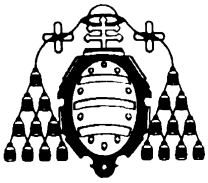

# *UNIVERSIDAD DE OVIEDO*

**MASTER UNIVERSITARIO EN BIOTECNOLOGÍA ALIMENTARIA**

# "Análisis bioinformático de la secuencia de un plásmido de resistencia-virulencia de *Salmonella enterica* obtenida por técnicas de primera, segunda y tercera generación"

**TRABAJO FIN DE MASTER**

**POR**

**Xenia Vázquez Sánchez**

**Julio, 2016**

**Master en Biotecnología Alimentaria**
**Universidad de Oviedo**
**C/Julián Clavería s/n. 33071 Oviedo. España**
**Tel. 985106226. Fax 985103434. http://www.unioviedo.es/MBTA**

**PROFESORES TUTORES:**

Dra. Mª DEL ROSARIO RODICIO RODICIO, Profesora titular de Microbiología del Departamento de Biología Funcional (Universidad de Oviedo) y el Dr. VÍCTOR MANUEL LADERO LOSADA del Instituto de Productos Lácteos de Asturias (IPLA-CSIC).

**CERTIFICA:**

Que Dña. **XENIA VÁZQUEZ SÁNCHEZ** ha realizado bajo mi dirección el Trabajo Fin de Master al que corresponde la presente memoria en el contexto de los estudios del Master Universitario en Biotecnología Alimentaria, 10ª promoción curso 2015-2016.

Oviedo, 11 de Julio de 2016

Fdo. Dra. Mª del Rosario Rodicio Fdo. Dr. Víctor Manuel Losada

VºBº

Manuel Rendueles de la Vega
Coordinador del Master en Biotecnología Alimentaria

# Agradecimientos

*En primer lugar, quisiera darle las gracias a la Dra. Mª del Rosario Rodicio Rodicio y a Víctor Ladero Losada por la confianza depositada en mí, y darme la oportunidad de realizar este proyecto, así como todo el apoyo y dedicación recibida para lograrlo.*

*A Raquel García Fierro por enseñarme todo lo que he aprendido durante este tiempo, por la paciencia y toda la ayuda brindada durante estos dos años, por enseñarme a tener paciencia cuando las cosas no salen bien y recibirme siempre con una sonrisa aunque su día fuese agobiante. Por el interés y hacerme sentir una más, gracias.*

*A Vane, Javi y Nacho por hacerme sentir como en casa y ayudarme siempre que lo he necesitado.*

*A mis amigas Ana Belén, Ana María, Paula, Priscila y Soraya por contagiarme vuestro optimismo y apoyarme siempre que lo he necesitado, incluso en la distancia. Por estar conmigo en los buenos momentos y no abandonarme durante estos años tan duros. Gracias por no dejarme caer y hacerme salir del "caparazón" en el que me había encerrado. Por enseñarme que los mejores amigos no son aquellos con los que compartes más tiempo, sino aquellos que por mucho tiempo que pase siempre están ahí para tenderte la mano cuando los necesites.*

*A todos mis compañeros del Grado en Biología y del Máster Biotecnología alimentaria por las risas y los buenos momentos vividos, porque sin ellos no habría sido lo mismo llegar hasta aquí, y por ello son una parte imprescindible que no pueden faltar.*

*A toda la gente que ha pasado por mi vida, profesores, antiguos compañeros, amigos, el club de tenis de Grado, etc., por ayudarme a forjar la persona que soy hoy en día, y que sin las cuales tal vez no sería posible escribir hoy estas líneas.*

*Y por último, pero no menos importante, a mi familia por hacer posible esto y confiar siempre en mí. En especial a mamá porque estos dos años no te lo he hecho fácil, lo sé, pero tú siempre has sabido estar ahí para no dejarme caer y ayudarme a seguir cuando no me creía con fuerzas para ello.*

# Índice

# Resumen
# Abstract

# RESUMEN

Este trabajo se centró en el plásmido de resistencia-virulencia pUO-STmRV1 característico del clon español variante monofásica 4,[5],12:i:- de *Salmonella enterica*. Dicha variante es un importante patógeno transmitido por alimentos. El objetivo del trabajo fue aplicar distintas herramientas bioinformáticas para comparar la información sobre la secuencia del plásmido obtenida por técnicas de secuenciación de primera (Sanger), segunda (Illumina y Roche 454) y tercera (SMART-PacBio) generación. Los datos aportados por Sanger y Roche 454 habían sido combinados en un trabajo previo. En el caso de Illumina y SMART-PacBio la secuencia del plásmido se obtuvo como parte de la secuencia del genoma completo de la bacteria que lo contiene.

En la primera parte del trabajo se evaluó la cantidad y calidad de las secuencias obtenidas por estas técnicas. Los mejores resultados se consiguieron con SMRT-PacBio, que proporcionó la secuencia completa, en un solo contig, tanto del cromosoma (4907248 pb) como del plásmido (202354 pb). La secuencia del plásmido obtenida por Illumina, Roche 454 y Sanger/Roche 454 constituyó el 82.7%, 78,5% y 97,7% de la secuencia de PacBio, distribuidas en 21, 30 y 3 contigs, respectivamente. En segundo lugar se llevó a cabo el ensamblaje de las lecturas Illumina (cromosoma y plásmido) mediante el ensamblador *de novo* Velvet y el programa VelvetOptimiser. La optimización manual del parámetro k-mer de VelvetOptimizer dio lugar a 153 contigs, que corresponden al 98,3% de la secuencia PacBio utilizada como referencia. La comparación del ensamblaje optimizado con el ensamblaje automático realizado por una empresa de bioinformática reveló que, aunque el porcentaje de secuencia fue similar en los dos casos (98,3% versus 98,5%), el número de contigs fue menor (153 vs 185) y la calidad de la secuencia considerablemente superior en el primero. Finalmente, se realizó una comparación entre la anotación funcional de la secuencia de pUO-STmRV1, realizada de forma automática con el servidor RAST y de forma manual. RAST identificó 261 genes pero solo asignó una posible función a 52 de ellos (15%). La anotación manual detectó un menor número de genes (214) pero permitió asignar una función a 135 (63%).

Por todo ello, pese a que los sistemas de automatización suponen un gran avance en el análisis de secuencias, aún presentan limitaciones importantes, siendo de capital importancia la supervisión y comprobación de los resultados por parte de un especialista.

# ABSTRACT

This work focused on the resistance-virulence plasmid pUO-STmRV1, characteristic of the Spanish clone of the monophasic 4,[5],12:i:- variant of *Salmonella enterica*. This variant is a major foodborne pathogen. The aim of the study was to apply different bioinformatic tools to compare sequence data from the plasmid generated by different technologies of first (Sanger), second (Illumina and Roche 454) and third (SMART-PacBio) generation. The data provided by Sanger and Roche 454 has been combined in a previous work. For Illumina and SMART-PacBio the plasmid sequence was obtained as part of the sequence of the entire genome of the bacterium that carries it.

In the first part of the work quantity and quality of the sequences obtained by these techniques were evaluated. The best results were achieved with SMRT-PacBio, which provided the complete sequence of both the chromosome (4907248 bp) and the plasmid (202354 bp) in single contigs. The sequence of the plasmid obtained by Illumina, Roche 454 and Sanger/Roche 454 was 82.7%, 78.5% and 97.7% of the sequence of PacBio, distributed in 21, 30 and 3 contigs, respectively. In the second part, the Illumina reads (chromosome and plasmid) were assembled *de novo* using Velvet and VelvetOptimiser. Manual optimization of the k-mer parameter of VelvetOptimizer resulted in 153 contigs, which correspond to 98.3% of the PacBio sequence used as reference. Comparison of the optimized assembly with the automatic assembly performed by a bioinformatics company revealed that, although the percentage of sequence was similar in both cases (98.3% versus 98.5%), the number of contigs was lower (153 vs 185) and the quality of the sequence significantly higher in the former. Finally, a comparison between the functional annotations of the sequence pUO-STmRV1, performed automatically with the RAST server and manually, was done. RAST identified 261 genes but a possible function was only assigned to 52 of them (15%). The manual annotation identified a smaller number of genes (214) but assigned a function to 135 (63%).

Therefore, although automation systems represent an important breakthrough in sequence analysis, they still have important limitations, being crucial the monitoring and verification of the results by a specialist.

# Lista de figuras

# Lista de tablas

# Introducción

*Salmonella enterica* constituye una de las principales causas de gastroenteritis bacteriana en el mundo, siendo el segundo patógeno más frecuente notificado al Sistema de Información Microbiológica en España desde el año 2006, tan sólo por detrás del género *Campylobacter* (Figura 1) (Boletín Epidemiológico Semanal). En el año 2014, según los datos publicados en el Boletín Epidemiológico Semanal se notificaron al Sistema de Información Microbiológica (SIM) un total de 5008 casos de salmonelosis humana, lo que representa un ligero aumento respecto a años anteriores los años anteriores.

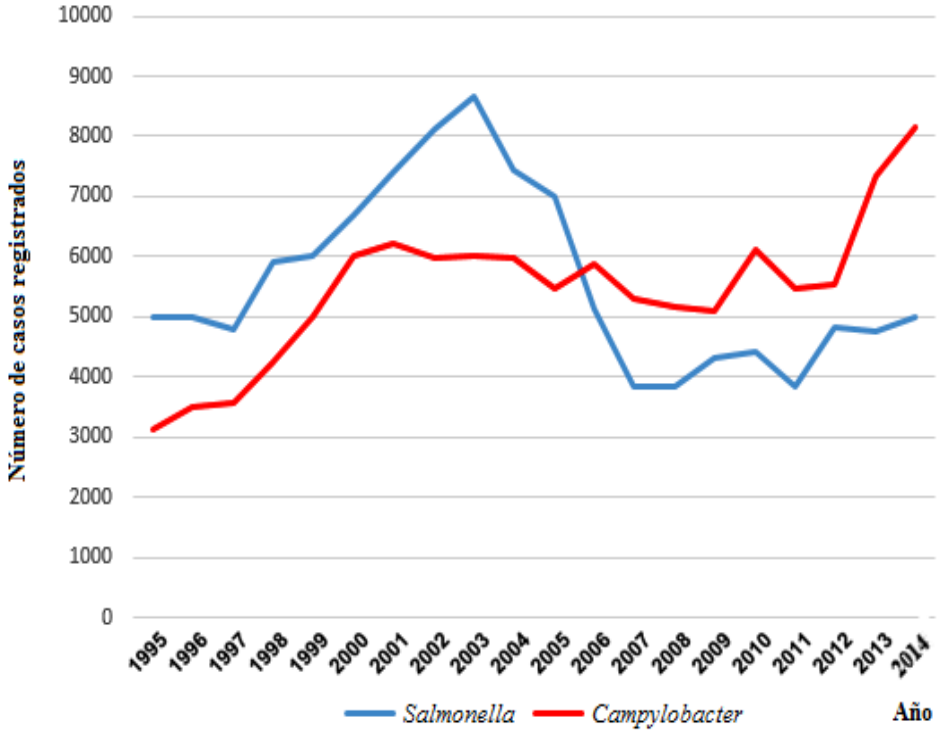


Figura 1. Evolución de los casos de salmonelosis y campylobacteriosis en España durante el periodo 1995-2014. En la gráfica se representa el número de casos registrados por año publicados en el Boletín Epidemiológico Semanal.

La mayor parte de los casos de salmonelosis registrados tanto en España como en otros países europeos se deben a los serotipos Enteritidis y Typhimurium. Sin embargo, la incidencia de estos dos serotipos presenta tendencias diferentes. Mientras que los casos atribuidos a Enteritidis han disminuido considerablemente en los últimos años, los causados por Typhimurium han aumentado (Figura 2). Por ello, el serotipo Typhimurium presenta un especial interés.

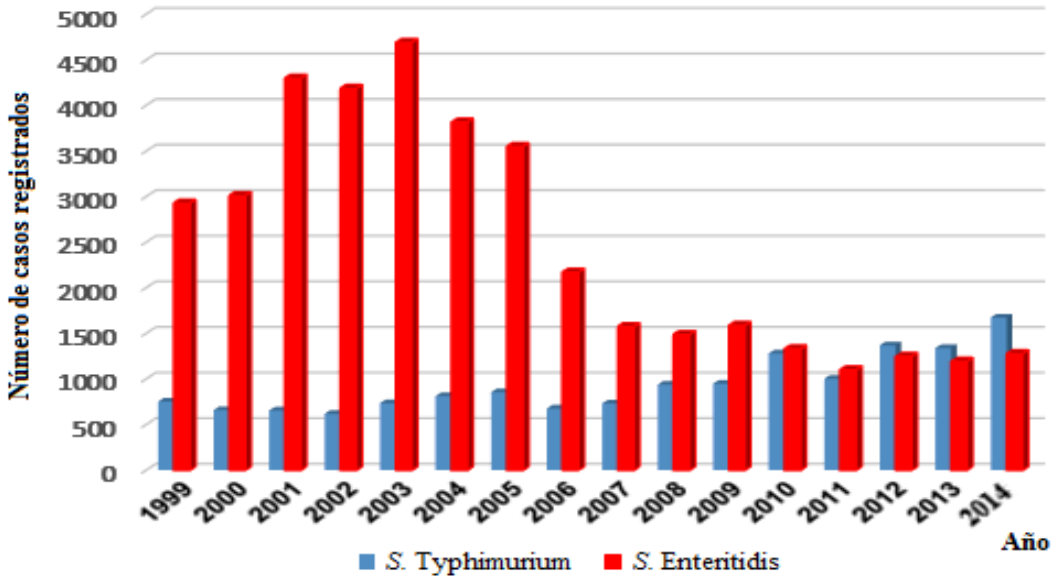


Figura 2. Casos de salmonelosis registrados durante el periodo 1999-2014 causados por los serotipos Enteritidis y Tyhimurium (Fuente: Boletín Epidemiológico Semanal).

Además, en las últimas décadas, se ha detectado en diversos países de Europa, incluyendo España, una variante monofásica del serotipo Typhimurium o serotipo 4,[5],12:i:-, que se caracteriza por carecer del antígeno flagelar de segunda fase. Esta variante representa una grave amenaza para la salud pública debido al aumento de su incidencia en diversos países, donde ocasionan miles de casos de salmonelosis al año. En España, de acuerdo con los datos publicados en el Boletín Epidemiológico Semanal, esta variante es el agente causal de entre 40-83 casos de salmonelosis al año.

Como se verá más adelante, esta variante monofásica de Typhimurium se puede clasificar en tres tipos en función de sus perfiles de resistencia, el clon europeo, el clon español y el recientemente descrito clon sud-europeo, todos ellos multirresistentes.

El clon español, detectado inicialmente en nuestro país (Echeita *et al*., 1999), se encuentra ampliamente distribuido en la península Ibérica. La mayoría de los aislamientos de este clon son resistentes a ampicilina, cloranfenicol, estreptomicina, gentamicina, sulfonamidas, tetraciclina y trimetoprim, aunque existen variaciones (Echeita *et al*., 1999; Guerra *et al*., 2001; Antunes *et al*., 2011; García *et al*., 2011; 2013). Los genes responsables de las resistencias son *bla*$_{\text{TEM-1}}$, *cmlA1*, *aadA1*, *aadA2*, *aac(3)-IV*, *sul1*, *dfrA12* y *tet*(A), respectivamente (Guerra *et al*., 2001; García *et al*., 2011; 2013). Estos genes se localizan en plásmidos de gran tamaño (110 a 220 Kb) del grupo de incompatibilidad IncA/C (± IncN) que, en su mayoría portan el locus *spv* (*spvRABCD*)

junto con otros genes característicos del plásmido de virulencia pSLT específico del serotipo Typhimurium (Guerra *et al*., 2001; García *et al*., 2011; García *et al*., 2013). Contienen, además, numerosos elementos genéticos móviles que aportan los genes de resistencia, como integrones de clase 1 de tipo *sul1* y *sul3*, versiones intactas o delecionadas de elementos genéticos transponibles (Tn*3*, Tn*21* y Tn*1721*) y secuencias de inserción (IS*CR2*, IS*CR3*, IS*Ecp1* e IS*26* (García *et al*., 2011; García *et al*., 2013).

Para conocer en profundidad tanto la organización como las funciones conferidas por los plásmidos IncA/C del clon español se procedió a la secuenciación del plásmido pUO-STmVR1, seleccionado como representante del grupo (García, Tesis Doctoral, 2013). Este plásmido, de 200 Kb aproximadamente, se encontró en el aislamiento más antiguo del clon español detectado en Asturias (LSP 389/97). Para la secuenciación del plásmido se utilizó inicialmente una combinación de secuenciación cíclica de Sanger y pirosecuención Roche 454, de primera y segunda generación, respectivamente. Hasta el momento la secuencia disponible comprende 197 kb. El 30% se corresponde con parte del genoma conservado común a plásmidos A/C, mientras que el 70% restante lo componen regiones accesorias. Se identificaron un total de 219 *orfs*, de las que 77 constituyen el genoma core siendo responsables de las funciones biológicas básicas del plásmido, como replicación, mantenimiento y transferencia por conjugación. Las regiones accesorias, seis en total, incluyen principalmente genes y elementos genéticos implicados en resistencia, genes de virulencia y genes relacionados con replicación y mantenimiento de plásmidos de otros grupos de incompatibilidad como IncN (García , Tesis Doctoral, 2013).

A pesar de los esfuerzos realizados, el gran tamaño y la elevada complejidad de pUO-STm-VR1, que contiene numerosas regiones repetidas, impidieron conseguir la secuencia completa. Para tratar de completar la secuencia del mismo y profundizar en el conocimiento del clon español, se recurrió a la secuenciación del genoma completo de un aislamiento escogido como representante de dicho clon (LSP 398/97) utilizando para ello métodos de segunda (Illumina) y tercera (SMART-PacBio) generación.

## Objetivos

En base a lo expuesto, el **objetivo general** del presente trabajo fue aplicar distintas herramientas bioinformáticas para comparar e integrar la información sobre la secuencia de pUO-STmRV1 obtenida mediante distintas plataformas.

Los **objetivos concretos** fueron los siguientes:

- Evaluar la cantidad y calidad de la secuencia del plásmido pUO-STmRV1 obtenida usando los distintos métodos de secuenciación.
- Ensamblar la secuencia del genoma del aislamiento LSP 398/97, escogido como representante del clon español de la variante monofásica *S. enterica* 4,[5],12,i:-, obtenida con la plataforma Illumina. Comparar dicho ensamblaje con el realizado por una empresa de bioinformática.
- Comparar las secuencias de pUO-STmRV1 obtenidas mediante las diferentes técnicas de secuenciación.
- Anotar la secuencia del plásmido.

# Consideraciones teóricas

## 2.1. Género *Salmonella*: Características generales y clasificación

El género *Salmonella* pertenece al filum *Proteobacteria*, orden *Enterbacteriales*, familia *Enterbacteriaceae*. Incluye bacterias Gram negativas de morfología bacilar, generalmente móviles mediante flagelos perítricos (Figura 3A), cuyo tamaño oscila entre 0,7 x 2 y 1,5 x 5 µm (Corral y Perea, 1992). Son organismos anaerobios facultativos, con metabolismo oxidativo y fermentativo, siendo capaces de fermentar compuestos carbonados, especialmente glucosa, como fuente de carbono con producción de gas, pero no fermentan lactosa. Tienen la capacidad de utilizar el sulfato del medio reduciéndolo, con la consiguiente formación de ácido sulfhídrico, el cual les confiere el color negro que caracteriza a sus colonias si las condiciones del medio son las adecuadas (Figura 3B). Además presentan resistencia a sales biliares, lo cual permite utilizar medios selectivos para su aislamiento. Se trata de organismos oxidasa negativos y catalasa positivos. Los miembros de este género se encuentran ampliamente distribuidos, localizándose en el tracto intestinal de mamíferos, aves, reptiles e insectos; algunos están especialmente adaptados a su hospedador estableciendo una relación de comensalismo. Otros, sin embargo, invaden el tracto gastrointestinal del hospedador, en el cual actúan como patógenos, causando un cuadro clínico complejo y en múltiples ocasiones grave, como es el caso de las infecciones en humanos. Pueden sobrevivir en un amplio rango de temperaturas (5-45ºC) aunque su temperatura óptima se encuentra entre los 35 y 37ºC. Son capaces de tolerar un rango de pH entre 4,5-9, creciendo con mayor facilidad con valores de pH de 6,5-7,5 (Mossel *et al.*, 2002).


