TÍTULO

VARIABILIDAD NATURAL de la ISLA GENÓMICA 1 DE *Haloquadratum walsbyi*

AUTORA

Ana Belén Martín Cuadrado

Esta edición electrónica ha sido realizada en 2013

Tutores	Francisco Rodríguez Valera y Enrique Viguera Mínguez
Curso	*Máster en Bioinformática*
ISBN	978-84-7993-571-9
©	Ana Belén Martín Cuadrado
©	De esta edición: Universidad Internacional de Andalucía
Fecha documento	Abril 2012

Reconocimiento-No comercial-Sin obras derivadas

Usted es libre de:

- Copiar, distribuir y comunicar públicamente la obra.

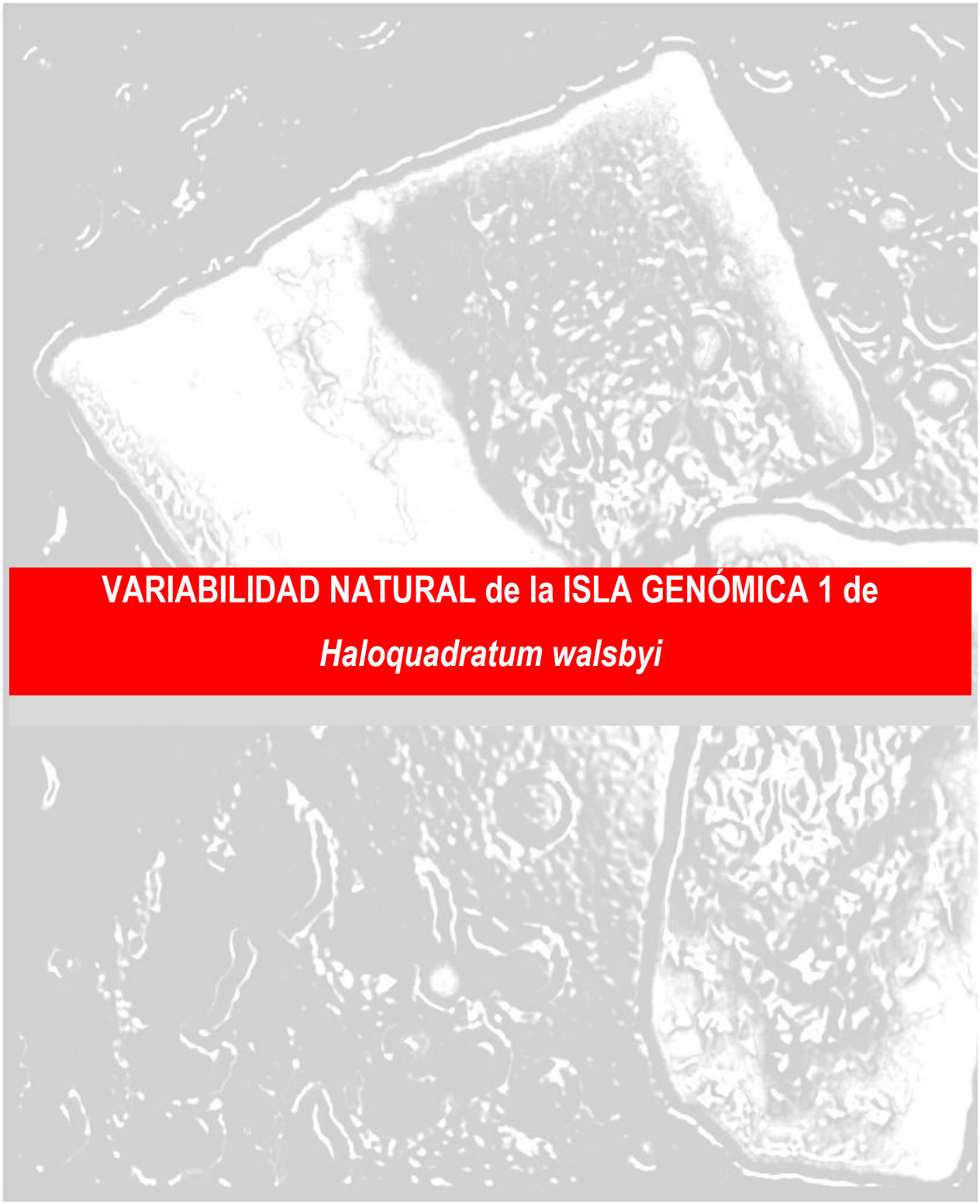
Bajo las condiciones siguientes:

