence typing as a replacement for serotyping in Salmonella enterica. *PLoS Pathogens*, *8*(6), e1002776.

Agbor, T. A., y McCormick, B. A. (2011). Salmonella effectors: important players modulating host cell function during infection. *Cellular Microbiology*, *13*(12), 1858–1869.

Allman, E. S., Degnan, J. H., y Rhodes, J. A. (2011). Determining species tree topologies from clade probabilities under the coalescent. *Journal of Theoretical Biology*, *289*, 96–106.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., y Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., y Wheeler, D. L. (2005). GenBank. *Nucleic Acids Research*, *33*(suppl_1), D34–D38.

Betancor, L., Yim, L., Fookes, M., Martinez, A., Thomson, N. R., Ivens, A., ... others. (2009). Genomic and phenotypic variation in epidemic-spanning Salmonella enterica serovar Enteritidis isolates. *BMC Microbiology*, *9*(1), 237.

Brennan, R. G., y Link, T. M. (2007). Hfq structure, function and ligand binding. *Current Opinion in Microbiology*, *10*(2), 125–133.

Brenner, F., Villar, R., Angulo, F., Tauxe, R., y Swaminathan, B. (2000). Salmonella nomenclature. *Journal of Clinical Microbiology*, *38*(7), 2465–2467.

Bruen, T., y Bruen, T. (2005). PhiPack: PHI test and other tests of recombination.

McGill University, Montreal, Quebec.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, *17*(4), 540–552.

Degnan, J. H., y Rosenberg, N. A. (2006). Discordance of species trees with their most likely gene trees. *PLoS Genetics*, *2*(5), e68.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797.

Falush, D., Torpdahl, M., Didelot, X., Conrad, D. F., Wilson, D. J., y Achtman, M. (2006). Mismatch induced speciation in Salmonella: model and data. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *361*(1475), 2045–2053.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., y Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, *391*(6669), 806–811.

Fröhlich, K. S., Haneke, K., Papenfort, K., y Vogel, J. (2016). The target spectrum of SdsR small RNA in Salmonella. *Nucleic Acids Research*, *44*(21), 10406–10422.

Fu, L., Niu, B., Zhu, Z., Wu, S., y Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152.

Gadagkar, S. R., Rosenberg, M. S., y Kumar, S. (2005). Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, *304*(1), 64–74.

Gong, H., Vu, G.-P., Bai, Y., Chan, E., Wu, R., Yang, E., ... Lu, S. (2011). A

Salmonella small non-coding RNA facilitates bacterial invasion and intracellular replication by modulating the expression of virulence factors. *PLoS Pathogens*, *7*(9), e1002120.

Gottesman, S., y Storz, G. (2011). Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harbor Perspectives in Biology*, *3*(12), a003798.

Guy, L., Roat Kultima, J., y Andersson, S. G. (2010). genoPlotR: comparative gene and genome visualization in R. *Bioinformatics*, *26*(18), 2334–2335.

Hébrard, M., Kröger, C., Srikumar, S., Colgan, A., Händler, K., y Hinton, J. C. (2012). sRNAs and the virulence of Salmonella enterica serovar Typhimurium. *RNA Biology*, *9*(4), 437–445.

Issenhuth-Jeanjean, S., Roggentin, P., Mikoleit, M., Guibourdenche, M., De Pinna, E., Nair, S., ... Weill, F.-X. (2014). Supplement 2008–2010 (no. 48) to the White–Kauffmann–Le Minor scheme. *Research in Microbiology*, *165*(7), 526–530.

Jacob, F., y Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, *3*(3), 318–356.

Jacobsen, A., Hendriksen, R. S., Aaresturp, F. M., Ussery, D. W., y Friis, C. (2011). The Salmonella enterica pan-genome. *Microbial Ecology*, *62*(3), 487.

Kacharia, F. R., Millar, J. A., y Raghavan, R. (2017). Emergence of New sRNAs in Enteric Bacteria is Associated with Low Expression and Rapid Evolution. *Journal of Molecular Evolution*, *84*(4), 204–213.

Kishino, H., Miyata, T., y Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *Journal of Molecular Evolution*,

31(2), 151–160.

Kröger, C., Colgan, A., Srikumar, S., Händler, K., Sivasankaran, S. K., Hammarlöf, D. L., ... others. (2013). An infection-relevant transcriptomic compendium for Salmonella enterica Serovar Typhimurium. *Cell Host Microbe*, *14*(6), 683–695.

Leaché, A. D., Banbury, B. L., Felsenstein, J., Oca, A. N.-M. de, y Stamatakis, A. (2015). Short tree, long tree, right tree, wrong tree: new acquisition bias corrections for inferring SNP phylogenies. *Systematic Biology*, *64*(6), 1032–1047.

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, *50*(6), 913–925.

Lindgreen, S., Umu, S. U., Lai, A. S.-W., Eldai, H., Liu, W., McGimpsey, S., ... others. (2014). Robust identification of noncoding RNA from transcriptomes requires phylogenetically-informed sampling. *PLoS Computational Biology*, *10*(10), e1003907.

Lu, Z., Zhang, Q. C., Lee, B., Flynn, R. A., Smith, M. A., Robinson, J. T., ... others. (2016). RNA duplex map in living cells reveals higher-order transcriptome structure. *Cell*, *165*(5), 1267–1279.

Majowicz, S. E., Musto, J., Scallan, E., Angulo, F. J., Kirk, M., O'brien, S. J., ... Enteric Disease "Burden of Illness" Studies, I. C. on. (2010). The global burden of nontyphoidal Salmonella gastroenteritis. *Clinical Infectious Diseases*, *50*(6), 882–889.