Figura 3. Género *Salmonella*. (A) Célula de *Salmonella* con morfología bacilar y flagelación perítrica (Tomado de http://www.sosenfermero.com/noticias-de-salud/profesionales/salmonella/). (B) Colonias de color negro de *Salmonella enterica* debido a la producción de $H_2S$ en agar Xylose Lysine Desoxycholate (XLD) (Tomado de http://www.ispch.cl/lab_amb/serv_lab/salmonella.html).

Antiguamente se consideraba la existencia de numerosas especies dentro del género *Salmonella*, hoy en día se acepta que el género *Salmonella* consta de dos especies, *S. enterica* y *S. bongori*. La primera especie está dividida en seis subespecies: *S. enterica* subespecie *enterica* (subespecie I), *S. enterica* subespecie *salamae* (subespecie II), *S. enterica* subespecie *arizonae* (subespecie IIIa), *S. enterica* subespecie *diarizonae* (subespecie IIIb), *S. enterica* subespecie *houtenae* (subespecie IV) y *S. enterica* subespecie *indica* (subespecie VI). *S. bongori* es el miembro más antiguo del género y se compone únicamente por la subespecie V (Tindall *et al*., 2005; Grimont y Weill, 2007). En función de sus características antigénicas *S. enterica* y *S. bongori* se pueden clasificar en serogrupos y serotipos.

*Salmonella* presenta una estructura antigénica principalmente compuesta por tres tipos de antígenos: antígenos somáticos (O), antígenos flagelares (H) y antígenos capsulares (K o Vi). En función de estos antígenos, siguiendo el esquema Kauffmann-White Le Minor, se han identificado 67 serogrupos definidos por los antígenos O mayores y más de 2600 serotipos establecidos en función de las diferentes combinaciones de antígenos O, K y H, de los cuales más de 1500 pertenecen a la subespecie I de *S. enterica* (Grimont y Weill, 2007).

Los **antígenos somáticos (O)** se localizan en la membrana externa de la pared celular de bacterias Gram negativas. Son moléculas de naturaleza polisacarídica, termoestables y alcohol resistentes, que constituyen la fracción más externa del lipopolisacárido (LPS) (Figura 4). Se pueden distinguir dos tipos de antígenos O, los mayores y los menores. Los antígenos O mayores son conocidos también como factores principales y determinan el grupo antigénico o serogrupo. Los antígenos O menores son conocidos también como factores secundarios. Pueden ser compartidos por varios serogrupos. Cambios químicos producidos en un antígeno mayor pueden originar uno menor, aunque también se pueden producir por conversión fágica, un proceso conocido como exteriorización ligada a la presencia de un bacteriófago.

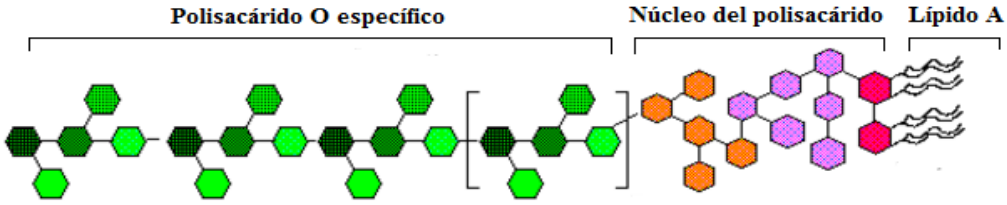

Figura 4. Estructura del lipopolisacárido donde se observa la región correspondiente al antígeno O (Tomado de http://ftp.utalca.cl/profesores/raherrera/biorom2008/contenido/cibertexto/hc/sugar35b.htm).

Los **antígenos flagelares (H)** se deben a diferencias en la flagelina, proteína estructural de los flagelos bacterianos. Son termolábiles y vienen definidos por la región central de los monómeros de flagelina debido a su variabilidad. La gran mayoría de los serotipos son bifásicos, es decir, expresan dos tipos de antígenos flagelares, denominados fases. El antígeno flagelar puede aparecer de forma alternativa en fase I o específica, que es característica de serotipo, o en fase 2, menos específica, ya que puede presentarse en varios serotipos. Es importante destacar que en un determinado momento una bacteria sólo puede expresar un antígeno flagelar concreto. El mecanismo de cambio de fase se debe a la presencia de una recombinasa codificada por el gen *hin*, capaz de invertir reversiblemente un fragmento de 993 pb que contiene el gen *hin* y el promotor del operón *fljAB* (Figura 5). En un sentido se expresan los genes *fljA* y *fljB*, los cuales codifican la flagelina de segunda fase, y un represor del gen *fliC*, que codifica la flagelina de primera fase y se encuentra localizado en una región diferente del cromosoma. Cuando se invierte el fragmento, estos genes no se expresan, por lo que *fliC* no estará reprimido y se produce el antígeno de primera fase. El cambio de fase se relaciona con la evasión de la respuesta inmune, ya que si ésta reconoce la flagelina de primera fase, la bacteria puede escapar cambiando de fase y produciendo la flagelina de segunda fase.

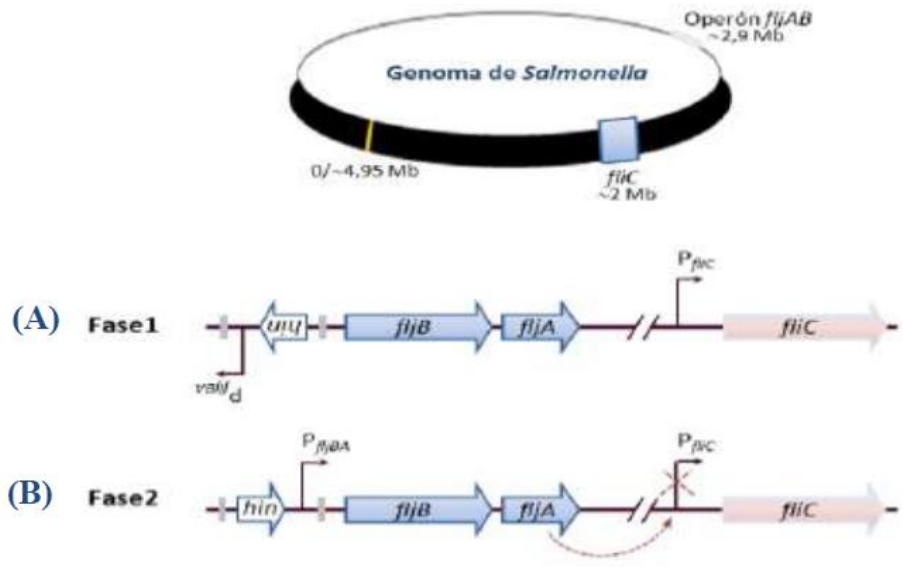


Figura 5. Localización de los genes que codifican las flagelinas de fase 1 y de fase 2 en el cromosoma de *Salmonella enterica* (A) y mecanismo de alternancia de fases (B). (tomado de García, Tesis doctoral, 2013).

El **antígeno capsular (K o Vi)** localizado en la cápsula bacteriana es un polisacárido termolábil. Solamente ha sido descrito en tres serotipos altamente invasivos, Typhi, Paratyphi *C* y algunos aislamientos de Dubin. Se trata del antígeno de virulencia (Vi), localizado en el locus *viaB* de la isla de patogenicidad SPI-7, gracias al cual las bacterias pueden evadir el sistema inmune del hospedador (Raffatellu *et al.*, 2006; Raffatellu *et al.*, 2008).

Desde el punto de vista médico se distinguen los serotipos tifoideos y no tifoideos. Dentro de los tifoideos se incluyen *Salmonella enterica* serotipo Typhi (*S.* Typhi), *S.* Paratyphi A, B, C y *S.* Sendai. Están ampliamente adaptados al ser humano y son responsables de la fiebre tifoidea o paratifoidea. Como excepción, *S.* Paratyphi B tiene amplio rango de hospedador y no causa fiebre entérica por lo que se considera junto con los serotipos no tifoideos. Estos serotipos originan normalmente gastroenteritis, aunque también son capaces de producir infecciones locales extraintestinales e infecciones sitémicas. Algunos de ellos son altamente específicos como es el caso de *S.* Gallinarum, *S.* Dublin *S.* Choleraesuis que colonizan a gallinas, cerdos y vacas, respectivamente, aunque también pueden causar infecciones en el ser humano. Sin embargo, la mayoría de los serotipos no tifoideos no presentan un hospedador específico y están ampliamente distribuidos (Uzzau et al., 2000; Fierer y Guinei, 2001).

## 2.2. Enfermedades causadas por *Salmonella*, mecanismos de trasmisión y factores de virulencia

*Salmonella* es un importante patógeno de seres humanos, siendo la mayoría de las infecciones causadas por serotipos de *S. enterica* subsp. *enterica*. Desde el punto de vista médico, éstos se clasifican en tifoideos y no tifoideos. Dentro de los tifoideos se incluyen Typhi, Paratyphi (A, B y C) y Sendai. Están específicamente adaptados al ser humano y son responsables de enfermedades invasivas graves, conocidas como fiebres entéricas, tifoideas o paratifoideas. Por el contrario, la mayoría de los serotipos no tifoideos tienen un amplio rango de hospedador, pudiendo colonizar o causar enfermedad en seres humanos y diversos animales (Corral y Perea, 1992; Fierer y Guiney, 2001).

Los serotipos no tifoideos son bacterias intracelulares facultativas que se transmiten por la vía fecal-oral. La cadena de transmisión es compleja (Figura 6). La bacteria llega a los animales domésticos a través de diferentes vías: contaminación de pastos y aguas, piensos elaborados con restos de animales enfermos o portadores de la bacteria, contacto directo con animales silvestres portadores o con sus heces, entre otras. Además, los actuales sistemas de cría intensiva de ganado y aves facilitan la propagación de la bacteria tanto de unos animales a otros como a sus productos (leche y huevos). Desde los animales *Salmonella* puede llegar al ser humano por diferentes vías:

1. Consumo de leche y derivados lácteos con contaminación fecal.
2. Consumo de huevos contaminados o de sus derivados, siendo este uno de los vehículos de transmisión más importantes. Normalmente la contaminación tiene lugar de forma externa durante la puesta, debido a que el oviducto y la cloaca desembocan en el mismo conducto. Sin embargo, en ocasiones la contaminación es transovárica, ocurriendo durante la formación del huevo.
3. Consumo de carne de mamíferos y aves, así como subproductos, cuya contaminación transcurre durante el proceso de evisceración en los mataderos donde la carne entra en contacto con el contenido intestinal.
4. Contaminación exógena o cruzada de cualquier tipo de alimento que, en principio no contiene a la bacteria, sino que puede llegar a él a través del contacto con utensilios y superficies contaminadas, agua incorrectamente tratada o por los manipuladores de alimentos.
5. Por último, pero no menos importante, también se puede producir una transmisión fecal- oral de persona a persona.

En seres humanos, dependiendo del serotipo y del estado del hospedador, *Salmonella* puede ser eliminada por el sistema inmune del hospedador sin llegar a causar enfermedad, establecer un estado de portador asintomático o provocar distintos tipos de enfermedad.
La enfermedad más frecuente causada por los serotipos no tifoideos es la inflamación intestinal aguda conocida comúnmente como salmonelosis, que puede afectar al intestino delgado (enteritis) y/o al intestino grueso (enterocolitis). Después de un periodo de incubación de 6-48 horas, aparecen síntomas como náuseas y vómitos, seguidos de dolor abdominal y deposiciones diarreicas. Normalmente cursa con fiebre, ya que se trata de una bacteria enteroinvasiva, y cólicos abdominales. En individuos adultos sanos la

enfermedad es generalmente autolimitada y no precisa tratamiento con agentes antimicrobianos. Sin embargo, en niños, ancianos y personas inmunodeprimidas las infecciones intestinales pueden evolucionar a infecciones extraintestinales focales o sistémicas, potencialmente graves e incluso mortales, en cuyo caso sí es necesaria la administración de antimicrobianos. Entre los más de 2000 serotipos no tifoideos identificados hasta el momento, unos pocos causan la mayoría de las infecciones en seres humanos, siendo Enteritidis y Typhimurium los más frecuentes.

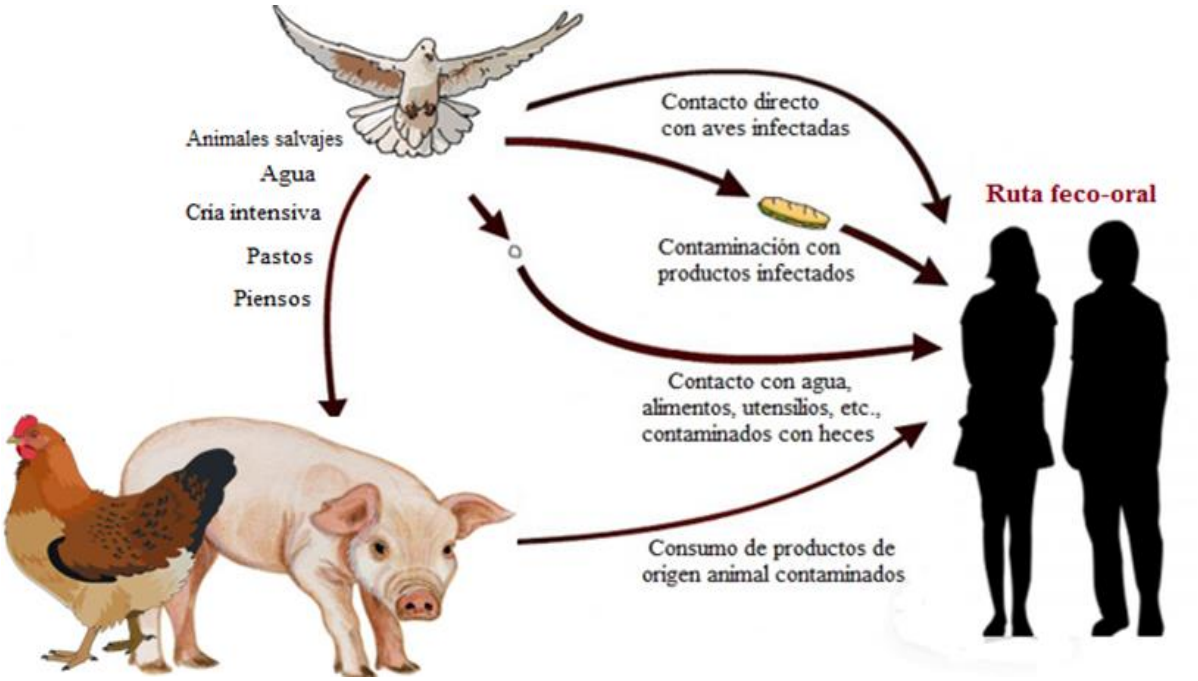


Figura 6. Rutas de transmisión de los serotipos no tifoideos de *Salmonella*.

La capacidad de causar enfermedad depende en gran medida de la dosis infectiva (entre 15 y 20 bacterias, en el caso de los serotipos no tifoideos) y de los factores de virulencia (Figura 7). Estos últimos están codificados por genes localizados en el cromosoma bacteriano o en plásmidos (Schmidt y Hensel, 2004).

Los factores de virulencia de los serotipos no tifoideos se pueden clasificar en dos grupos: aquellos que codifican estructuras de superficie, y factores específicos que modifican la fisiología de las células del hospedador y/o protegen a la bacteria de la respuesta inmune (Fierer y Guiney, 2001). Entre los primeros se encuentran el lipopolisacarido, los flagelos (ya mencionados anteriormente) y las fimbrias.

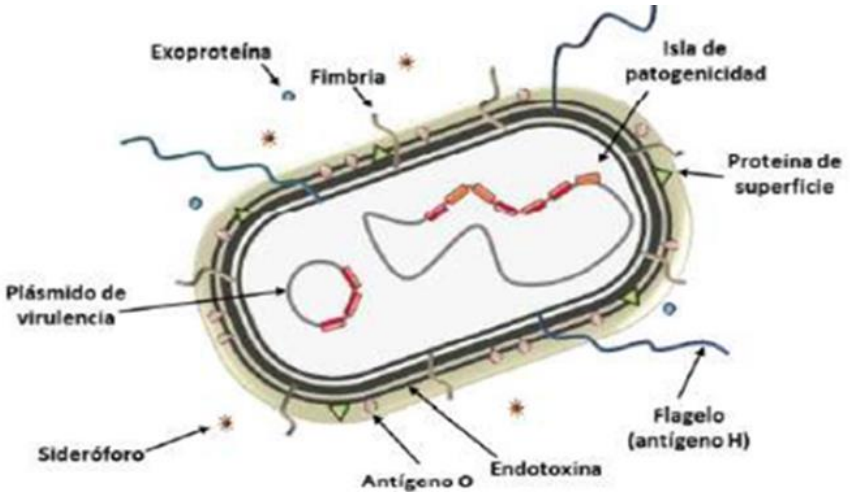


Figura 7. Representación esquemática de los diferentes factores de virulencia de *Salmonella enterica* (tomado de García, Tesis doctoral, 2013).

Las **fimbrias** son estructuras proteicas situadas en la superficie de las bacterias y formadas por la polimeración de proteínas denominadas fimbrinas. Son importantes en la adhesión a componentes del hospedador, la colonización e invasión celular, el desarrollo de biofilms, etc. (Gibson *et al.*, 2007). El análisis del genoma completo de *S.* Typhimurium puso de manifiesto la existencia de 13 operones, cuyos genes codifican las subunidades estructurales y las enzimas que participan en la formación y ensamblaje de las fimbrias (Humphries *et al.*, 2003; Weening *et al.*, 2005). Muchos de los factores específicos están codificados por **islas de patogenicidad** constituidas por largas agrupaciones de genes de virulencia, localizadas en el genoma de bacterias patógenas, pero ausentes en bacterias no patógenas de la misma especie o especies relacionadas. En *Salmonella* reciben el nombre de SPIs (*Salmonella* Pathogenicity Islands), siendo SPI-1 y SPI-2 las más importantes (Schmidt y Hensel, 2004).

SPI-1 (40 Kb) presenta un contenido guanina-citosina (GC) inferior al que tiene de media el genoma de *Salmonella* (52%). No se encuentra asociada a ningún gen de ácido ribonucleico de transferencia (ARNt), y es necesaria para el proceso de infección intestinal (Marcus *et al*., 2000; Schmidt y Hensel, 2004). Codifica un sistema de secreción de tipo 3 (SPI-1 SST3), el cual introduce una serie de proteínas efectoras, no todas codificadas por SPI-1, al interior celular. Las proteínas efectoras SopE, SopE2 y SptP median la entrada de la bacteria en los enterocitos, a través de la modificación del citoesqueleto de actina. Otro grupo de proteínas efectoras, SopA, SopB y SopD, son

responsables de los síntomas intestinales debido a que activan cascadas de transducción provocando la liberación de citiquinas proinflamatorias.

SPI-2 (también de 40 Kb) también presenta un contenido GC inferior al del genoma de *Salmonella*, se asocia al gen *valV,* el cual codifica el ARNt$^{val}$ específico para la valina. Codifica un segundo SST3 (SPI-2 SST3), cuyos efectores están relacionados con la supervivencia y multiplicación de la bacteria en los enterocitos y macrófagos, siendo esencial en el establecimiento de infecciones sistémicas (Schmidt y Hensen, 2004).

Otros genes de virulencia se localizan en loci más pequeños, denominados **islotes de patogenicidad** o se presentan aislados en el cromosoma bacteriano. Los productos de estos genes están implicados en la supervivencia o multiplicación en el interior de los macrófagos, codifican una enterotoxina e intervienen en la captación del hierro o en la regulación (Fierer y Guiney, 2001).

Finalmente, existen **plásmidos de virulencia**, aunque solamente se encuentran en siete serotipos no tifoideos pertenecientes a la subespecie *enterica.* Todos ellos contienen el locus *spv* (*Salmonella* plasmid virulence) que es importante para la multiplicación intracelular de la bacteria durante la infección sistémica **(**Fierer y Guiney, 2001; Fierer y Guiney, 2011). Consta de cinco genes (*spvRABCD*) regulados por las condiciones que se dan en el interior de los macrófagos. El plásmido de virulencia del serotipo Typhimurium tiene 94 Kb y se denomina pSTV o pSLT, siendo Typhimurium LT2 la cepa tipo del género *Salmonella.*

## 2.3. Resistencia a antimicrobianos

La existencia de aislamientos resistentes y multirresistentes (resistentes a tres o más antimicrobianos pertenecientes a diferentes familias) de *S. enterica* ha aumentado mucho en los últimos años debido en gran parte, al uso abusivo de sustancias antimicrobianas. La capacidad de las bacterias para resistir a dichas sustancias constituye un evento evolutivo constante; es decir, son capaces de desarrollar mecanismos para evadir la acción de los antimicrobianos, lo cual representa una gran amenaza tanto para animales como para humanos (Torres *et al.,* 2010). Los mecanismos empleados para ello se pueden clasificar en dos grupos (Cantón *et al*., 2003; Alekshun y Levy, 2007):

**1) Mecanismos de resistencia innata** o **intrínseca,** en los cuales son las propias características específicas de la bacteria las responsables de su resistencia natural.