- **Reconocimiento**. Debe reconocer los créditos de la obra de la manera. especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).
- **No comercial**. No puede utilizar esta obra para fines comerciales.
- **Sin obras derivadas**. No se puede alterar, transformar o generar una obra derivada a partir de esta obra.
- *Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra.*
- *Alguna de estas condiciones puede no aplicarse si se obtiene el permiso del titular de los derechos de autor.*
- *Nada en esta licencia menoscaba o restringe los derechos morales del autor.*

MÁSTER EN BIOINFORMÁTICA

UNIVERSIDAD INTERNACIONAL DE ANDALUCÍA. UNIA.

PROYECTO FIN DE MÁSTER. ***ABRIL 2012.***



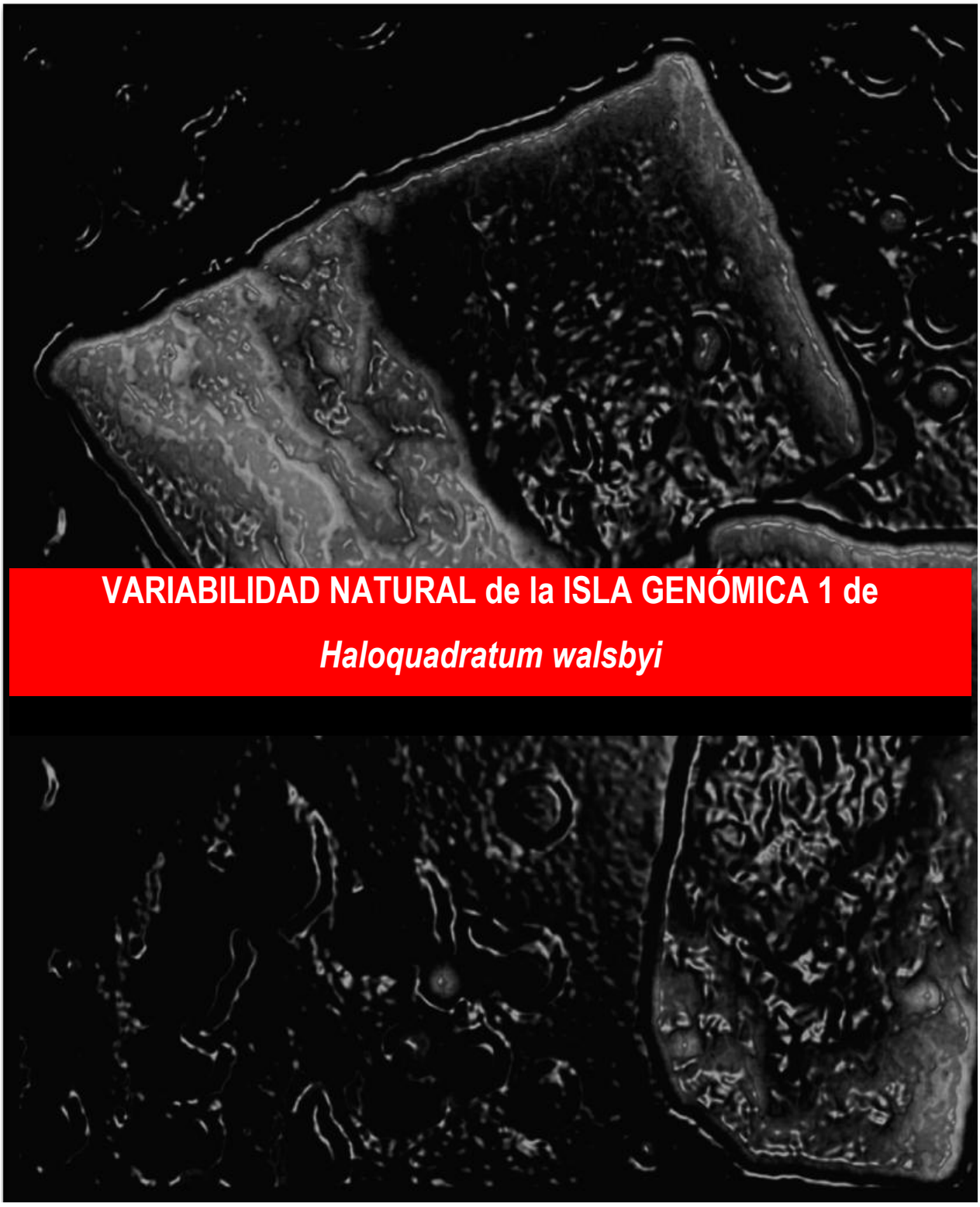
VARIABILIDAD NATURAL de la ISLA GENÓMICA 1 de *Haloquadratum walsbyi*