McClelland, M., Sanderson, K. E., Clifton, S. W., Latreille, P., Porwollik, S., Sabo, A., ... others. (2004). Comparison of genome degradation in Paratyphi A and Typhi, human-restricted serovars of Salmonella enterica that cause typhoid. *Nature Genetics*, *36*(12), 1268–1274.

Minh, B. Q., Nguyen, M. A. T., y Haeseler, A. von. (2013). Ultrafast approximation for

phylogenetic bootstrap. *Molecular Biology and Evolution*, *30*(5), 1188–1195.

Miyakoshi, M., Chao, Y., y Vogel, J. (2015). Regulatory small RNAs from the 3' regions of bacterial mRNAs. *Current Opinion in Microbiology*, *24*, 132–139.

Padalon-Brauch, G., Hershberg, R., Elgrably-Weiss, M., Baruch, K., Rosenshine, I., Margalit, H., y Altuvia, S. (2008). Small RNAs encoded within genetic islands of Salmonella typhimurium show host-induced expression and role in virulence. *Nucleic Acids Research*, *36*(6), 1913–1927.

Page, A. J., Taylor, B., Delaney, A. J., Soares, J., Seemann, T., Keane, J. A., y Harris, S. R. (2016). SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial Genomics*, *2*(4).

Paradis, E., Claude, J., y Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*, 289–290.

Peer, A., y Margalit, H. (2011). Accessibility and evolutionary conservation mark bacterial small-RNA target-binding regions. *Journal of Bacteriology*, *193*(7), 1690–1701.

Peer, A., y Margalit, H. (2014). Evolutionary patterns of Escherichia coli small RNAs and their regulatory interactions. *Rna*, *20*(7), 994–1003.

Perkins, T. T., Kingsley, R. A., Fookes, M. C., Gardner, P. P., James, K. D., Yu, L., ... others. (2009). A strand-specific RNA–Seq analysis of the transcriptome of the typhoid bacillus salmonella typhi. *PLoS Genetics*, *5*(7), e1000569.

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https:

//www.R-project.org/

Raghavan, R., Groisman, E. A., y Ochman, H. (2011). Genome-wide detection of novel regulatory RNAs in E. coli. *Genome Research*, *21*(9), 1487–1497.

Raghavan, R., Kacharia, F. R., Millar, J. A., Sislak, C. D., y Ochman, H. (2015). Genome rearrangements can make and break small RNA genes. *Genome Biology and Evolution*, *7*(2), 557–566.

Schatten, H., y Eisenstark, A. (2007). *Salmonella: methods and protocols*. Springer Science y Business Media.

Skippington, E., y Ragan, M. A. (2012). Evolutionary dynamics of small RNAs in 27 Escherichia coli and Shigella genomes. *Genome Biology and Evolution*, *4*(3), 330–345.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313.

Storz, G., Vogel, J., y Wassarman, K. M. (2011). Regulation by small RNAs in bacteria: expanding frontiers. *Molecular Cell*, *43*(6), 880–891.

Tabara, H., Grishok, A., y Mello, C. C. (1998). RNAi in C. elegans: soaking in the genome sequence. *Science*, *282*(5388), 430–431.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences*, *17*(2), 57–86.

Timme, R. E., Pettengill, J. B., Allard, M. W., Strain, E., Barrangou, R., Wehnes, C., ... Brown, E. W. (2013). Phylogenetic diversity of the enteric pathogen Salmonella enterica subsp. enterica inferred from genome-wide reference-free SNP characters. *Genome Biology and Evolution*, *5*(11), 2109–2123.

Updegrove, T. B., Shabalina, S. A., y Storz, G. (2015). How do base-pairing small RNAs

evolve? *FEMS Microbiology Reviews*, *39*(3), 379–391.

Uzzau, S., Brown, D. J., Wallis, T., Rubino, S., Leori, G., Bernard, S., ... Olsen, J. E. (2000). Host adapted serotypes of Salmonella enterica. *Epidemiology y Infection*, *125*(2), 229–255.

Wang, J., Liu, T., Zhao, B., Lu, Q., Wang, Z., Cao, Y., y Li, W. (2015). sRNATarBase 3.0: an updated database for sRNA-target interactions in bacteria. *Nucleic Acids Research*, *44*(D1), D248–D253.

Warnes, G. R., Bolker, B., Bonebakker, L., Gentleman, R., Liaw, W. H. A., Lumley, T., ... Venables, B. (2016). *gplots: Various R Programming Tools for Plotting Data*. Retrieved from https://CRAN.R-project.org/package=gplots

Warnow, T. (2017). *Computational phylogenetics: An introduction to designing methods for phylogeny estimation*. Cambridge University Press.

Yachison, C. A., Yoshida, C., Robertson, J., Nash, J. H., Kruczkiewicz, P., Taboada, E. N., ... others. (2017). The validation and implications of using whole genome sequencing as a replacement for traditional serotyping for a national Salmonella reference laboratory. *Frontiers in Microbiology*, *8*, 1044.

Yoshida, C. E., Kruczkiewicz, P., Laing, C. R., Lingohr, E. J., Gannon, V. P., Nash, J. H., y Taboada, E. N. (2016). The Salmonella in silico typing resource (SISTR): an open Web-accessible tool for rapidly typing and subtyping draft Salmonella genome assemblies. *PLoS One*, *11*(1), e0147101.

Zhang, C., Sayyari, E., y Mirarab, S. (2017). ASTRAL-III: Increased Scalability and Impacts of Contracting Low Support Branches. In *RECOMB International Workshop on Comparative Genomics* (pp. 53–75). Springer.