**2) Mecanismos de resistencia adquirida,** en los cuales la resistencia se consigue bien por una mutación de un gen cromosómico cuya transferencia es vertical, esto es, de una bacterias a sus descendientes, u horizontal, entre bacterias más o menos relacionadas, presentes en un mismo nicho. Esta última es la principal responsable de la resistencia y multirresistencia a antimicrobianos. La transferencia horizontal se sirve de elementos genéticos móviles, entre los que se encuentran el sistema integrón-casete génica, los transposones, los plásmidos y las islas genómicas.

Los **integrones** son sistemas naturales de clonación y expresión, que incorporan una o más casetes génicas y las convierten en genes funcionales. Las casetes génicas son pequeños elementos móviles formados por una pauta abierta de lectura (*orf*) carente de promotor. Se localizan en la región variable del integrón, que frecuentemente codifica resistencia a antimicrobianos. En los integrones de clase 1, que son los más frecuentes, la región variable se encuentra flanqueada por segmentos 5' y 3' constantes (5'CS y 3'CS). El segmento 5'CS aporta un gen que codifica una integrasa (*intI*) y uno o más promotores para la expresión de las casetes (Figura 8A). El segmento 3'CS contiene genes de resistencia a compuestos derivados de amonio cuaternario (*qacEΔ1*), sulfonamidas (*sul1*) y una *orf* de función desconocida (*orf5*) (Fluit y Schmidt, 2004). Existen también integrones de clase 1 con segmentos 3'CS diferentes, que contienen el gen *sul3* en vez del gen *sul1* de resistencia a sulfonamidas y que se clasifican en varios tipos dependiendo de la región variable (Figura 8B).

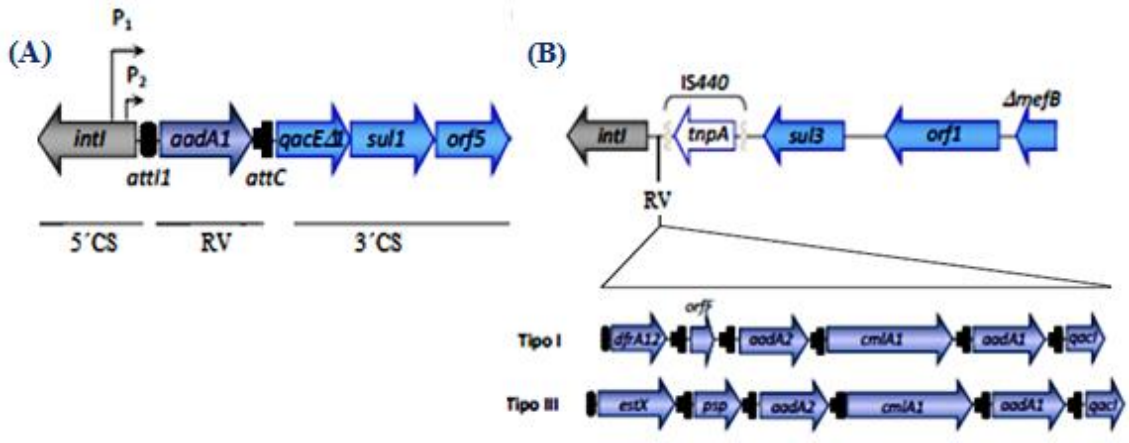


Figura 8. Representación esquemática de integrones de clase 1 de tipo *sul1* (A) y de tipo *sul3* (B). RV: Región Variable; CS: Segmento Conservado; IS: Secuencia de Inserción; *orf*: pauta abierta de lectura (Open Reading Frame) (modificado de García, Tesis doctoral, 2013).

Los **transposones** son elementos genéticos móviles que poseen genes que codifican una enzima transposasa y proteínas reguladoras, que conforman la maquinaria necesaria para promover su propia translocación desde un sitio donador a otro receptor. Además, pueden contener genes de resistencia a antimicrobianos, desempeñando un papel fundamental en bacterias Gram negativas (Kleckner, 1981). Se pueden clasificar en dos grupos:

1) Transposones de clase 1 o compuestos, los cuales presentan una estructura modular formada por dos copias de la misma secuencia de inserción flanqueando una región central variable, la cual incluye genes de resistencia a antimicrobianos (Figura 9A). Las secuencias de inserción son los elementos genéticos transponibles más sencillos, que sólo contienen los genes necesarios para su propia transposición delimitados por repeticiones terminales invertidas.
2) Transposones de clase 2 o complejos, que están flanqueados por repeticiones terminales invertidas de 38-40 Kb, y contienen genes que codifican una transposasa (*tnpA*) y una resolvasa (*tnpR*), junto con una secuencia *res*, todos ellos implicados en transposición, además de genes accesorios que pueden conferir resistencia a antimicrobianos (Figura 9B).

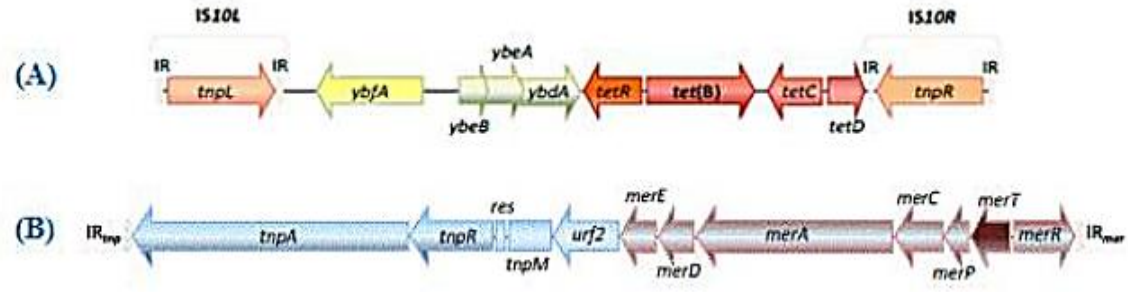


Figura 9. Representación esquemática del transposón compuesto *Tn10* (A) y del transposón complejo *Tn21* (B) (tomado de García, Tesis doctoral, 2013).

Los **plásmidos** son elementos génicos extracromosómicos formados por ácido desoxirribonucleico (ADN) de doble cadena, generalmente circulares y con capacidad de replicación autónoma, cuya herencia es estable. Además de los genes implicados en funciones esenciales para su replicación y mantenimiento, pueden llevar genes accesorios que codifican funciones metabólicas, de virulencia (como pSLT de Typhimurium, mencionado anteriormente), y de resistencia a antimicrobianos. Los plásmidos que no aportan una función conocida a la bacteria que los contiene se denominan crípticos. Los plásmidos se encuentran en casi todas las especies bacterianas y su tamaño es variable,

de 2-400 Kb (Bennett, 2008; Caratolli, 2009). Estos elementos pueden diseminarse de unas bacterias a otras a través de la conjugación, mecanismo replicativo por el cual el plásmido se transfiere de una bacteria donadora que lo mantiene, a otra receptora. Pueden integrarse en el cromosoma bacteriano y desempeñan una función esencial en la transmisión horizontal de la resistencia.

En *S. enterica* es frecuente encontrar plásmidos de resistencia pertenecientes a distintos grupos de incompatibilidad (IncF, IncA/C, IncL/M, IncI1, IncR e IncP; Caratolli, 2009). Los plásmidos de un mismo grupo no pueden mantenerse de forma estable en una línea celular debido a que compiten por la maquinaria biosintética de la célula hospedadora. Sin embargo, se puede dar el caso de que una misma bacteria presente dos o más plásmidos compatibles (Benett, 2008).

Cabe destacar que los plásmidos de resistencia pueden ganar genes de virulencia y algunos plásmidos de virulencia específicos de serotipo, como pSLT de Typhimurium, han adquirido genes de resistencia. Se originan así, plásmidos híbridos de virulencia-resistencia, permitiendo la coselección y co-transferencia de ambas funciones (Rodicio *et al*., 2011).

## 2.4. Características de la variante monofásica del serotipo Typhimurium

En las últimas décadas se ha detectado, tanto en España como en muchos otros países, una variante monofásica del serotipo Typhimurium que carece del antígeno flagelar de segunda fase, debido a distintas mutaciones y/o delecciones, que afectan a la región del operón *fljAB*. Su fórmula antigénica es 4,[5],12,i:-, donde 4, [5] (5 puede estar presente o no) y 12 son antígenos somáticos, i es el antígeno flagelar de primera fase y – indica la ausencia del antígeno flagelar de segunda fase (1,2 en el caso de Typhimurium).

Esta variante representa un riesgo para la salud pública a nivel mundial, debido al aumento de su incidencia en todos los continentes (Switt *et al*., 2009; EFSA, 2010; Echeita *et al*., 2011). En función de los perfiles de resistencia a antimicrobianos la variante monofásica se puede clasificar en varios tipos, entre los que destacan el clon europeo y el clon español, ambos multirresistentes.

El **clon europeo** es una de las principales causas de infección intestinal en diversos países de Europa, entre los que se incluyen España, Francia, Alemania y Reino Unido, entre otros. Los aislamientos de este clon pertenecen, en su mayoría a los fagotipos DT193 y DT120 y son típicamente resistentes a ampicilina, estreptomicina, sulfonamidas y tetraciclina. Estas resistencias se deben a la presencia de los genes *bla*$_{\text{TEM-1}}$, *strAB, sul2* y *tet*(B), localizados en el cromosoma bacteriano y organizado en dos regiones (RR1 y RR2; Figura 10) flanqueadas por copias de la secuencia de inserción IS*26* (Lucarelli *et al*., 2010). RR1, con 8721 pb, contiene los genes *bla*$_{\text{TEM-1}}$ (presente en una variante delecionada del transposón Tn*3*), *strA*, *strB* y *sul2*, mientras que RR2 de 14587 pb porta el gen *tet*(B) en una variante defectiva del transposón Tn*10*. Los aislamientos del clon europeo carecen de integrones y del plásmido de virulencia específico de Typhimurium, pero pueden contener plásmidos crípticos de pequeño tamaño.

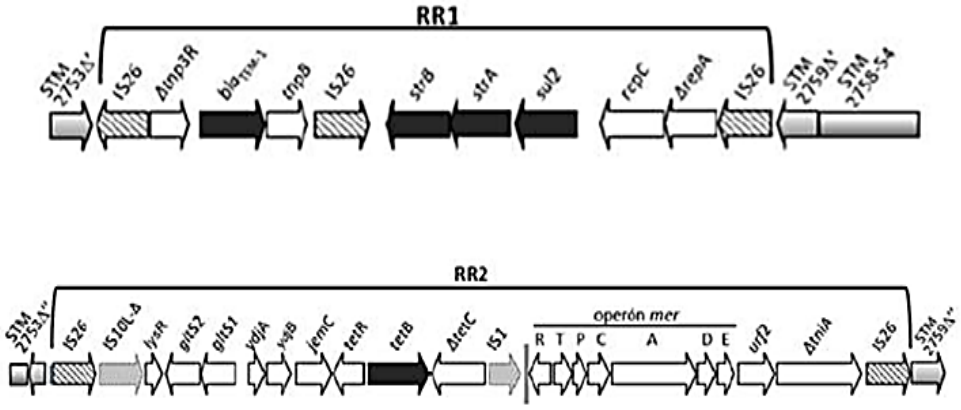


Figura 10. Representación gráfica de las regiones RR1 y RR2 del clon monofásico europeo (tomado de García, Tesis doctoral, 2013).

Por su parte, el **clon español** tiene una mayor incidencia en la Península Ibérica. Los aislamientos de este clon pertenecen mayoritariamente al fagotipo U302, aunque se han asignado también a otros fagotipos como DT193 y DT208. En general son resistentes a un mayor número de antimicrobianos que los pertenecientes al clon europeo, incluyendo ampicilina, cloranfenicol, estreptomicina, gentamicina, sulfonamidas, tetraciclina y trimetoprim (Echeita *et al*., 1999; Guerra *et al*., 2001; Antunes *et al*., 2011; García *et al*., 2011; 2013). Los genes responsables de estas resistencias son *bla*$_{\text{TEM-1}}$, *cmlA1*, *aadA2*, *aac(3)-IV*, *sul1*, y *tet*(A) y *dfrA12*, respectivamente. Estos genes se localizan en plásmidos de gran tamaño (110 a 220 Kb) del grupo de incompatibilidad IncA/C ± IncN, que en su mayoría portan genes del locus *spv* (*spvRABCD*), junto con otros genes característicos de plásmidos de virulencia específicos de serotipo (García *et al*., 2011). En cuanto a

elementos genéticos móviles, estos plásmidos, cuyo representante es pUO-STm-RV1 de aproximadamente 200 Kb, portan tres integrones de clase 1, (dos de tipo *sul1* y otro *sul3* tipo III), versiones intactas o deleccionadas de elementos genéticos transponibles (Tn*3*, Tn*21* y Tn*1721*), así como numerosas copias de la secuencia de inserción IS*26*.

Recientemente se ha descrito un tercer grupo de aislamientos monofásicos con los genes *cmlA1*, *aadA1*/*aadA2*, *sul3* y *dfrA12*, que confieren resistencia a cloranfenicol, estreptomicina, sulfonamidas y trimetoprim, respectivamente. Dichos genes son aportados por integrones *sul3* de tipo I y se localizan en plásmidos del grupo de incompatibilidad IncR de entre 80 a 160 Kb, que contienen también genes de virulencia, incluidos los pertenecientes al locus *spv* (García *et al*., 2014). Estos aislamientos fueron posteriormente asignados al clon sud-europeo que, en base a las deleciones detectadas en la región *fljAB*, podrían estar evolutivamente relacionados con los aislamientos americanos (Mourão *et al*., 2014). Sin embargo, los aislamientos de América carecen de plásmidos IncR y muestran un menor grado de resistencia o son susceptibles (Switt *et al*., 2009).

Finalmente, también se han descrito aislamientos con plásmidos portadores de los replicones IncFIB e IncFIIs, característicos de los plásmidos de virulencia específicos de serotipo. Contienen los genes *spvC*, *rck*, *mig5* y *srgB* de pSLT, del cual podrían haber evolucionado mediante la adquisición de múltiples genes de resistencia. Confieren resistencia frente a ampicilina, cloranfenicol, estreptomicina, sulfonamidas y trimetoprim, gracias a los genes $bla_{\text{TEM-1}}$, *cmlA1*, *aadA1*, *aadA2*, *sul3* y *dfrA12*, respectivamente. El gen $bla_{\text{TEM-1}}$ se encuentra vinculado a Tn*2*, mientras que el resto de genes se asocian a un integrón *sul3* de tipo I. Estos plásmidos de virulencia-resistencia coexisten con otro de 20 Kb, que aporta los genes *strA*, *sul*2 y *tet*(A) de resistencia a estreptomicina, sulfonamidas y tetraciclina, respectivamente (García *et al*., 2014).

## 2.5. Métodos de secuenciación de ADN

Estos métodos tienen como objetivo el conocimiento de la secuencia de nucleótidos de un segmento o una molécula de ADN. La obtención de la secuencia de ADN ha tenido un enorme impacto en el campo de la biología molecular y en otras muchas áreas de la Biología y la Medicina. Sin embargo, a partir de los años 70, tuvo lugar el desarrollo de métodos rápidos y automatizados de secuenciación que sustituyeron a los anteriores, permitiendo que la secuenciación completa de genomas se esté convirtiendo en una potente e importante herramienta para las investigaciones epidemiológicas y permita en un futuro, realizar una identificación precisa y caracterización de los aislamientos bacterianos.

### 2.5.1. Métodos clásicos de secuenciación o de primera generación

Las técnicas pioneras en este campo fueron el método de la secuenciación química de Maxam y Gilbert (Maxam y Gilbert, 1980) y el método enzimático de Sanger (Sanger y Coulson, 1975). Estos dos métodos fueron desarrollados a partir de los años 70 imponiéndose el de Sanger sobre el de Maxam y Gilbert por ser más sencillo y preciso.

El **método enzimático de Sanger** también se conoce como método de los terminadores o método de los didesoxiribonucleótidos (ddNTPs). Los ddNTPs son análogos estructurales de los desoxirribonucleótidos (dNTPs) que pueden ser incorporados por las ADN polimerasas a las cadenas de ADN que están siendo sintetizadas. Sin embargo, en su carbono 3’ no contienen el grupo hidroxilo (-OH), por lo que su incorporación impide la formación del siguiente enlace fosfodiéster, bloqueando la síntesis.

El protocolo inicial de Sanger consistía en preparar 4 mezclas de reacción que contienen el ADN en cadena sencilla, la ADN polimerasa, una mezcla de desoxirribonucleótidos (dNTPs), uno de los cuales se encontraba marcado radioactivamente con $^{35}S$ y en cada mezcla de reacción se añade uno de los ddNTPs. En cada mezcla de reacción se obtiene una variedad de fragmentos marcados de diversos tamaños que comparten el mismo extremo 5’ y presentan el mismo ddNTPs en el extremo 3’. Posteriormente, estas mezclas se separan por electroforesis en geles de poliacrilamida desnaturalizantes, obteniéndose un patrón de bandas que reflejan la distribución de los ddNTPs, y por tanto de los dNTPS análogos, en el ADN recién sintetizado (Figura 11).

Corriendo las cuatro reacciones en paralelo en el mismo gel, y tras su visualización mediante autorradiografía, se obtiene un patrón de bandas a partir del cual se puede obtener la secuencia del ADN (Sanger y Coulson 1975).

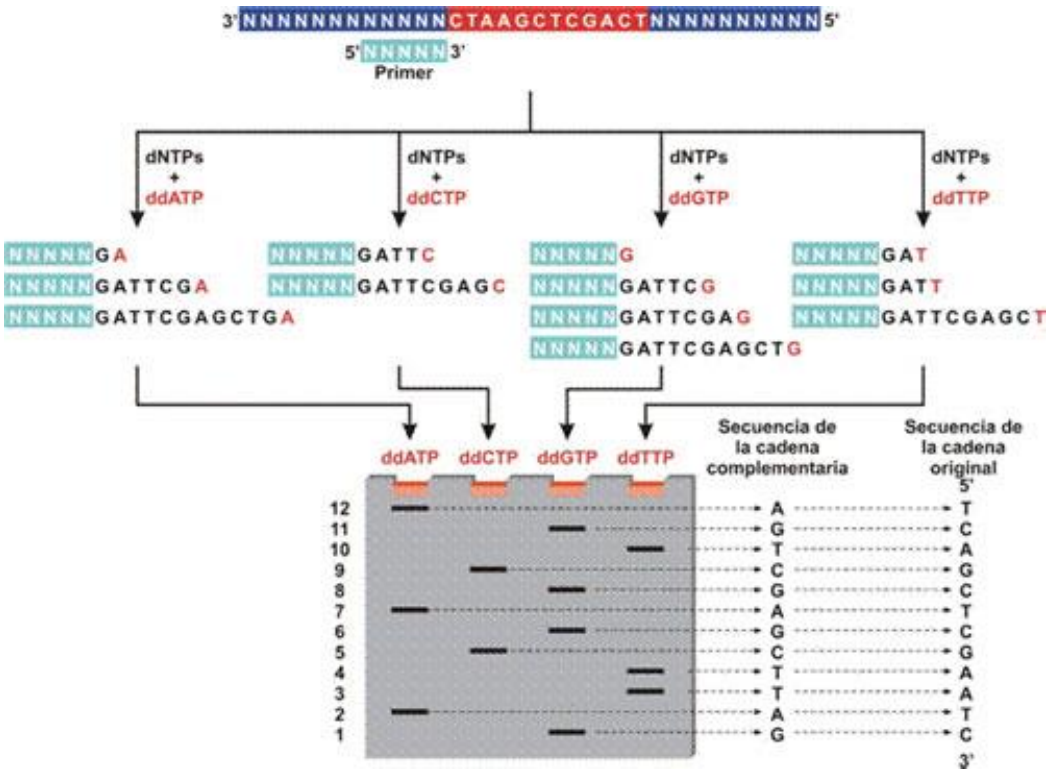


Figura 11. Representación esquemática del método de secuenciación enzimática de Sanger ( Tomado de http://ulum.es/secuenciacion-sanger/).