Trabajo realizado por: **Ana Belén Martín Cuadrado**

Tutores:

Francisco Rodríguez Valera, Catedrático de Microbiología de la Universidad Miguel Hernández

Enrique Viguera Mínguez, Profesor Titular de la Universidad de Málaga



VARIABILIDAD NATURAL de la ISLA GENÓMICA 1 de *Haloquadratum walsbyi*

ÍNDICE

A Vega, por no dejarme dormir con sus besos.

A "Papá", por hacerme más fuerte.

A mi familia, por quererme.

ABREVIATURAS

ANI: Media de identidad nucleotídica

ALP: “adhesin-like protein”, putativa adhesina

CSG: glicoproteína de la pared celular, del inglés “cell Surface glycoprotein”

CR30: cristalizador 30 de la salina “Bras del Port” en Santa Pola, Alicante

HGT: “horizontal gene transfer”, transferencia genética horizontal

MID: “multiple identifier”, marcador usado en la pirosecuenciación

MCSG: glicoproteína mayoritaria de la pared celular o mayoritaria de la capa S, del inglés “mayor cell surface glycoprotein”

OTU: Unidad Taxonómica Operacional

ORF: “open reading frame”, marco de lectura abierta

VLP: “virus like particle”, partículas virales

W/V: Relación Peso / Volumen

RESUMEN

La metagenómica ha revelado una inesperada variabilidad dentro de una especie procariótica que ha llevado a formular el concepto de "pan-genoma" para definir el repertorio de genes de una especie. El genoma flexible o adaptativo, a diferencia del denominado "core-genoma", no se ha conservado en todos los aislados y muchas veces aparecen en islas genómicas de tamaño variable. Estas islas genómicas varían de una cepa a otra, aún incluso si proceden de un mismo entorno. Teóricamente, en un metagenoma, si la especie es muy abundante y el nivel alcanzado de secuenciación es el de saturación del sistema, sería posible encontrar todos los clones o linajes representantes de la especie bacteriana en ese hábitat. Las mayores diferencias entre estos se deberían a las islas genómicas. Un patrón muy bien conservado en todos los genomas estudiados es la presencia de muchos genes relacionados con los componentes extracelulares de la pared celular: lipo-polisacárido, capa S, cápsulas, ... etc, en las islas genómicas. Partiendo de esta idea, y usando como modelo de hábitat el cristalizador de una salina solar, se han analizado varios fragmentos genómicos pertenecientes a varios linajes co-existentes de la archaea cuadrada *Haloquadratum walsbyi*, la cual domina la flora microbiana de aguas hipersalinas. En concreto, el área del genoma estudiada es la isla genómica rica en genes que codifican proteínas implicadas en la envuelta celular de esta archaea (denominda GI1 en este trabajo). A partir de una librería de fósmidos ambiental construida a partir de una muestra recogida en el mismo cristalizador (CR30) donde el primer aislado de *H. walsbyi* se obtuvo (cepa HSBQ001), se han secuenciado 7 fósmidos que cubren GI1, donde se encuentrasn, entre otros, el gen para la "proteína mayoritaria de la capa S (MCSG)" y otras glicoproteínas de la superficie celular (CSGs). Además de estos fragmentos, hemos sumado en el análisis, las islas GI1 presentes en los genomas y la descrita previamente presente en el fósmido eHw-559.

Los resultados indican que GI1 contiene genes que codifican componentes de la envoltura celular y éstos presentan una variabilidad muy alta. En todos los casos, estos fragmentos genómicos reclutan muy poco en el metagenoma SS37, obtenido a partir de una muestra recogida 4 años después de la muestra con se construyó la librería y aproximadamente dos respecto la del primer aislado HSBQ001. La ausencia de homólogos en el metagenoma indica la co-existencia de varios clones de *H. walsbyi* con un conjunto de proteínas diferentes para construir la envoltura exterior de las células. El hallazgo de una versión casi exacta de la isla genómica del aislado australiano, C23, indica que los clones de *H. walsbyi* están ampliamente distribuidos y que persisten a lo largo del tiempo (al menos durante 4 años). La variabilidad encontrada de glicoproteínas de superficie celular en cada uno de los fósmidos probablemente refleja la presión a la que este tipo de proteínas están sometidas: ser la diana de fagos, que en este hábitat son diez veces más abundantes que en cualquier otro descrito. De acuerdo con el modelo de "diversidad constante" propuesto por Rodriguez-Valera *et al.* (2009), la coexistencia de varias versiones de esta región ayudaría a distribuir entre los diferentes clones de la comunidad la probabilidad de ser reconocido e infectado por un tipo de fago que reconozca alguna de las estructuras expuestas específicas de clon o linaje. Por lo tanto, en *H. walsbyi*, esta variabilidad refleja una evasión de los fagos mediante el mantenimiento de la microdiversidad. Luego para explicar la coexistencia de tan alto número de fagos y de células de *H. walsbyi*, debe de existir una alta diversidad de genes de reconocimiento de fagos, muy probablemente las CSGs localizadas en

las islas genómicas. Las descritas en el presente trabajo podrían ser una pequeña representación de cómo pueden llegar a variar estas dianas víricas.

INTRODUCCIÓN

INTRODUCCIÓN

El presente trabajo presenta el análisis bioinformático de varios fósmidos pirosecuenciados pertenecientes a una zona muy concreta del genoma de la archaea cuadrada hiperhalófila *Haloquadratum walsbyi*, la isla genómica 1 (GI1). Esta área se fue definida como tal en el trabajo de Cuadros-Orellana et al. (2007) usando como referencia el genoma de *H. walsbyi* DSM 16790. En este trabajo, se usará GI1 refiriéndose siempre al área del genoma de *H. walsbyi* rica en genes que codifican proteínas implicadas en la envuelta celular, indistintamente de la cepa y de si es o no la primera de las islas genómicas detectadas.

En este apartado, Introducción, se comentarán los siguientes aspectos:

- Los datos más relevantes conocidos hasta la fecha sobre la comunidad procariota presente en el cristalizador de una salina solar, y más concretamente, en el cristalizador CR30 de la salina "Bras del Port" en Santa Pola (Alicante), estudiado durante muchos años en el laboratorio donde se ha desarrollado el presente trabajo.
- Los datos existentes publicados sobre *H. walsbyi*: genomas que se encuentran secuenciados y los metagenomas que se han obtenido a partir de la biomasa recogida en el cristalizador CR30.
- Los aspectos más importantes del pan-genoma bacteriano.
- Los diferentes métodos de secuenciación "de nueva generación" y el empleado en la secuenciación de los fósmidos.