Posteriormente, con el descubrimiento de la reacción en cadena de la polimerasa (PCR), el proceso anterior fue sustituido por la secuenciación cíclica o automática. Este método emplea dNTPs marcados con fluorocromos diferentes, lo cual permite llevar a cabo las reacciones de secuenciación en el mismo tubo, donde tienen lugar rondas sucesivas de desnaturalización, hibridación y extensión de la cadena por la DNA polimerasa termoestable (Taq) en un termociclador, dando como resultado una amplificación lineal de los productos de extensión. Estos productos se separan por electroforesis capilar capilar para la obtención de su secuencia de nucleótidos. Estas mejoras en el procedimiento de secuenciación, en los reactivos usados y en la tecnología de separación de los fragmentos de ADN generados fueron de suma importancia para la realización del Proyecto Genoma Humano. Sin embargo, este proyecto puso también de manifiesto las limitaciones del método de Sanger para obtener información a gran escala, lo que condujo al desarrollo de nuevos métodos de secuenciación englobados bajo la denominación

general de secuenciación de siguiente generación (NGS; Next Generation Sequencing). El avance en el desarrollo de nuevos métodos de secuenciación ha hecho que se haya actualizado la denominación de los mismos a métodos de segunda generación (SGS) y de tercera generación (TGS), considerando la secuenciación de Sanger como de primera generación.

### 2.5.2. Métodos de secuenciación de Segunda Generación (SGS, Second-Generation Sequencing)

La búsqueda de nuevas tecnologías que permitiesen conseguir un mayor rendimiento con una minimización del coste y el tiempo, han dado lugar a las tecnologías SGS (NGS), que se conocen también como "secuenciación masiva paralela" ya que pueden secuenciar múltiples cadenas de ADN al mismo tiempo. Las más utilizadas son Roche/454 (Microsynth AG, Balgach, Suiza ), primera disponible como producto comercial en 2005, e Illumina-Solexa (EEUU; comercializada por Solexa en 2006 y adquirida por Illumina en 2007; BGI, Hong Kong). Ambas serán comentadas a continuación.

Estas tecnologías presentan una serie de etapas comunes que son en términos generales, la preparación del ADN molde, la secuenciación y detección de la señal, y el análisis de la información obtenida. La combinación única de protocolos específicos es lo que distingue a una tecnología de otra y determina el tipo de datos producidos a partir de cada una. Sin embargo, todas incluyen una serie de pasos comunes: 1) preparación del ADN molde, 2) secuenciación y 3) análisis bioinformático de las secuencias, el cual se verá más adelante.

**1) Preparación del ADN molde**

En ambas tecnologías, el ADN molde consiste en una colección de los fragmentos que se quieren secuenciar, denominada librería. Para la obtención de ésta, se fragmenta el ADN genómico al azar, dando lugar a segmentos de pequeño tamaño aunque variable dependiendo de la plataforma SGS. A continuación, se ligan adaptadores cortos en sus extremos, los cuales permiten por un lado la unión de las moléculas a una superficie sólida, y por otro lado el apareamiento de los iniciadores (cebadores o primers) que serán utilizados en la amplificación del ADN molde y su posterior secuenciación de la librería

(Ansorge, 2009). En los métodos SGS, la etapa de amplificación es necesaria para obtener suficiente intensidad de la señal que permitirá la detección fiable de las bases incorporadas.

Los métodos que se utilizan en la amplificación del ADN molde son la PCR en emulsión (emPCR) y la amplificación en fase sólida.

1.1) **La PCR en emulsión** (emPCR), utilizada por Roche 454, consiste en crear una librería de fragmentos a los que se ligan en sus extremos adaptadores que contienen sitios de reconocimiento para iniciadores universales. Después de la ligación, el ADN se desnaturaliza para que se separe en hebras sencillas y es capturado sobre perlas de acrilamida de 28 µm, bajo unas condiciones que favorezcan la unión de un fragmento de ADN por perla (Figura 12). La mezcla de reacción consiste en una emulsión de agua-aceite que es creada para encapsular las perlas-fragmento de ADN en gotitas acuosas individuales. La amplificación por PCR se desarrolla dentro de estas pequeñas gotas para conseguir perlas que contengan varios miles de copias del mismo fragmento. Después de la etapa de amplificación, las perlas pueden ser depositadas en pocillos individuales de placas "PicoTiter" en las que puede llevarse a cabo la siguiente etapa de secuenciación (Metzker, 2010; Mardis, 2007).

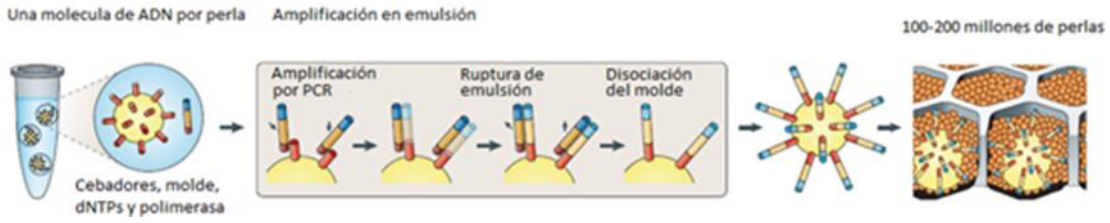


Figura 12. Estrategia de inmovilización del ADN molde en la tecnología Roche 454 (tomado de Metzker, 2010).

1.2) La **PCR en fase sólida** es utilizada por Illumina. En este caso, los fragmentos de ADN molde se ligan por ambos extremos a adaptadores y, después de un paso de desnaturalización, son inmovilizados por un extremo sobre un soporte sólido. Esto sucede, ya que la superficie del soporte esta densamente cubierta con oligonucleótidos directos y reversos, unidos covalentemente, que son complementarios a los adaptadores específicos que se han ligado a los fragmentos de ADN de la librería. La hibridación de estas moléculas de ADN de cadena sencilla a los oligonucleótidos de la superficie del soporte, se produce por calentamiento seguido de una etapa de enfriamiento. A

continuación, cada fragmento de cadena sencilla, inmovilizado por uno de sus extremos a la superficie, crea una estructura de "puente" por hibridación de su extremo libre con el adaptador complementario localizado también en la superficie del soporte (placa de PicoTiter).

Posteriormente tiene lugar la amplificación, en la que los adaptadores de la superficie actúan como cebadores universales para la amplificación por PCR. Después de varios ciclos de PCR se obtienen grupos aleatorios (clusters) de alrededor de 1000 copias de fragmentos de ADN de cadena sencilla sobre la superficie sólida de la placa (Figura 13) (Ansorge, 2009, Mardis, 2007).

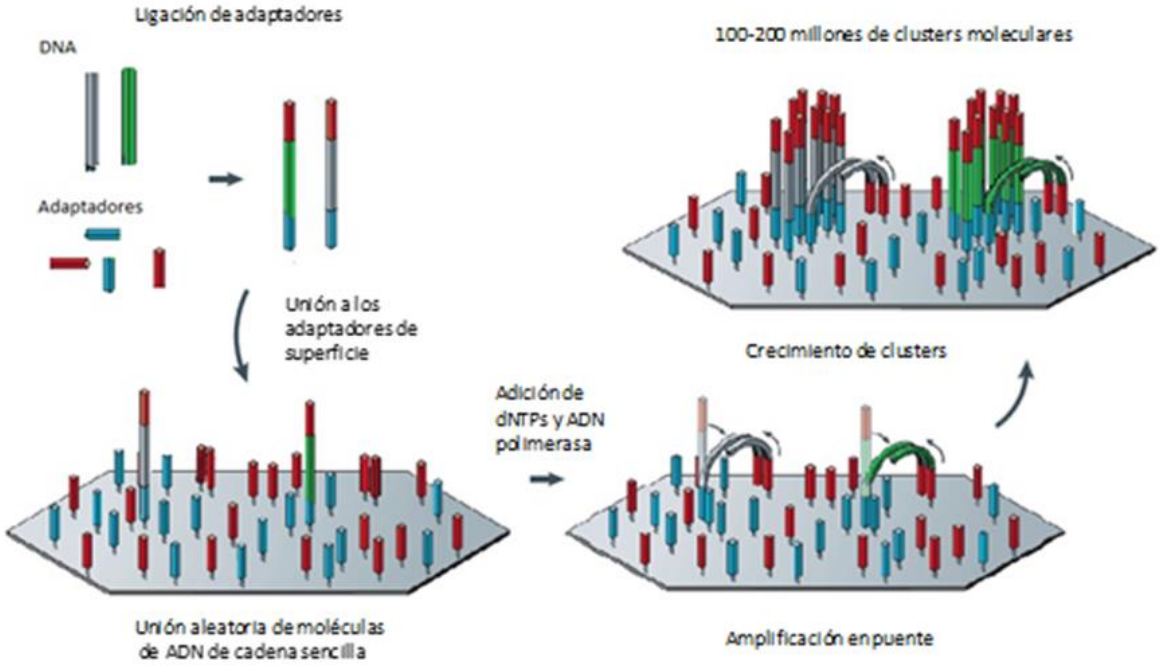


Figura 13. Estrategia de inmovilización del ADN molde en la tecnología Illumina/Solexa PCR en fase sólida (tomada de Metzker 2010).

## 2) Secuenciación

Las tecnologías SGS se caracterizan por las diversas implementaciones de secuenciación masiva paralela en arrays o placas. La tecnología 454 utiliza los Secuenciadores genómicos 454 de Roche y la tecnología Solexa los analizadores de Genomas Illumina.

La **tecnología 454/Roche** utiliza los Secuenciadores Genómicos 454 y permite lecturas de 400-500 pares de bases (pb). Se fundamenta en el principio de la pirosecuenciación, un método bioluminiscente no electroforético, basado en la liberación del pirofosfato inórganico (PPi) que se produce cuando la ADN polimerasa incorpora un nuevo nucleótido a la cadena de ADN en proceso de síntesis. El PPi es convertido de manera proporcional en luz visible, al ser usado como sustrato en una cascada de reacciones enzimáticas (Shendure y Ji, 2008). En esta cascada participan cuatro enzimas: la ADN polimerasa I (fragmento Klenow), la ATP sulfurilasa y la luciferasa interviniendo, además, la aspirasa. Durante la pirosecuenciación, el templado se hibrida con iniciador incubándose junto con los enzimas DNA polimerasa, ATP sulfurilasa, luciferasa y aspirasa, además de los sustratos adenosina-5'-fosfosulfato (APS) y D- luciferina. A continuación, la DNA polimerasa cataliza la incorporación de un dNTP complementario tras la adición de uno de los cuatro dNTPs, liberándose como resultado de la incorporación PPi. La cantidad de PPi liberado es proporcional al número de dNTPs incorporados. La ATP sulfurilasa convierte en presencia de APS el PPi formado en ATP (Adenosín trifosfato), el cual permite la conversión de D-luciferina en oxiluciferina por acción de la enzima luciferasa. En esta reacción, se produce la liberación de luz visible en una cantidad que es proporcional a la cantidad de ATP presente, pero además cada señal luminosa también es proporcional al número de dNTPs incorporados y a la cantidad de ADN molde. Durante este proceso, es necesaria la eliminación de los dNTPs no incorporados y de los restos de ATP por acción del enzima aspirasa. Finalmente, se adiciona un nuevo dNTP iniciando el siguiente ciclo y así sucesivamente.

Tras la elaboración de las librerías, o sea de la mezcla de fragmentos cortos flanqueados por adaptadores y capturados en la superficie de las perlas (Figura 14), se procede a romper la emulsión y las perlas se tratan con agentes desnaturalizantes para eliminar las hebras que no se han unido. A continuación, cada perla individual, que contiene múltiples copias del mismo amplicón, es depositada en un pocillo de la placa "PicoTiter" al que se añade, además, la ADN polimerasa, el cebador y los sustratos para las otros enzimas que participarán en la reacción: la adenosina-5'-fosfosulfato (APS; sustrato de la sulfurilasa) y D-luciferina (sustrato de la luciferasa).

El ADN molde se unirá al cebador cuyo extremo 3’ será utilizado por la ADN polimerasa para iniciar la síntesis de ADN. Los cuatro nucleótidos se añadirán de uno en uno, de una manera cíclica y de forma iterativa. Así, al añadir un dNTP complementario al molde, se extiende la cadena en crecimiento y después la síntesis se detiene hasta la adicción del siguiente. En cada pocillo se colocan también perlas de menor tamaño portadoras de la ATP sulfurilasa y de la luciferasa. El recubrimiento con estas últimas asegura que las perlas, con el ADN molde adherido, permanezcan en los pocillos durante la secuenciación.

Finalmente, una vez colocadas las perlas en la placa PicoTiter se coloca ésta en el secuenciador 454. Este incluye tres componentes principales, la cámara de flujo, donde se sitúa la placa PicoTiter, un subsistema de fluidos (fluidic subsystem), que hace pasar los reactivos de secuenciación, tampones y nucleótidos a través de los pocillos de la placa PicoTiter y un sensor (Margulies *et al*., 2006). El flujo secuencial de nucleótidos en un orden predeterminado permite la secuenciación en paralelo de los cientos de miles de perlas, cada una con millones de copias del mismo fragmento de ADN molde. Si un nucleótido complementario al molde llega al pocillo, el enzima ADN polimerasa extenderá la cadena añadiendo dicho nucleótido. La incorporación de uno o más nucleótidos resulta en una serie de reacciones en cascada que generan una señal luminosa. La señal luminosa es registrada por el sensor de una cámara CCD (Charge-Coupled Device o dispositivo de carga acoplada). Dicha señal será proporcional al número de nucleótidos incorporados. Finalmente, un ordenador controla al secuenciador y muestra el orden e intensidad de los picos de luz captada por la cámara en forma de flujograma (Flowgram). Los picos revelan la secuencia de ADN generada.

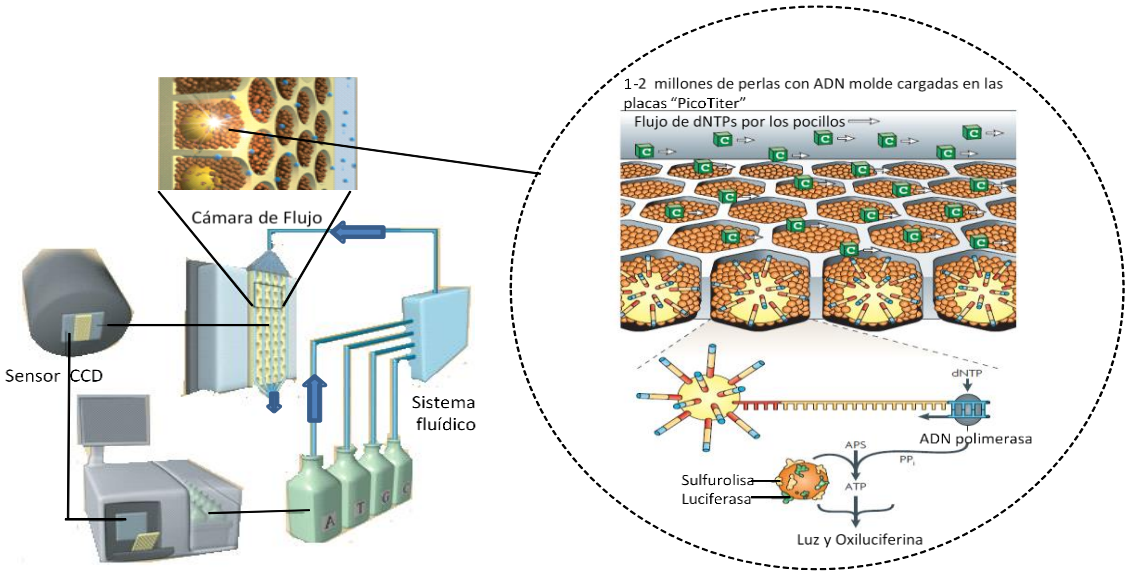


Figura 14. Pirosecuenciación utilizando la plataforma 454/Roche (tomada de Metzker 2010).

La tecnología de **secuenciación Illumina/Solexa** se basa en la Terminación Cíclica Reversible (CRT, Cyclic Reversible Termination), permitiendo lecturas de entre 35-150 pb. Durante el proceso de síntesis, la incorporación en la cadena de ADN del nucleótido terminador, así como su posición sobre la superficie de soporte, se detecta y se identifica a través de su marcador fluorescente por la cámara CCD (Figura 15). La terminación de la síntesis del ADN después de la adición de un único nucleótido es una característica importante de la secuenciación CRT. Tras la incorporación se lavan los terminadores restantes, y se capta la fluorescencia para determinar la identidad del nucleótido incorporado. A continuación, se eliminan tanto el grupo terminador responsable de la interrupción de la síntesis como el marcador fluorescente y se realiza un nuevo lavado antes de pasar al siguiente ciclo de síntesis donde se repite el proceso (Ansorge, 2009).

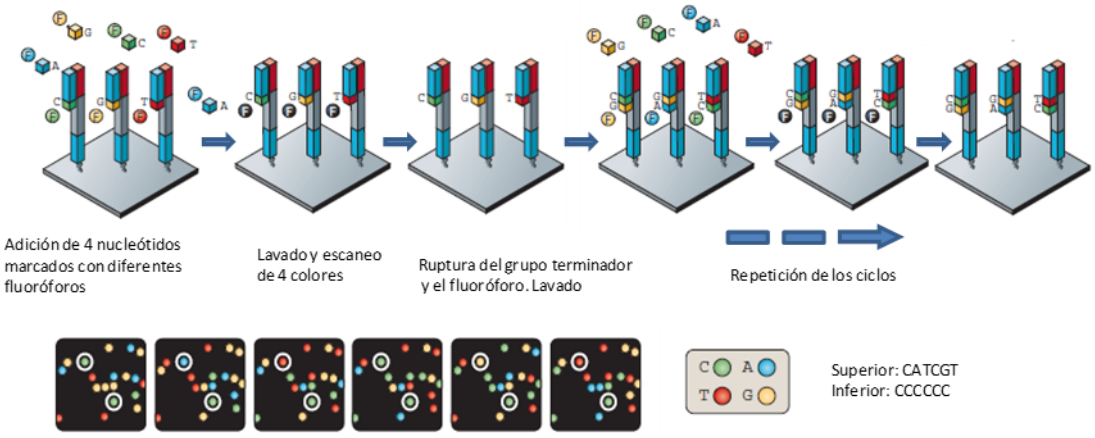


Figura 15. Método de la Terminación Cíclica Reversible (CRT) (tomado de Metzker, 2010).

### 2.5.3. Métodos de secuenciación de Tercera generación (TGS, Third-Generation Sequencing)

Las plataformas de secuenciación de tercera generación se caracterizan fundamentalmente por dos aspectos:

1) No requieren la etapa de amplificación previa a la secuenciación. Esto disminuye el tiempo de preparación de la muestra de ADN y elimina los posibles errores asociados a la amplificación por PCR.

2) La señal, bien sea fluorescente (PacBio), química (Ion Torrent) o eléctrica (tecnologías Nanopore) es detectada en tiempo real (Zhang *et al*., 2011). De todas ellas, sólo se comentará la plataforma PacBio, ya que fue la empleada en el presente trabajo.

La **plataforma PacBio** es un sistema de secuenciación de moléculas individuales (Single-Molecule Real-Time: SMRT), desarrollado por Pacific Biosciences (USA), que utiliza la tecnología Zero Mode Waveguide (ZMV) para observar la polimerización a tiempo real sin previa amplificación del ADN molde (Eid *et al*., 2009). Es la única plataforma que permite lecturas largas a partir de templados individuales, de unos 7- 10 Kb en el año 2014. Para llevar a cabo la secuenciación, el ADN es fragmentado al azar y, a continuación reparado en los extremos. Después, se añade adenina en el extremo 3' del ADN genómico fragmentado, lo que facilita la ligación posterior de un adaptador que es un oligonucleótido de ADN de cadena sencilla, que forma una estructura de horquilla intramolecular. Los fragmentos de ADN molde, son estructuralmente una molécula lineal, pero los adaptadores crean una molécula topológicamente circular.

La célula o chip SMRT donde se llevará a cabo la secuenciación, alberga una disposición configurada de zero-mode Waveguides (ZMWs) que son orificios circulares de 70 nm de diámetro y 100 nm de profundidad que están nanofabricados sobre una superficie de vidrio (Figura 16A). Dentro de cada celdilla se crea un volumen de observación lo suficientemente pequeño (20 zeptolitros) como para detectar la incorporación de un único nucleótido. La célula SMRT se prepara para la inmovilización de la polimerasa mediante el recubrimiento de la superficie con estreptavidina. La preparación de la reacción de secuenciación requiere la incubación de la ADN polimerasa biotinilada con moléculas del ADN molde con su cebador.

Los productos acoplados son entonces inmovilizados a la célula SMRT utilizando una interacción biotina-estreptavidina. Cuando comienza la reacción de secuenciación, la polimerasa anclada incorpora a la cadena de ADN creciente, nucleótidos marcados cada uno de ellos con un fluoróforo diferente. Durante dicho evento, el nucleótido fluorescente se localiza en el sitio activo de la polimerasa y próximo a la superficie de vidrio del ZMW. Por debajo de éste, una cámara de alta resolución registra la fluorescencia del nucleótido que se incorpora.

Durante la reacción de incorporación, el fluoróforo es liberado del nucleótido y esa disociación disminuye la señal fluorescente. Así, mientras que la polimerasa sintetiza una copia de la cadena molde, los sucesivos eventos de incorporación de nucleótidos son registrados (Figura 16B) (Eid *et al*., 2009). La clave del proceso es que dentro del orificio se crea un volumen de observación lo suficientemente pequeño como para detectar la incorporación de un único nucleótido. La luz emitida por un láser, ilumina solamente la parte inferior del pocillo, de forma que cuando la ADN polimerasa incorpora un determinado nucleótido este será excitado emitiendo una determinada señal fluorescente que será recogida por un detector, permitiendo su identificación. Sin embargo la luz no alcanza a los dNTPs no incorporados, de modo que éstos, al no ser excitados, no emitirán fluorescencia.

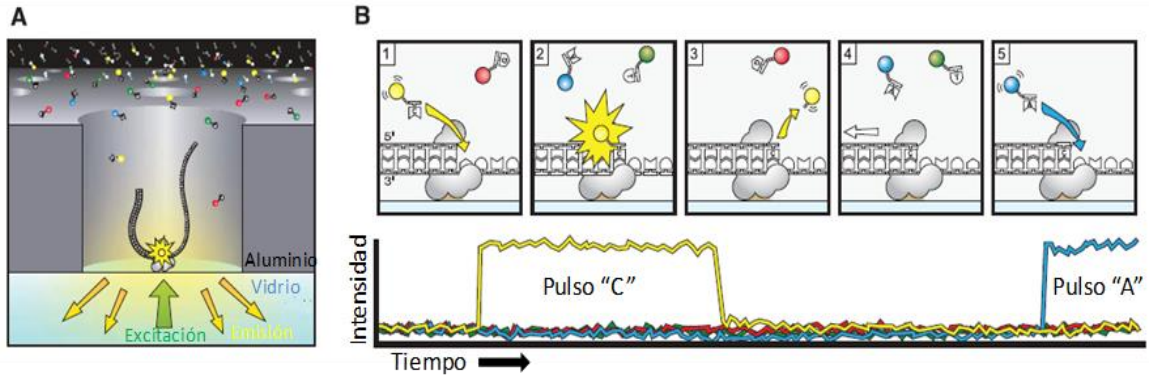


Figura 16. Plataforma PacBio. A) Vista lateral de una nano-estructura ZMW; B) Esquema de la incorporación de nucleótidos marcados con fluoróforos (tomada de Eid *et al*., 2009).

### 2.5.4. Análisis de las secuencias

Las estrategias empleadas en la secuenciación de fragmentos relativamente pequeños mediante las técnicas de segunda y tercera generación constituyen un reto para el análisis. Estos métodos generan grandes cantidades de información, hasta varios millones de lecturas, por lo que se ha hecho necesario el desarrollo tanto de sistemas de almacenamiento masivo de datos como de herramientas informáticas para su análisis. Esto condujo al desarrollo de una nueva disciplina de análisis informático que por ser aplicada al estudio de secuencias biológicas se ha llamado bioinformática.

Además de los retos antes mencionados, el análisis de secuencias se enfrenta al problema de la existencia de diferencias entre las distintas plataformas (formato de datos, longitud de lectura, etc.), lo cual incide en la diversificación informática y en la dificultad de utilizar sistemas únicos automatizados para el análisis de los mismos. En la actualidad, existen una gran diversidad de herramientas bioinformáticas que son de gran utilidad para el alineamiento y ensamblaje de genomas, así como para realizar una predicción y anotación funcional de la información contenida en estos genomas. Debido a esta diversidad, es sumamente importante su manejo y correcta interpretación, teniendo que tener en cuenta en cuenta el origen del DNA, tanto la especie a estudiar como la tecnología de secuenciación, antes de aplicarlas.

# Material y métodos

## 3.1. Secuencias del plásmido

Como se comentó anteriormente, el punto de partida de este trabajo fueron las secuencias del plásmido pUO-STmRV1, característico del clon monofásico español de *S. enterica*, obtenidas por cuatro métodos diferentes: Secuenciación cíclica de Sanger (Macrogen Europa), pirosecuenciación Roche 454 (Applied Sciences, Basilea), CRT Illumina (BGI, Hong Kong) y SMRT PacBio (Pacific Biosystems, USA). En los dos últimos casos, la secuencia del plásmido se consiguió como parte de la secuencia del genoma completo de la bacteria portadora (LSP 398/97).

Cabe recordar que la secuenciación del plásmido mediante Sanger y Roche 454 y su posterior ensamblaje y anotación fueron realizados en el contexto de la Tesis Doctoral de García (2013). Además, el ensamblaje preliminar de las secuencias obtenidas por Illumina y PacBio fue llevado a cabo por la empresa bioinformática Era7 y por Pacific Biosystems, respectivamente. En la Tabla 1 se indica el número de contigs obtenido en cada caso.

Con objeto de mejorar los resultados anteriores y comparar los ensamblajes automatizados con aquellos optimizados para la secuencia en cuestión, en el presente TFM se utilizaron distintas herramientas bioinformáticas, encaminadas al ensamblaje y anotación de las secuencias disponibles.

| **Método de secuenciación** | **Número de contigs** | **Secuencia contenida** |
|---|---|---|
| Sanger/Roche 454 | 3 | Plásmido |
| Roche 454 | 30 | Plásmido |
| CRT Illumina-Era7 | 185 | Genoma |
| SMRT PacBio | 2 | Genoma |

Tabla 1. Métodos de secuenciación empleados para conseguir la secuencia del plásmido pUO-STmRV1 y número de contigs conseguidos con cada uno de ellos.

## 3.2. Ensamblaje de *novo* de las lecturas obtenidas por Illumina

La secuenciación Illumina generó lecturas paired-end de 90 pb a partir de una librería con de fragmentos de 500 pb. La cantidad total de datos filtrados aportada por BGI fue de 1

Gb. Previamente al ensamblaje, se llevó a cabo un análisis de la calidad de la secuencia mediante el programa informático fastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

El programa fastQC permite analizar la calidad de la secuencia generada por plataformas de NGS. Permite analizar la calidad de las secuencias, adjudicándole un índice de calidad (score) tanto a las secuencias individuales como a las bases de cada secuencia en función de su posición. Esto permite decidir si se utiliza la secuencia completa o si por ejemplo se elimina los extremos si la calidad en estos disminuye. Es especialmente útil en el caso de secuencias superiores a los 50 nucleótidos. Por otro lado también realiza un análisis de secuencias altamente representadas que suelen estar generadas por errores en el proceso de secuenciación. FastQC tiene como objetivo principal proporcionar un análisis global de la calidad de las secuencias antes de utilizarlas con otros programas.

El ensamblaje de *novo* de las secuencias obtenidas se realizó con los paquetes de software Velvet (Zerbino y Birney, 2008) y VelvetOptimiser 2.2.5 (Zerbino, 2010) (www.ebi.ac.uk/~zerbino/velvet_latest.tgz). Velvet es un ensamblador de *novo*, que fusiona las lecturas de las secuencias que se solapan, en secuencias contiguas de mayor tamaño denominadas contigs, sin necesidad de ningún genoma de referencia (Edwards y Holt, 2013). Este programa requiere que el conjunto de lecturas individuales generadas durante la secuenciación masiva se encuentren en un solo archivo, en formato FASTA o FASTAQ. En nuestro caso, en concreto, se utilizaron los archivos de las secuencias en formato FASTQ. Por otro lado, el conjunto de datos de las lecturas de extremos emparejados a menudo se proporcionan como archivos separados, por lo que hay que tenerlo en cuanta antes de realizar el ensamblado.

El funcionamiento de Velvet se basa en la comparación, mediante el uso de algoritmos, de las secuencias obtenidas, creando gráficos o tablas de relaciones ponderadas para determinar cuáles solapan con otras y en qué medida lo hacen (Zerbino y Birney, 2008). El proceso ocurre típicamente en dos etapas: “hashing” y construcción de un gráfico Bruijin, que se consiguen gracias a los ejecutables Velvetg y Velveth respectivamente (Figura 17). Por un lado, Velvetg lee los archivos con las secuencias creando una librería

de lecturas fragmentadas a partir una determinada longitud (k-mer). El k-mer es un parámetro preestablecido por el usuario, que define de forma exacta los alineamientos locales entre las lecturas. Posteriormente, Velvetg lee estos alineamientos construyendo el gráfico de Bruijn a partir de ellos y corrigiendo errores. Finalmente lo simplifica y resuelve las repeticiones basándose en los parámetros proporcionados por el usuario (Zerbino, 2010). La variable k-mer se encuentra limitada por la longitud de las lecturas a analizar, ya que debe permitir un pequeño solapamiento. Por lo general suele usarse un valor de k = de 21 a 25 pb. A continuación, para cada k-mer observado en el conjunto de lecturas, se construye una tabla *hash* o matriz asociativa, en la cual se identifica un dato o parámetro y la posición del mismo, registrando el ID de la primera lectura encontrada que contiene ese k-mer y su posición dentro de la lectura. Cada k-mer se graba simultáneamente con su secuencia inversa complementaria, lo cual permite reescribir cada lectura como un conjunto de k-mers originales junto con solapamientos con las lecturas previamente fragmentadas. Esta nueva representación de las lecturas de la secuencia se conoce como hoja de ruta o ROADMAP.

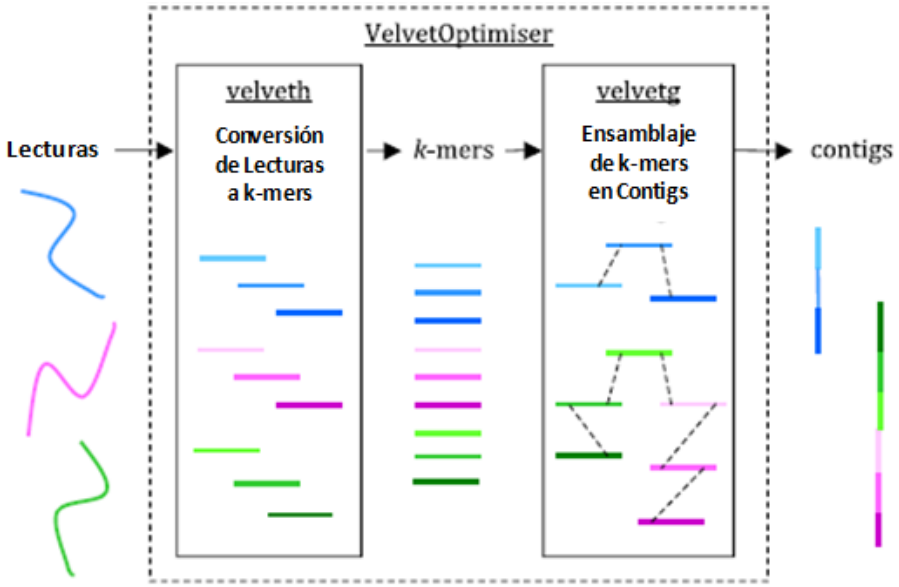


Figura 17. Lecturas ensambladas en contigs usando Velvet y VelvetOptimiser en dos etapas: En la primera, Velveth convierte las lecturas en k-mers usando tablas hash. En la segunda, Velvetg ensambla los k-mers solapantes en contigs mediante el gráfico de Bruijn (Edwards y Holt, 2013).

Por otro lado, el programa crea una segunda base de datos con la información opuesta, registrándose para cada lectura cuáles de sus k-mers originales han solapado con lecturas posteriores. El conjunto ordenado de los k-mers originales de esta lectura se corta cada

vez que comienza o termina el solapamiento con otra lectura, creándose un nodo de cada secuencia ininterrumpida de k-mers originales.
Finalmente, estas lecturas son rastreadas a través de un gráfico usando las hojas de ruta. Al final se obtiene un gráfico en el que cada nodo representa una serie de k-mers solapantes. La secuencia de estos nucleótidos corresponde a la secuencia del nodo. El nodo gemelo, unido directamente por debajo o por encima, representa la serie inversa del conjunto de k-mers complementarios a los k-mers del nodo inicial. Es importante tener en cuenta que las secuencias asociadas a un nodo y su gemelo no necesitan ser inversas complementarias una de la otra. La secuencia ensamblada se obtiene recorriendo el gráfico desde el extremo inicial hasta el final, concatenando la primera base del k-mer almacenado en cada uno de los nodos recorridos, excepto el k-mer del nodo final que se une por completo al resultado.

A pesar de que el Velvet puede funcionar con lecturas generadas a partir de un solo extremo de los fragmentos de la librería, el uso de lecturas de extremos emparejados (paired-end) es recomendable para obtener contigs largos, especialmente en secuencias con regiones repetitivas. Velvet proporciona lecturas diferentes en función de la longitud, emparejamiento y su librería. Esta primera distinción la realiza entre lecturas largas y cortas.

Cuando un ensamblador utiliza gráficos de Bruijn, como es el caso de Velvet, deben tenerse en cuenta una serie de variables con el fin de producir los contigs más óptimos. Las distintas plataformas de secuenciación producen fragmentos de diferente longitud y calidad, por lo que determinados k-mers serán mejores para unas u otras, lo cual pone de manifiesto la importancia de una buena selección de la longitud de los k-mers. Los k-mers más pequeños aumentan la posibilidad de encontrar solapamiento entre dos lecturas debido a errores de secuenciación e incrementan también el número de repeticiones ambiguas en el gráfico. Por ello, es necesario buscar un equilibrio entre la sensibilidad ofrecida por los k-mer de menor tamaño y la especificidad asociada a los k-mer de mayor tamaño. Otras variables a considerar cuando se utiliza Velvet incluyen la cobertura (esto es, el número de veces que se espera que cada base esté representada en cada una de las lecturas individuales), la longitud del tamaño de los fragmentos de las librerías de extremos emparejados y el valor de corte de la cobertura mínima (profundidad de lectura). Los programas de ensamblaje que utilizan gráficos de Bruijn no distinguen entre un error

y una variante real si existen evidencias de error, lo cual sucede con niveles más altos de cobertura. En este caso, se puede seleccionar un subconjunto de las lecturas y utilizarse para el ensamblaje (Edwards y Holt, 2013).

El ajuste de parámetros necesarios para el ensamble de *novo* con Velvet puede automatizarse de forma eficaz usando VelvetOptimiser (Zerbino, 2010). Este programa está diseñado para evaluar varios de los parámetros utilizados en Velvet de forma que minimiza el tiempo de computación realizando una evaluación de cada una de las opciones a optimizar, seleccionando aquellas que generen los mejores valores en función de objetivo fijado por el usuario, generalmente minimizar el número de contigs maximizando el tamaño de los mismos. De hecho, VelvetOptimiser permite conseguir rápidamente largas secuencias continuas, a partir de un conjunto de datos de lectura cortas obtenidas mediante técnicas de secuenciación de segunda generación, siendo el más eficaz en el caso de Illumina (Zerbino, 2010). El parámetro optimizado con VelvetOptimizer en el presente trabajo fue la longitud de los k-mer (35-117). También se indicaron el formato de los archivos (FASTQ), el tamaño de las secuencias (Short paired), y que las secuencias a ensamblar se encuentran en archivos separados (separate), pero corresponden a fragmentos de extremos emparejados. Todos estos parámetros vienen integrados en la orden introducida en el VelvetOptimiser, la cual se indica a continuación:

**perl VelvetOptimiser.pl -s 35 -e 117 -f '-shortPaired -fastq.gz -separate archivo.fastq.gz archivo.fastq.gz**

## 3.3. Comparación de las secuencias obtenidas por diferentes técnicas de secuenciación

Las secuencias obtenidas mediante los métodos de Sanger, 454/Roche e Ilumina fueron comparadas frente a la secuencia obtenida con la plataforma PacBio, usada como referencia, utilizando el servidor online CONTIGuator (http://contiguator.sourceforge.net). Se trata de una herramienta de software para mapear contigs en comparación con un genoma de referencia, lo cual permite visualizar un mapa de contigs indicando la pérdida y/o ganancia de elementos genéticos, así como finalizar genomas cuya secuencia se encuentra en más de un contig. Utiliza una secuencia de comandos que combinan las rutinas de una de las herramientas más utilizadas, Abacas,

refinando los resultados mediante la herramienta de comparación Artemis (ACT, Artemis Comparison Toon). El enfoque del programa permite al usuario una visualización clara de la estructura genómica mediante un análisis comparativo.
Este programa requiere un archivo FASTA que contenga todos los contigs, y otro archivo FASTA con la secuencia de referencia. La primera etapa del análisis es una comparación de alta similitud (megablast run) que proporciona el perfilado inicial de los contigs. Este paso es útil para poner de manifiesto los contigs que son homólogos o divergentes a la secuencia de referencia. Este alineamiento inicial se utiliza después como entrada para el script de perl Abacas, uno para cada replicón de referencia, que lleva a cabo una comparación con el programa MUMmer permitiendo crear una molécula ensamblada, pseudocontig, basada en el genoma de referencia (Kurtz *et al*., 2004). Esta molécula es utilizada y corregida por el CONTIGuator para crear el mapa de comparación.

Por tanto, los contigs son mapeados respecto a un genoma de referencia utilizando una combinación del Blast y MUMmer generando un mapa visible con ACT. En este mapa el genoma de referencia se localiza en la parte superior y el pseudocontig en la parte inferior (Figura 18). En este mapa se indican los alineamientos existentes entre la secuencia de referencia y los contigs (regiones rojas), así como las regiones no homólogas entre ambos (regiones blancas). También se indica la orientación en la que se alinean los contigs con la secuencia, los contigs con alineación problemáticos (color rojo) los resultados de la comparación de las secuencias realizado mediante tblastn.

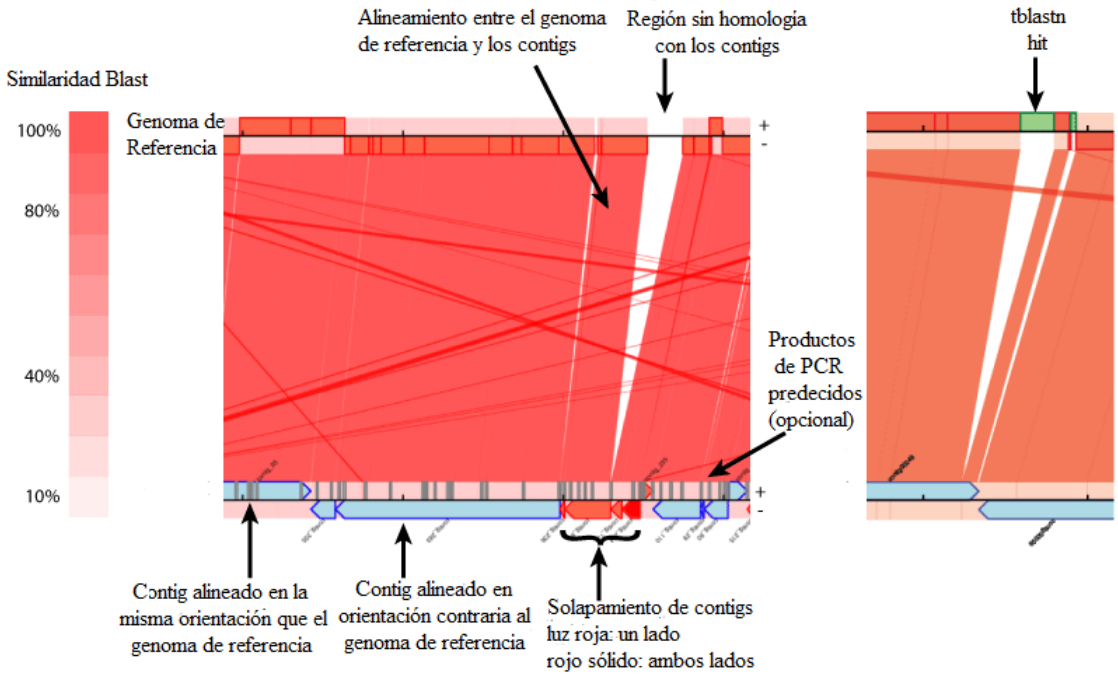


Figura 18. Representación de un mapa ACT.