1. AMBIENTES ACUÁTICOS HIPERSALINOS Y MICROORGANISMOS ASOCIADOS

Se consideran ambientes hipersalinos aquellos cuya salinidad supera la del agua de mar (aproximadamente 3,5 %). Estos ambientes se pueden clasificar en dos tipos:

- **Talasohalinos:** en este tipo de ambientes la proporción iónica es semejante a la del agua de mar puesto que se originan por la evaporación de la misma. El ion predominante es el sodio y los aniones predominantes, el cloruro y el sulfato. Normalmente, en este tipo de ambientes el pH es neutro o ligeramente básico. Los lagos salinos de origen marino como el Gran Lago Salado de Utah (Estados Unidos) y las salinas solares pertenecen a este tipo de ambientes hipersalinos.
- **Atalasohalinos**: la proporción de iones es diferente a la encontrada en el agua de mar. Este tipo de ambiente está dominado principalmente por los iones K^+, Mg^{2+} y Na^+. Dentro de estos ambientes están los lagos salados alcalinos, ricos en cationes Mg^{2+} y Ca^{2+}, que contienen elevadas concentraciones de carbonato/bicarbonato y presentan valores de pH superiores a 10. Ejemplos de estos ambientes hipersalinos atalasohalinos son el Mar Muerto (Israel, Jordania y Palestina) donde el Mg^{2+} y el Ca^{2+} son los cationes más abundantes y el sulfato se encuentra en bajas proporciones, el lago alcalino Magadi (Kenia), el lago Chaka (Michina *et al.*), el algo alcalino Wadi Natrum (Egipto) y el Lago Mono (California).