## 3.4. Anotación automática de la secuencia de pUO-STmRV1

La anotación de las secuencias ensambladas consiste en primer lugar en identificar los posibles genes en las secuencias de los contigs obtenidos. Posteriormente se realiza una comparación de los genes identificados, tanto a nivel de secuencia nucleotídica como aminoacídica, con otras secuencias, generalmente de bases de datos públicas, para establecer si existe suficiente homología entre la nueva secuencia y la de referencia (contenida en la base de datos), con el fin de identificar y asignar una función a los genes de la secuencia en estudio. Este es un proceso complejo que requiere la utilización de métodos bioinformáticos basados en diferentes aproximaciones y el acceso a bases de datos, lo que ha provocado el desarrollo de herramientas bioinformáticas que han de ser ejecutadas para poder obtener la máxima información de una secuencia genómica.

Se realizó una anotación automática provisional de la secuencia del plásmido, obtenida mediante el método de secuenciación de PacBio, utilizando la plataforma online Rapid Annotations using Subsystems Technology versión 2.0 (RAST) (http://rast.nmpdr.org/rast.cgi). El servidor RAST fue creado en el año 2007 y se publicó una descripción de su tecnología en el año 2008. Esta plataforma es un servicio

automatizado diseñado para identificar y anotar rápidamente los genes de un genoma procariota completo, tanto si está en un único contig como si está en varios. Para ello, utiliza una biblioteca de subsistemas basados en familias de proteínas que garantizan un alto grado de consistencia en la asignación de identidad (FIGfams). Además, realiza un análisis de los genes y subsistemas del genoma, mediante técnicas de análisis comparativas y de clasificación en grupos funcionales.

El servidor RAST identifica de forma automática los genes y realiza una anotación funcional a través de dos estrategias: identificaciones basadas en los subsistemas, es decir, se basa en el reconocimiento de grupos funcionales en los genes presentes en la secuencia a analizar. Además, complementa esta información con una segunda estrategia basada en comparación e integración de sistemas. Este hecho, permite por ejemplo, la obtención de una posible reconstrucción metabólica, por lo que constituye un buen punto de partida como inicio para un trabajo de anotación más amplio.

## 3.5. Anotación optimizada de la secuencia del plásmido

El software informático Clone Manager 9 (http://www.scied.com/pr_cmbas.htm) permite, entre otras muchas aplicaciones, comparar dos secuencias de nucleótidos gracias al comando “Align” y al subcomando “Compare two sequences” Este procedimiento fue utilizado para comparar la secuencia PacBio con la secuencia de cada uno de los tres contigs obtenidos por la combinación de Sanger/Roche 454 (García, Tesis Doctoral, 2013). Dichos contigs habían sido previamente anotados de forma manual (García, Tesis Doctoral, 2013). Una vez establecidas las regiones de homología, los tres contigs fueron combinados en el orden correcto, mediante el comando “Clone” y el subcomando “Join sequences”, que permite mantener las anotaciones de las secuencias unidas.

# Resultados y discusión

## 4.1. Secuencia obtenida con los diferentes métodos de secuenciación

El primer objetivo de este trabajo fue evaluar la cantidad y calidad de la secuencia del plásmido pUO-STmRV1 obtenida usando métodos de secuenciación de primera, (Sanger) segunda (Roche 454 y Illumina) y tercera (PacBio) generación. Como se comentó anteriormente, en la Tabla 2 se puede observar un resumen de los resultados obtenidos con cada uno de ellos, junto con el coste y el tiempo invertido. Los datos correspondientes a Sanger y Roche 454 fueron combinados durante la Tesis Doctoral de García (2013).

| | **Sanger/Roche 454** | **Roche 454** | **Illumina** | **PacBio** |
|---|---|---|---|---|
| **Secuencia (pb)** | 197649 | 158917 | 167415 | 202354* |
| **% del total de la secuencia** | 97,7 | 78,5 | 82,7 | 100% |
| **Número de contigs** | 3 | 30 | 185* | 2* |
| **Coste** | Incalculable | 2000 E | 470 $* | 2520 $* |
| **Tiempo** | 3 años | 1 mes | 1 mes* | 2 meses* |

Tabla 2. Características de las secuencias de pUO-STmRV1 obtenidas con los diferentes métodos empleados.

*Datos correspondientes al genoma completo de la cepa LSP 389/97 del clon monofásico español (cromosoma y plásmido).

### 4.1.1. Secuencia obtenida por el método Sanger

Para obtener la secuencia del plásmido por el método Sanger se requiere la utilización de diversas técnicas de Biología Molecular, incluyendo extracción del plásmido, clonación de fragmentos del mismo, diseño de iniciadores, amplificación por PCR y secuenciación de los clones y fragmentos obtenidos. En el caso de un proyecto de secuenciación como el de este trabajo, un plásmido de bajo número de copia y de gran tamaño (200 Kb, aproximadamente), las dificultades aumentan. Por un lado, resulta muy difícil extraer la cantidad de plásmido requerida para obtener los clones que serán posteriormente secuenciados. Éstos son fragmentos del plásmido conseguidos por digestión con determinados enzimas de restricción e insertados en un vector de clonación. En segundo lugar, aunque alguno de los clones pudo seleccionarse gracias a la presencia de genes de resistencia, en la mayor parte de los casos no existió selección. De manera ideal, los fragmentos clonados deben solapar entre ellos y contener toda la

secuencia del plásmido. La secuenciación de cada uno de ellos se inicia utilizando cebadores que se unen al vector, siendo después necesario el diseñando nuevos cebadores en base a la secuencia que se va generando (DNA walking). Los huecos que quedan se rellenan mediante amplificación por PCR y secuenciación de los amplicones. De esta manera se van obteniendo segmentos contiguos del plásmido denominados contigs. En el caso del plásmido pUO-STmRV1, este trabajo se prolongó durante un periodo de tres años y las secuencias obtenidas fueron finalmente combinadas con las aportadas por Roche 454 (ver apartado 4.1.2). Sin embargo, cabe destacar aquí que, a pesar del gran esfuerzo realizado y del tiempo invertido, no se pudo resolver la secuencia total del plásmido. Por otro lado el coste es incalculable, ya que a las secuencias realizadas (cuyo número se desconoce) es necesario sumar el importe de los cebadores y de los reactivos de extracción del plásmido, clonación y amplificación por PCR.

### 4.1.2. Secuencia obtenida por el método Roche 454

Otra de las estrategias para conseguir la secuencia del plásmido fue mediante el uso de una tecnología de segunda generación, en concreto la pirosecuenciación de Roche 454. En este caso, la mayor dificultad estribó en la obtención de suficiente material de partida, esto es, de plásmido puro sin contaminación con DNA cromosómico y lo más intacto posible. El hecho de que el plásmido pUO-STmRV1 sea de gran tamaño y de bajo número de copia obligó a optimizar el método de extracción. La secuencia se obtuvo tras el envío del DNA plasmídico a la empresa Microsynth AG.

El ensamblaje de la secuencia obtenida por esta metodología se realiza directamente con un programa específico de la propia compañía Roche (GS De Novo Assembler version 2.5.3). Como se observa en la Tabla 2, se obtuvieron 158917 pb que representan un 78,5% de la secuencia del plásmido. El importe fue de 2000 euros y el tiempo necesario para obtener la secuencia de aproximadamente un mes. La secuencia quedó contenida en 30 contigs, que fueron combinados con los obtenidos por el método de Sanger. Esto permitió conseguir un total de 197649 pb distribuidas en 3 contigs, lo que supone el 97,7% de pUO-STmRV1 (Tabla 2). Existen varios motivos para no haber conseguido la secuencia completa. Además del tamaño, destaca la existencia de numerosas regiones repetidas, que pueden enmascararse en el análisis e impiden el ensamblaje. Por otro

lado, la propia estructura física del DNA puede hacer que haya regiones con mayores dificultades de acceso al método de secuenciación, que no se vean representadas en la secuencia final. A pesar de no disponer de datos del importe económico de la secuenciación por el método de Sanger, el uso de tecnologías de segunda generación reduce considerablemente el tiempo y el coste para obtener secuencia. La fiabilidad a nivel de nucleótido es mucho mayor en el caso de estas tecnologías, ya que reduce el número de errores y estos se ven reducidos por el hecho de que cada nucleótido se lee varias veces (el coverage medio fue de 40).

### 4.1.3. Secuencia obtenida por el método Illumina

Con esta metodología de segunda generación el número de secuencias obtenidas es muy alto, concretamente en este caso de 11,1 millones de lecturas realizados, con un tamaño de 90 nucleótidos por lectura. Además aportan información estructural ya que se secuencia una librería genómica de 500 pb por ambos extremos (secuenciación de extremos pareados). Esto hace que el coverage total sea muy alto (cercano a 200). El inconveniente es que se tiene que realizar la secuenciación completa del genoma de la cepa, tanto del cromosoma como del plásmido, lo que dificultó el análisis de este último. En la Tabla 2 se puede ver un resumen de los datos obtenidos. El número de contigs es muy alto porque incluye tanto los contigs correspondientes al cromosoma como los correspondientes al plásmido. En este caso no es posible realizar una comparación con los resultados obtenidos previamente. Si se puede evaluar el coste, considerablemente inferior al de los dos métodos anteriores (470 dólares USA; 270 por la construcción de la librería y 200 más por la secuenciación CRT), lo mismo que el tiempo. Cabe destacar además que, al no tener que realizar la extracción del DNA plasmídico, el proceso se simplifica mucho.

Illumnia es la técnica de elección cuando se quiere obtener una información básica sobre el genoma de la cepa a estudiar (Ansorge, 2009). Generalmente, como se puede ver en la Tabla 1, el principal problema es la obtención de un alto número de contigs lo que dificulta la organización de los mismos y la asignación a cromosoma o plásmido. Este alto número de contigs se debe a la realización de librerías de fragmentos de pequeño tamaño (500 pb) que no permiten la resolución y asignación de regiones repetidas con un tamaño superior a 500 pb. Entre ellas se incluyen las secuencias de

inserción, los transposones, las repeticiones génicas, los operones ribosomales, etc. Este hecho es especialmente grave en el caso de los plásmidos, estructuras génicas de alta variabilidad y movilidad que adquieren fácilmente secuencias de inserción y otros elementos móviles, como transposones, que se acumulan en la secuencia del plásmido de forma totalmente aleatoria.

### 4.1.4. Secuencia obtenida por el método PacBio

Esta metodología de tercera generación está recomendada para la finalización de genomas debido al gran tamaño de las lecturas individuales (hasta 9 Kb) (Eid *et al*., 2009). De esta forma, se elimina el principal inconveniente de las secuencias generadas por la tecnología Illumina, esto es, la imposibilidad de resolver secuencias repetidas de tamaño superior a 500 pb. Además, al igual que en el caso de Illumina, la secuenciación del DNA total y no sólo del plásmido, reduce considerablemente el trabajo, el tiempo y el coste de la extracción del DNA molde. Como se observa en la Tabla 2 se obtuvieron un total de 5109602 pb en dos contigs, correspondiendo el mayor de ellos de 4907248 pb al cromosoma de la cepa LSP 389/97 y el segundo de 202354 pb al plásmido pUO-STmRV1. De acuerdo con ello, el genoma completo de LSP 389/97 es de 5109602 pb. Cabe destacar que de todos los métodos de secuenciación evaluados en este TFM el que mejores resultados ha obtenido fue PacBio. De hecho, la secuencia PacBio fue utilizada como referencia para evaluar la calidad de las secuencias obtenidas por el resto de métodos, así como de los distintos ensamblajes realizados.

Como inconvenientes, el coste de esta tecnología (2520 dólares; 735 dólares por la construcción de la librería y 1785 dólares por la secuenciación) es mayor que el de las técnicas de segunda generación, pero los resultados obtenidos compensan claramente la diferencia y el tiempo requerido fue solo de un mes (Tabla 2).

## 4.2. Ensamblaje de *novo* de las secuencias obtenidas por Illumina

En este TFM se llevó a cabo el ensamblaje de las lecturas obtenidas mediante la tecnología Illumina y se comparó con el realizado por una empresa de bioinformática (Era7 Information Technologies, Granada, España), con un coste de 175 euros.

### 4.2.1. Ensamblaje realizado por una empresa comercial

En general las empresas de bioinformática ofrecen un servicio de ensamblaje automatizado mediante la aplicación de un software específico que permite la realización de varios ensamblajes con unos pocos cambios en parámetros optimizados para la mayor parte de los genomas. Esto hace que no consideren las peculiaridades que pueda tener un genoma particular, por ejemplo la existencia de repeticiones y estructuras organizativas complejas. Por otro lado, la optimización maximiza generalmente la longitud de los contigs finales a expensas de perder calidad en la información estructural. Se puede permitir, por ello, la unión de contigs en estructuras superiores, denominadas scaffolds, que podrían no ser reales. En el caso concreto de la secuencia de la cepa LSP 389/97, la empresa realizó el ensamblaje utilizando un software propio adaptado al programa Velvet y como parámetro principal de optimización maximizó la longitud de los contigs. De esta manera, el número de contigs obtenido para el genoma completo de la cepa fue de 185 (Tabla 1).

Para evaluar la calidad del ensamblaje, se utilizó el programa CONTIGuator para realizar un alineamiento de los contigs de Era7 con la secuencia del cromosoma de LSP 389/97 obtenida por PacBio, incluida como referencia. En la Figura 19A se puede ver el mapa ACT, con los contigs de Era7 en la parte inferior y la secuencia PacBio en la parte superior. En la figura se observan varias regiones cuyo ensamblaje por Era7 no coincide con el de PacBio. El cromosoma ensamblado por Era7 incluye 4832564 pb distribuidas en 36 contigs, cuyo orden es 17, 28, 71, 8, 131, 92, 156, 15, 42, 37, 16, 144, 9, 30, 5, 4, 99, 24, 172, 6, 129, 19, 114, 27, 18, 161, 167, 49, 74, 22, 35, 36, 50, 127, 118 y 89. El cromosoma de Era7 corresponde al 98,5% del de PacBio que se encuentra totalmente ensamblado.

### 4.2.2. Ensamblaje realizado mediante el programa VelvetOptimiser

Como paso previo al ensamblado de las secuencias se procedió a analizar la calidad de las mismas con el programa FastQC. La calidad de las secuencias fue muy buena como se observa en la Figura 20, donde la mayor parte de la secuencia tiene un score >20, valor que se considera óptimo. El score medio de las secuencias fue de 36,8, muy por

encima del valor óptimo (score=20). Además se comprobó que no existen Ns en ninguna de las posiciones de los 100000080 de secuencias analizadas. Por otro lado, no se encontraron secuencias con repeticiones, ni altamente repetidas, ni sobre representadas, constituyendo por tanto unas secuencias de alta calidad. Debido a este hecho no fue necesario realizar ningún filtrado por calidad ni por tamaño previo al ensamblado.

En este TFM se llevó a cabo un ensamblaje de las mismas secuencias generadas por Illumina mediante el ensamblador de novo Velvet y el programa VelvetOptimiser v.2.2.5. Se realizaron diversos ensayos de ensamblaje para optimizar los resultados. El criterio de optimización fue minimizar el N50, que por un lado trata de reducir el número de contigs, parámetro principal, maximizando la longitud de los contigs, parámetro secundario, tratando de obtener un equilibrio entre ambos. Además se ensayó un amplio rango de k-mers (de 21 a 117, con un paso de 2). El mejor ensamblaje obtenido respecto al número de contigs fue el correspondiente a un k-mer de 75, número superior al óptimo para este tipo de secuencias. Se obtuvieron un total de 153 contigs, número inferior al conseguido con el ensamblaje automático realizado por la empresa Era7 (185 contigs). El mapa ACT obtenido en el alineamiento de los 153 contigs con la secuencia PacBio del cromosoma de LSP 389/97 se muestra en la Figura 19B. De los 153 contigs solo 37 alinearon con la secuencia de cromosoma obtenida por PacBio. Estos contigs suponen un 98,3% de la secuencia con un total de 4827269 pb. Los contigs que alinearon y su orden fue 207, 25, 136, 18, 125, 95, 115, 20, 6, 32, 14, 161, 7, 176, 11, 9, 26, 23, 12, 190, 33, 121, 101, 4, 130, 79, 5, 53, 145, 199, 184, 39, 8, 54, 48, 126 y 37. Al comparar los ensamblajes realizados de forma automática y manual podemos ver que aunque el porcentaje de secuencia es similar (98,5 versus 98,3), la calidad es muy superior en el segundo caso, ya que la cantidad de DNA incorrectamente ensamblado es mucho menor. Los resultados obtenidos ponen de manifiesto que, pese a que los sistemas de automatización han supuesto un gran avance en el análisis de las secuencias, aún es necesario personal cualificado que permita una mayor optimización de los parámetros con el fin de obtener mejores resultados.

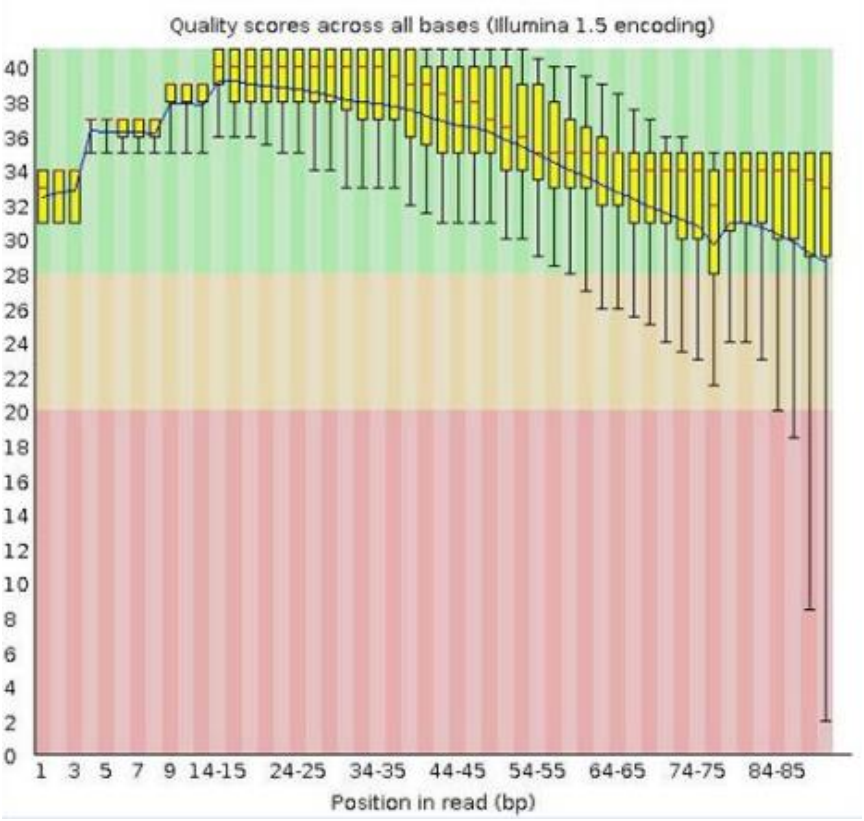


Figura 20. Gráfico de control de calidad de la secuencia obtenida por el método Illumina.

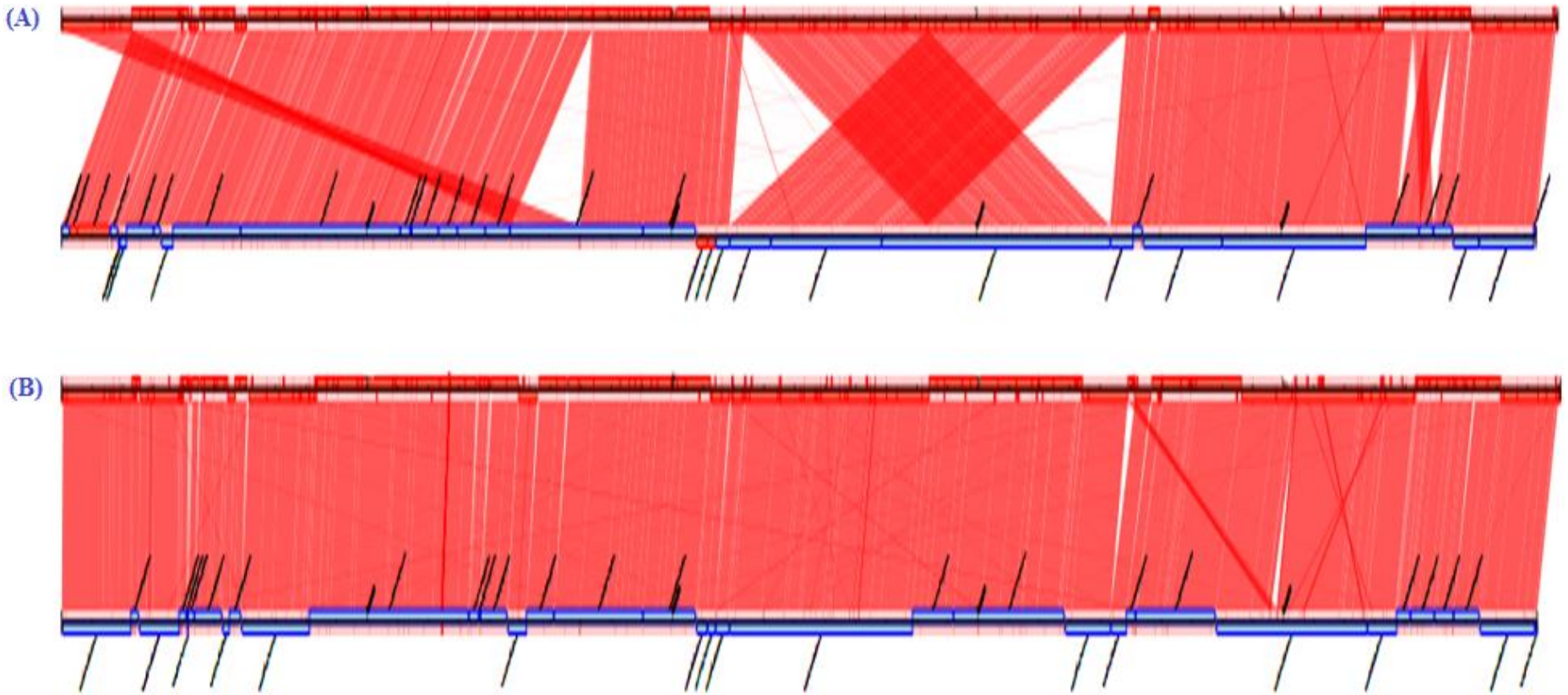


Figura 19. Comparación de los mapas ATC de la secuencia del cromosoma de LSP 389/97 obtenido mediante el método Illumina (parte inferior en A y B) y ensamblado de forma automática por Era7 (A) y con VelvetOptimiser (B). Como referencia se utilizó el cromosoma obtenido y ensamblado por PacBio (parte superior en A y B).

## 4.3. Comparación de las secuencias del plásmido obtenidas mediante las diferentes técnicas de secuenciación

La comparación de las secuencias del plásmido obtenidas por Illumina (ensambladas tanto por Era7 como en el presente TFM) y Roche 454e frente a la secuencia de pUO-STmRV1 obtenida por PacBio (incluida como referencia), permitió conocer el porcentaje de secuencia proporcionado por cada método, así como el número, identidad y orden de los contigs.

En la Figura 21A se observa el alineamiento (mapa ACT) de la secuencia obtenida por el método Illumina ensamblada por Era 7 (abajo) frente a la secuencia PacBio del plásmido (arriba). Se utilizó nuevamente el programa CONTIGuator, que permite obtener una estadística del alineamiento indicando el número de contigs alineados (en este caso pertenecientes al plásmido) y el número de contigs no alineados, que incluye tanto los que pertenecen al cromosoma, como aquellos que pueden presentar problemas incluida una baja calidad en el apareamiento. El número de contigs asignados al plásmido fue 21 de los 185, con un total de 167415 pb que corresponde al 82,7 % de la secuencia del plásmido. Los contigs alineados y el orden de alineamiento es el siguiente: 20, 70, 45, 26, 31, 44, 97, 64, 1, 51, 34, 2, 10, 63, 52, 12, 21, 39, 7, 25 y 72. En el ensamblaje también se observa que la orientación de algunos contigs difiere con respecto a la secuencia de referencia.

En la Figura 21B se puede ver el alineamiento (mapa ACT) de la secuencia PacBio del plásmido con los contigs resultantes del ensamblaje de las lecturas Illumina realizado en este trabajo. Solo 20 de los 153 contigs alinearon con la secuencia de referencia, mientras que el resto correspondería al cromosoma o habrían sido excluidos por razones técnicas. La secuencia correspondiente al plásmido constituye un 82,1% de la obtenida por PacBio, siendo los contigs alineados y su orden de alineamiento 28, 75, 19, 34, 36, 46, 68, 2, 55, 38, 3, 1, 67, 56, 15, 29, 41, 13, 31 y 16. Al igual que en el caso anterior, se observan regiones ausentes en el ensamblaje de las secuencias Illumina con respecto a la secuencia PacBio, así como contigs en orientación opuesta.

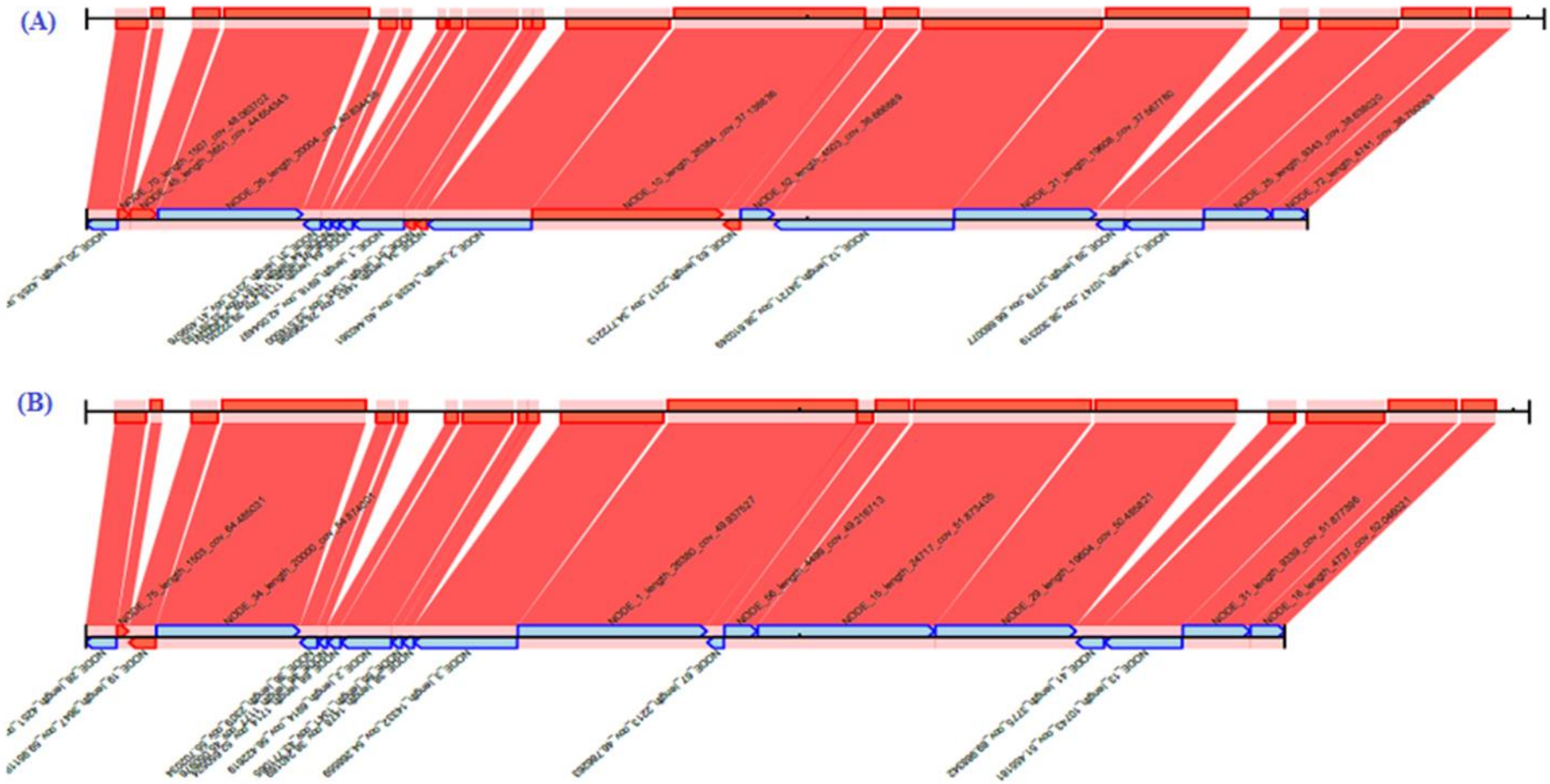


Figura 21. Mapas ACT comparativos de los contigs correspondientes al pUO-STmVR1 obtenidos por Illumina y ensamblados por Era7 (A) y con el programa VelvetOptimiser (B). La parte superior de las figuras corresponde al plásmido secuenciado y ensamblado por PacBio.

Finalmente, al comparar los ensamblajes realizados de forma automática y manual podemos ver que aunque el porcentaje de secuencia obtenida es similar (82,7 versus 82,1) la organización global resultó más ajustada a la realidad en el segundo caso (dos contigs problemáticos frente a seis; señalados en rojo en las Figuras 21A y 21B). Como se comentó anteriormente, el hecho de que los sistemas automáticos maximicen la longitud de los contigs hace que muchas veces sacrifiquen la exactitud estructural para conseguir contigs más largos.

En la Figura 22A se presenta el alineamiento de la secuencia generada por el método Roche 454 y la secuencia PacBio del plásmido. A pesar de que, en este caso, solo se secuenció el plásmido, únicamente 18 de los 30 contigs alinearon con la secuencia de referencia. Esto podría explicarse por una posible contaminación del plásmido con el DNA cromosómico de la bacteria. La secuencia Roche 454 del plásmido presenta múltiples regiones ausentes respecto a la de referencia y la orientación de algunos de los contigs no es la misma. Esta secuencia constituye un 75,5% de la obtenida por PacBio, siendo los contigs alineados y su orden 9, 19, 20, 11, 3, 16, 21, 17, 6, 14, 4, 1, 2, 10, 5, 7, 15 y 8.

Finalmente, en la Figura 22B se ilustra el alineamiento de la secuencia de pUO-STmRV1 obtenida por la combinación de los métodos Sanger/Roche 454 (abajo) con la secuencia PacBio (arriba). Todos los contigs del plásmido, en concreto tres, alinearon con secuencias homólogas contenidas en la de PacBio. En el mapa se observan dos regiones presentes en la secuencia PacBio que se encuentran ausentes en la secuencia Sanger/Roche 454, que estaba por tanto incompleta. Además, en los contigs aparecen varias regiones, de tamaño relativamente pequeño, que muestran homología con distintas posiciones dentro de la secuencia de PacBio, las cuales identifican secuencias repetidas. La secuencia obtenida por el método Sanger constituye el 97,7% de la secuencia PacBio, siendo el orden de alineamiento de los contigs, numerados de menor a mayor tamaño, 1, 3 y 2. De hecho, el mayor porcentaje de secuencia obtenido con respecto a la secuencia de PacBio resultó de la integración de los resultados generados por los métodos de Sanger y Roche 454. Cabe destacar, sin embargo, que el DNA localizado en el extremo 5’ del contig 2 no muestra homología con la región esperada de la secuencia PacBio, indicando un posible error en ensamblaje manual de las secuencias Sanger/Roche 454, llevado a cabo en la Tesis Doctoral de García (2013).

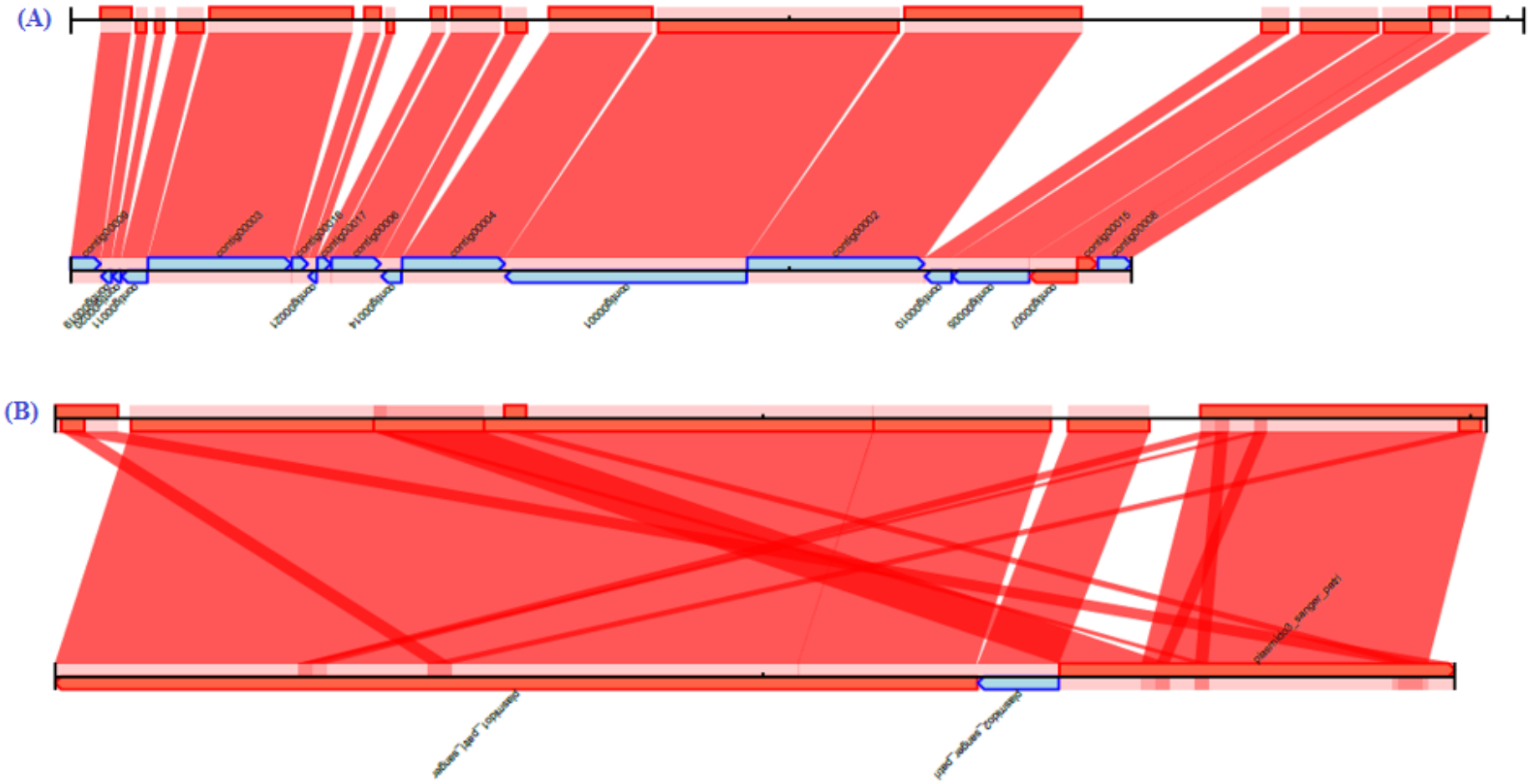


Figura 22. Mapas ACT comparativos de los contigs correspondientes al pUO-STmVR1 obtenidos y ensamblados por Roche (A) y de los contigs obtenidos por Sanger/Roche 454 y ensamblados manualmente (B). La parte superior de las figuras corresponde al plásmido secuenciado y ensamblado por PacBio.

## 4.4. Anotación de la secuencia de pUO-STmRV1

La secuencia del plásmido se anotó de dos maneras: anotación automática mediante el uso del servidor RAST e incorporación de la anotación manual realizada en la tesis doctoral de P. García a la secuencia PacBio con la ayuda del programa Clone.

### 4.4.1. Anotación con el sistema RAST

A partir de la secuencia plasmídica obtenida por el método de PacBio se realizó una anotación automática en el servidor RAST (http://rast.nmpdr.org/rast.cgi). Este servidor utiliza un sistema de anotación automática basada en relaciones filogenéticas como se explicó en el apartado 3.4.

Mediante este sistema se identificaron un total de 261 genes organizados en 14 subsistemas. Sin embargo, el sistema solo asignó una posible función a 52 de ellos, que representan un 15% de los genes detectados (Figura 23). De los genes anotados, el 81% (42) corresponde a sistemas implicados en la resistencia frente a antimicrobianos y a la producción de componentes tóxicos; el 11% (6) codifica funciones de transporte a través de membrana, respiración y respuesta al estrés; el 4% (2) corresponde a genes pertenecientes a fagos, profagos, elementos transponibles y/o plásmidos; y en el último 4% (2) se incluyen genes implicados en funciones de regulación y señalización celular.

Los datos proporcionados por este servidor no resultaron satisfactorios, debido al bajo porcentaje de genes a los que se les asigna una posible función, y a que a muchos de los genes identificados se les asigna la misma función. Por ello, cabe destacar que aunque puede suponer un buen punto de partida para obtener una anotación provisional, es necesario el uso de otras herramientas bioinformáticas que permitan obtener una anotación funcional más precisa de la secuencia. Sin embargo, aunque aquí no ha supuesto una herramienta de mucha utilidad en la anotación de la secuencia del plásmido, el servidor se encuentra realizando mejoras en el servicio, resaltando lo relativo a fagos, plásmidos y fragmentos cortos de ARN, por lo que no es descartable que en un futuro próximo pueda constituir una herramienta válida y fiable en la anotación no solo de plásmidos sino también de genomas completos.

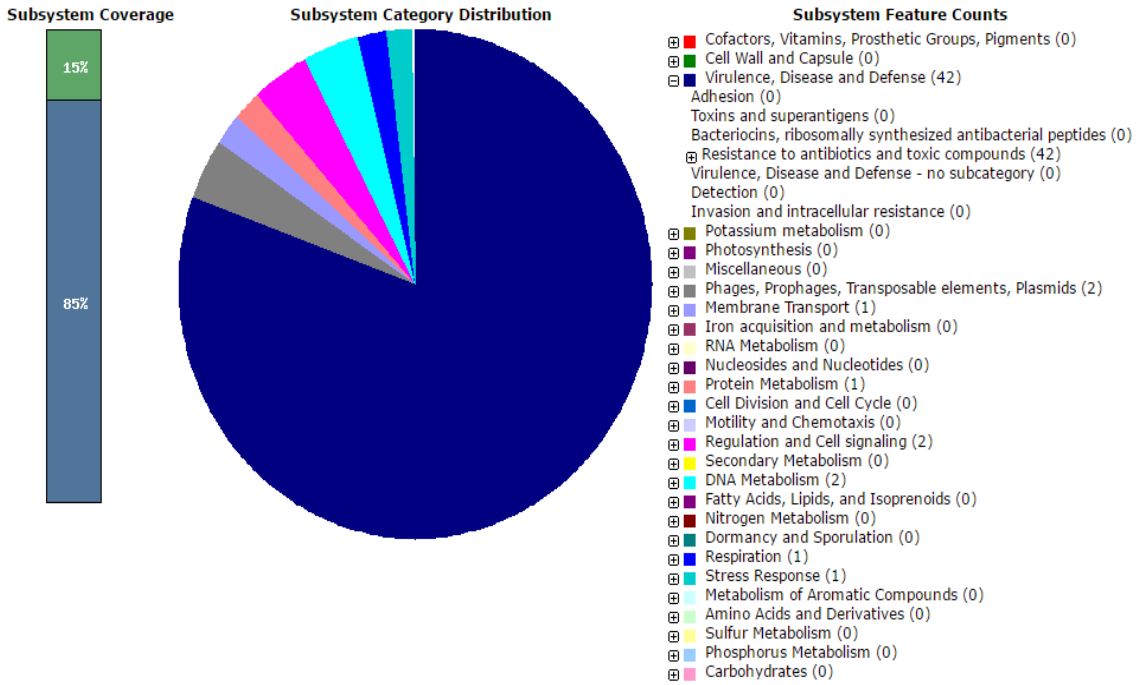


Figura 23. Representación gráfica de los genes anotados automáticamente con el servidor RAST. En la barra vertical se indica en verde el porcentaje de genes anotados cuya función ha sido identificada, mientras que en azul se observa el porcentaje de genes con función desconocida. [Imagen proporcionada por el servidor RAST (http://rast.nmpdr.org/rast.cgi)].

### 4.4.2. Anotación de la secuencia de pUO-STmRV1 con la ayuda del programa Clone Manager

Como consecuencia de las limitaciones observadas con el sistema de anotación automática y con la finalidad de obtener una información funcional válida y fiable, se incorporó a la secuencia PacBio del plásmido la anotación manual de los contigs obtenidos por Sanger/Roche 454 (García, Tesis doctoral, 2013), el software Clone Manager. Se trata de un software altamente flexible, que permite la simulación de experimentos de clonación, ayudando a su diseño, buscar regiones de la secuencia en estudio que presenten homología con otra secuencia de referencia, identificar pautas abiertas de lectura (*orfs*), encontrar y traducir genes dentro de la secuencia, anotar la función de las mismos y editar las secuencias de forma rápida y sencilla. Clone Manager, además de su propio formato de archivos (con extensión .cm5), puede utilizar archivos FASTA, GenBank, EMBL o secuencias de datos en archivos de texto ASCII, lo que aumenta su utilidad.

En este TFM, Clone Manager se utilizó para combinar los contigs obtenidos por Sanger/Roche 454, en el orden y orientación correctos deducidos por comparación con la secuencia PacBio. Durante este proceso se mantiene la anotación de los contigs, transfiriéndose a la secuencia combinada que representa el 97,7% de la secuencia total del plásmido (apartado 4.3).

Este método permitió identificar un menor número de genes (214) y asignarle una función al 63% de ellos (135), porcentaje muy superior al conseguido por RAST (Figura 24). Por otro lado, para facilitar la comparación de resultados, los genes con función inferida, se distribuyeron en ocho categorías funcionales que fueron establecidas por nosotros. Un 33% de estos genes (44) pertenecen a la categoría de genes implicados en funciones de resistencia a antimicrobianos, metales pesados o proteínas de inhibición de fagos. Así, se identificaron genes de resistencia frente a los antimicrobianos ampicilina, cloranfenicol, estreptomicina, gentamicina, sulfonamidas, tetraciclinas y trimetoprim, y frente a biociocidas, en concreto compuestos derivados de amonio cuaternario y diversos metales (mercurio, cobre, plata y arsénico). Otro 33% (44) corresponden a genes implicados en movilidad del DNA, que codifican integrasas y funciones de transposición asociadas con diversas secuencias de inserción (IS*CR2*, IS*CR3*, IS*Ecp1*, IS*Vsa3*, así como 14 copias de IS*26*) y transposones (como Tn*2*, Tn*21* y Tn*1721*). La tercera categoría funcional corresponde a genes plasmídicos que constituyen un 13% del total con función inferida (17). Estos genes codifican para funciones de mantenimiento, replicación y conjugación del plásmido. Un 12% (16) corresponde a la categoría funcional de genes implicados en el metabolismo bacteriano. Estos genes codifican oxidorreductasas, proteínas de transporte, permeasas, topoisomerasas, metiltranferasas, mutasas, peptidasas, nitrito reductasas, ATPasas, anhidrasa carbónica, etc. La categoría funcional de virulencia constituye un 5% (7) e incluye genes derivados del plásmido de virulencia específico del serotipo Typhimurium, importantes para la infección sistémica. El 4% de los genes (6) constituyen la categoría regulación, en la cual se encuentran genes implicados en la regulación de los sistemas celulares, como activadores transcripcionales o proteínas de unión al ADN. El 1% restante (1), corresponde a la última categoría denominada otros.

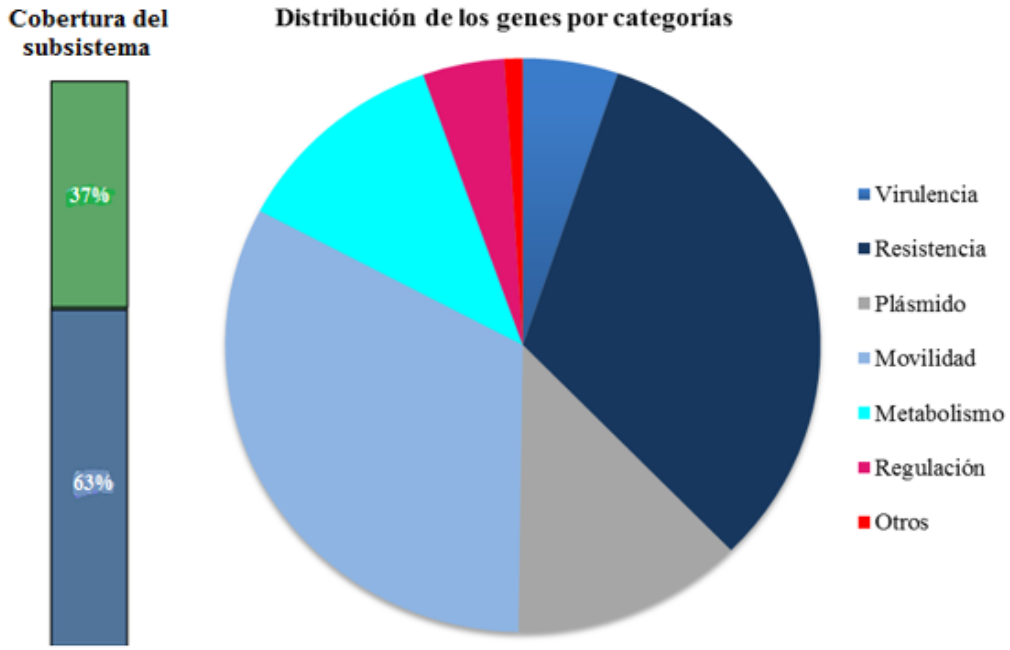


Figura 24. Distribución por categorías de los genes identificados de forma manual mediante el software Clone Manager 9. En la barra vertical se indica en verde el porcentaje de genes anotados cuya función ha sido identificada, mientras que en azul se observa el porcentaje de genes con función desconocida.

En resumen, aunque los servicios de anotación automática han supuesto un gran avance en el análisis de las secuencias acelerando el ritmo de trabajo, los resultados obtenidos apoyan la necesidad del uso de herramientas y análisis manuales para obtener información fiable y exhaustiva de la secuencia.

Finalmente, cabe destacar que los métodos automatizados han supuesto un gran ahorro en el tiempo y los costes de secuenciación, que han ido disminuyendo año tras año. Sin embargo, esta automatización requiere un mayor uso de herramientas bioinformáticas, las cuales se basan en softwares que pueden cometer errores. Por ello es de capital importancia la intervención de especialistas que revisen y, en su caso, corrijan manualmente los resultados tanto del ensamblaje como de la anotación.

# Conclusiones

1. El desarrollo de nuevas técnicas de secuenciación supone una disminución en el coste y el tiempo de obtención de resultados, permitiendo un estudio más rápido y completo de las características genéticas globales de los seres vivos.
2. La presencia de DNA repetido dificulta la obtención de secuencias completas.
3. Entre todos los métodos de secuenciación evaluados en este trabajo, solo la plataforma PacBio (tercera generación) proporcionó el genoma del aislamiento representativo del clon monofásico español de *S. enterica*: cromosoma y plásmido pUO-STmRV1, en secuencias únicas y completas.
4. Las herramientas bioinformáticas son imprescindibles para extraer de forma rápida y precisa la ingente cantidad de información aportada por las técnicas de secuenciación de segunda y tercera generación.
5. A pesar de las mejoras proporcionadas por la automatización de las herramientas bioinformáticas, se requiere personal cualificado con capacidad de optimizar los parámetros y realizar un análisis exhaustivo de los resultados.
6. Los sistemas de anotación automática tipo RAST constituyen una herramienta de partida útil pero resulta imprescindible una comprobación manual de la anotación obtenida y una mejora en la asignación de funciones.
7. Los métodos de secuenciación masiva han supuesto un gran avance en el conocimiento y seguimiento de los patógenos patógenos bacterianos transmitidos por alimentos, como es el caso de *S. enterica*.

# Lista de abreviaturas

**ACT**: Herramienta de comparación Artemis (Artemis Comparison Toon)

**ADN**: Ácido desoxirribonucleico

**ARN$_t$**: Ácido ribonucleico de transferencia

**APS**: Adenosina- 5'- fosfosulfato

**ATP**: Adenosín trifosfato

**CCD**: Dispositivo de carga acoplada (Charge-Coupled Device)

**CS**: Segmento conservado

**CRT**: Terminación cíclica reversible (Cyclic Reversible Termination)

**ddNTP**: Didesoxirribonucleótido

**dNTP** : Desoxirribonucleótido

**emPCR**: Reacción en cadena de la polimerasa en emulsión

**GC**: Guanina- citosina

**IS**: Secuencia de inserción

**LPS**: Lipopolisacárido

**NGS**: Secuenciación de siguiente generación (Next- generation sequencing)

**orf**: Pauta de lectura abierta (Open Read Frame)

**pb**: Pares de bases

**PCR**: Reacción en cadena de la polimerasa

**PPi**: Pirofosfato inorgánico

**RAST**: Rapid Annotations using Subsystems Technology

**RV**: Región variable

**SGS**: Secuenciación de segunda generación (Second- generation sequencing)

**SIM**: Sistema de Información Microbiológica

**SMRT**: Single-Molecule Real-Time

**SPI**: Islas de patogenicidad (*Salmonella* Patogenicity Islands)

**SST**: Sistema de secreción

**TGS**: Secuenciación de tercera generación (Third- generation sequencing)

**ZMV**: Zero Mode Waveguide

# Bibliografía

**Alekshum, M. N., S. B. Levy**, (2007). Molecular mechanisms of antibacterial multidrug resistance. Cell. **128**: 1037-1050.

**Ansorge, W. J.**, (2009). Next-generation DNA sequencing techniques. N Biotechnol. **25 (4)**: 195-203.

**Antunes, P., J. Mourão, N. Pestana, y L. Peixe**, (2011). Leakage of emerging clinically relevant multidrug-resistant *Salmonella* clones from pig farms. J Antimicrob Chemother. **66**: 2028- 2032.

**Bennett, P. M.**, (2008). Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria. Br J Pharmacol. **153**: S347-57.

**Boletín epidemiológico semanal (ISC) (**http://www.isciii.es/ISCIII/es/contenidos/fd-servicios-cientifico-tecnicos/fd-vigilancias-alertas/fd-enfermedades/enfermedade s - declaracion-obligatoria-informes-anuales.shtml) [Accesado el día 11 de enero de 2016].

**Cantón, R., T. M. Coque y F. Baquero**, (2003). Multi-resistance Gram-negative bacilli: from epidemics to endemics. Curr Opin Infect Dis. **16**: 315-325.

**Caratolli, A**., (2009). Resistance plasmid families in *Enterobacteriaceae*. Antimicrob Agents Chemother. **53**: 2227-2238.

**Clone Manager 9** (http://www.scied.com/pr_cmbas.htm).

**CONTIGuator (**http://contiguator.sourceforge.net) [Accesado el día 22 de mayo de 2016].

**Corral, J. L. y E.J. Perea**, (1992). *Salmonella*. En: Enfermedades Infecciosas y Microbiología Clínica. E. J. Perea (ed). Ed. Dogma, Barcelona. 647-661.

**Echeita, M. A., A. Aladuena, S. Cruchaga, y M. A. Usera**, (1999). Emergence and spread of an atypical *Salmonella enterica* subsp. *enterica* serotype 4, 5, 12:i:- strain in Spain. J Clin Microbiol. **37**: 3425.

**Echeita, M. A., S. Herrera, y C. Baamonde**, (2011). Gastroenteritis invasivas ¿algo nuevo?. Enferm Infect Microbiol Clin. **29**: 55-60.

**Edwars, D. J. y K. E. Holt**, (2013). Beginner's guide to comparative bacterial genome analysis usong next-generation sequence data. BioMed Central.

**Eid, J., A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clarks, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse,**

**Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zao, F. Zhong, J. Korlach y S. Turner**, (2009). Real-time DNA sequencing from single polymerase molecules. Science. **323**: 133-138.

**European Food Safety Authority (EFSA)**, (2010). Scientific Opinion on monitoring and assessment of the public health risk of "*Salmonella* Typhimurium-like" strains. EFSA J. **8**: 1826.

**FastQC (**http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) [Accesado el día 25 de mayo de 2016].

**Fierer, J. y D. G. Guiney**, (2001). Diverse virulence traits underlying different clinical outcomes of *Salmonella* infection. J Clin Invest. **107**: 775- 780.

**Fierer, J. y D. G. Guiney**, (2011). The role of the *spvC* gene in *Salmonella* Pathogenesis. Front Microbiol. **2**: 129.

**Fluit, A. C. y F. J. Schmidt**, (2004). Resistance integrons and super-integrons. Clin Microbiol Infect. **10:** 272-288.

**García, P**., (2013). Bases moleculares de la resistencia y la virulencia en la variante monofásica 4,[5],12:i:- de *Salmonella enterica* serotipo Typhimurium. Tesis doctoral. Universidad de Oviedo.

**García, P., B. Guerra, M. Bances, M. C. Mendoza y M. R. Rodicio**, (2011). IncA/C plasmids mediate antimicrobial resistance linked to virulence genes in the Spanish clone of the emerging *Salmonella enterica* serotype 4,[5],12:i:-. J Antimicrobiol Chemother. **66**: 543- 549.

**García, P., B. Malorny, E. Hauser, M. C. Mendoza y M. R. Rodicio**, (2013). Genetic types, gene repertoire, and evolution of isolates of the *Salmonella enterica* serovar 4,5,12:i:- Spanish clone assigned to different phage types. J Clin Microbiol. **51**: 973-978.

**García, P., K. Hopkins, V. García, J. Beutlich, M. C. Mendoza, J. Threlfall, D. Mevius, R. Helmunth, M. R. Rodicio y B. Guerra**, (2014). Diversity of plasmids encoding virulence and resistance functions in *Salmonella enterica* subsp. *enterica* serovar Typhimurium monophasic variant 4,[5],12:i:- strains circulating in Europe. PLoS ONE 9(2): e89635. doi: 10.1371/journal.pone.0089635.

**Grimont, P. A., y F. X. Weill**, (2007). Antigenic formulae of the *Salmonella* serovars. 9ª Edición. WHO collaborating Center for Reference and Research on *Salmonella*, Institute Pateur, Paris, France.

**Gibson, D. L., A. P. White, C. M. Rajotte y W. W. Kay**, (2007).AgfC and AgfE facilitate extracelular thin aggregative fimbriae synthesis in *Salmonella* Enteritidis. Microbiology **153**: 1131- 1140.

**Guerra, B., S. M. Soto, J. M. Argüelles, y M. C. Mendoza**, (2001). Multidrug resistance is mediated by large plasmids carrying a class 1 integron in the emergent *Salmonella enterica* serotype [4,5,12:i:-]. Antimicrob Agents Chemother. **45**: 1305-1308.

http://www.sosenfermero.com/noticias-de-salud/profesionales/salmonella/ [Accesado el día 22 de enero de 2016].

http://www.ispch.cl/lab_amb/serv_lab/salmonella.html [Accesado el día 22 de enero de 2016].

http://ftp.utalca.cl/profesores/raherrera/biorom2008/contenido/cibertexto/hc/sugar35b.htm [Accesado el día 23 de enero de 2016].

http://ulum.es/secuenciacion-sanger/ [Accesado el día 15 de abril de 2016].

**Humphries, A. D., M. Raffatellu, S. Winter, E. H. Weening, R. A. Kingley, R. Droleskey, S. Zhang, J. Figueiredo, S. Khare, J. Nunes, L. G. Adams, R. M. Tsolis y A. J. Baümler**, (2003). The use of flow cytometry to detect expression of subunits encoded by 11 *Salmonella enterica* serotype Typhimurium fimbrial operons. Mol Microbiol. **48**:1357-1376.

**Kleckner, N.**, (1981). Trasposable elements in prokariotes. Annu Rev Genet. **15**:341-404.

**Kurtz, S., A. Phillippy, A. L. Delcher, M Smoot, M Shumway, C. Antonescu y S. L. Salzberg**, (2004). Versatile and open software for comparing large genomes. Genome Biol. **5**: R12.

**Lucarelli, C., A. M. Dionisi, M. Torpdahl, L. Villa, C. Graziani, K. Hopkins, J. Threfall, A. Capriolli y I. Luzzi**, (2010). Evidence for a second genomic island conferring multidrug resistance in a clonal group of strains of *Salmonella enterica* serovar Typhimurium and its monophasic variant circulating in Italy, Denmark and the United Kingdom. J Clin Microbiol. **48**: 2103-2109.

**Marcus, S. L., J. H. Brumell, C. G. Pfeifer y B. B. Finlay**, (2000). *Salmonella* pathogenicity islands: big virulence in small packages. Microbes Infect. **2**: 145-156.

**Mardis, E. R.**, (2008). Next-generation DNA sequencing methods. Ann Rev Genomics Hum Genet **9**: 387-402.

**Margulies, E. H., C. W. Chen y E. D. Green**, (2006). Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparison. Trends Genet. **22(4)**: 187- 193.

**Maxam, A. M. y W. Gilbert**, (1980). Sequencing end-labeled DNA with base-specific chemical cleavages.

**Metzker, M. L.**, (2010). Sequencing technologies – the next generation. Nature Rev **11**: 31-46.

**Mossel, D. A. A., B. Moreno y C. B. Struijk**, (2002). Microbiología de los alimentos. 2ª edición. Ed Acribia, Zaragoza.

**Mourão, J., J. Machado, C. Novais, P. Antunes y L. Peixe**, (2014). Characterization of the emerging clinically-relevant multidrug-resistant Salmonella enterica serotype 4, [5],12:i:- (monophasic variant of S. Typhimurium) clones. J Clin Microbiol. **33**: 2249-2257.

**Raffatellu, M., D. Chessa, R. P. Wilson, C. Tükel, M. Akçelik y A. J. Baümler**, (2006). Capsuled-mediated inmune evasión: a new hypothesis explaining aspects of typhoid fever pathogenesis. Infect Inmum. **74**: 19-27.

**Raffatellu, M., R. P. Wilson, S. E. Winter y A. J. Bäumler**, (2008). Clinical Pathogenesis of typhoid fever. J Infect Dev Ctries. **2**: 260-266.

**RAST** http://rast.nmpdr.org/rast.cgi [Accesado el día 25 de mayo de 2016].

**Rodicio, M. R., A. Herrero, I. Rodríguez, P. García, I. Montero, J. Beutlich, R. Rodicio, B. Guerra y M. C. Mendoza**, (2011). Acquisition of antimicrobial resistance determinants by virulence plasmids specific for nontyphoid serovars of *Salmonella enterica*. Med Microbiol. **22**:55-65.

**Sanger, F. y A. R. Coulson**, (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polimerase. J Mol Biol. **94**: 441-448.

**Schmidt, H. y M. Hensel**, (2004). Pathogenicity islands in bacterial pathogenesis. Clin Microbiol Rev. **17**: 14- 56.

**Shendure, J. y H. Ji**, (2008). Next- generation DNA sequencing. Nat Biotechnol. **26 (10)**: 1135- 1145.

**Switt, A. I., Y. Soyer, L. D. Warnick, y M. Wiedmann**, (2009). Emergence, distribution, and molecular and phenotypic characteristics of *Salmonella enterica* serotype 4,5,12:i:-. Foodborne Pathog. Dis. **6:**407-415.

**Tindall, B. J., P. A. Grimont, G. M. Garrity, y J. P. Euzeby**, (2005). Nomenclature and taxonomy of the genus *Salmonella*. Int J Syst Evol Microbiol. **55**:521-524.

**Torres, C., M. A. Moreno, M. Zaragaza**, (2010). Prudent use of antimicrobial agents: not just for humans. Enferm Infecc Microbiol Clin. **28**: 669-671.

**Uzzau, S., D. J. Brown, T. Wallis, S. Rubino, G. Leori, S. Bernard, J. Casadesús, D. J. Platt y J. E. Olsen**, (2000). Host adapted serotypes of *Salmonella enterica*. Epidemiol Infect. **125**: 229-255.

**Velvet** (www.ebi.ac.uk/~zerbino/velvet_latest.tgz).

**Weening, E. H., J. D. Barker, M. C. Laarakker, A. D. Humphries, R. M. Tsolis y A. J. Baümler**, (2005). The *Salmonella enterica* serotypr Typhimurium *Ipf, bcf, stb, stc, std* and *sth* fimbrial operons are required for intestinal persistence in mice. Infect Immun. **73**: 3358-3366.

**Zerbino, D. R. y E. Birney**, (2008). Velvet: algorithms for the *novo* short read assembly using the Bruijn graphs. Genome Res **18(5)**:821-829.

**Zerbino, D. R.**, (2010). Using the Velvet de *novo* assembler for short-read sequencing technologies. Curr Protoc Bioinformatics **11**: Unit 115.

**Zhang, J., R. Chiodini, A. Badr y G. Zhang**, (2011). The impact of next-generation sequencing on genomics. J Genetics and Genomic. **38**: 95- 109.