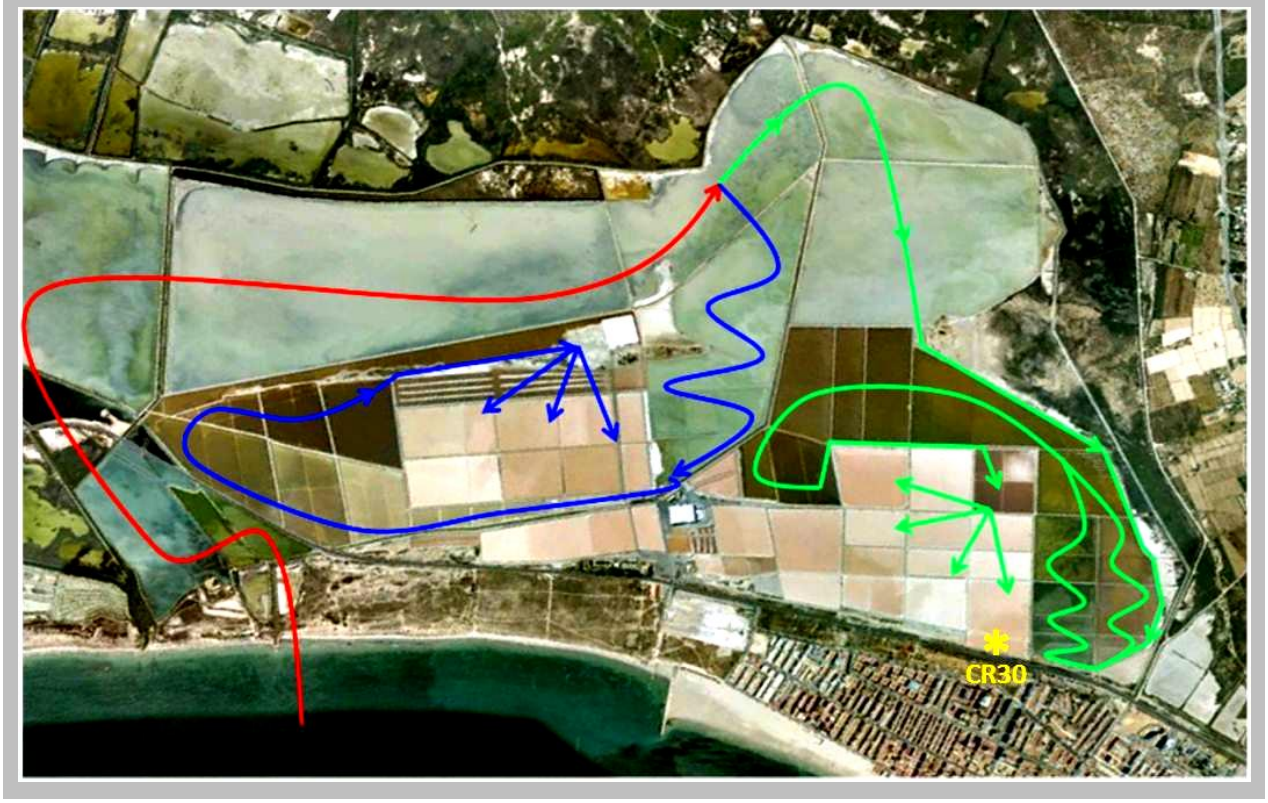
Estos ambientes hipersalinos son el hábitat de los microorganismos halófilos, tradicionalmente descritos como aquellos que tienen un óptimo de crecimiento a una concentración de sal de 50 g/l o superior y toleran, al menos, 100 g/l de sal (Oren 2008). Dentro de los microorganismos halófilos, se puede diferenciar entre halófilos extremos, que son los que crecen mejor en un medio que contenga entre 147 g/l y 306 g/l y halófilos moderados, que crecen mejor en un medio que contenga entre 30 y 147 g/l. Existen también microorganismos halotolerantes, que no necesitan sal para su crecimiento, pero que pueden crecen en presencia de la misma (Oren 2008). Los ambientes hipersalinos son ambientes extremos, en los cuales la elevada concentración de sales normalmente no es el único factor que limita la diversidad de la biota (Ventosa 2006). Otros parámetros que afectan a la diversidad en estos ambientes son la disminución de la concentración de oxígeno a medida que aumenta la salinidad y las altas o bajas temperaturas. Además, también pueden afectar a la biodiversidad encontrada en estos ambientes la presión, que puede ser alta o baja dependiendo del área geográfica en la que se encuentren, la baja disponibilidad de nutrientes, la radiación solar y/o la presencia de metales pesados y otros compuestos tóxicos (Rodríguez-Valera 1988). Al examinar la distribución de los microorganismos halófilos en el árbol filogenético de los seres vivos, basado en las secuencias de los genes del 16S y 18S rRNA, se ha visto que existen representantes de los mismos en los tres Dominios, *Archaea*, *Bacteria* y *Eukarya* (Oren 2008). Dentro del Dominio *Archaea*, los microorganismos hiperhalófilos se encuentran dentro del orden *Halobacteriales*, siendo algunos de sus representantes miembros de los géneros *Halobacterium*, *Haloarcula* y *Haloquadratum*. Todos ellos pertenecen al filo *Euryarchaeota*. Dentro del Dominio *Bacteria*, los microorganismos más halófilos se encuentran dentro del filo *Bacteroidetes*, aunque también existen representantes halófilos en los filos *Cyanobacteria*, *Proteobacteria*, *Firmicutes*, *Actinobacteria* y

Spirochaetes. Entre los representantes del Dominio *Eukarya*, se encuentran el alga halófila unicelular *Dunaliella* que es el principal productor primario en los ambientes hipersalinos (Oren 2005), el crustáceo *Artemia salina* y hongos como *Trimmatostromas salinum* .

1.1. Las salinas solares.

Las salinas solares, localizadas principalmente en zonas tropicales y subtropicales de todo el mundo, son sistemas donde se produce cloruro sódico a partir de agua de mar (Figura 1). Las salinas solares son ambientes hipersalinos talasohalinos que además de una elevada concentración de sales presentan otras características como la de estar sometidas a una fuerte irradiación solar, a fuertes oscilaciones de la temperatura entre el día y la noche y poseer un bajo contenido en oxígeno, ya que su solubilidad en estos medios tan concentrados en sales es muy baja (Rodriguez-Valera *et al.* 1983).

Figura 1. Funcionamiento de las salinas solares "Bras del Port" (Santa Pola, Alicante). En rojo se muestra la entrada de agua desde el Mediterráneo al circuito de la salina. En verde, el circuito del "Bras" y en azul el circuito del "Teniente". El asterisco amarillo, muestra la situación del cristalizador CR30.



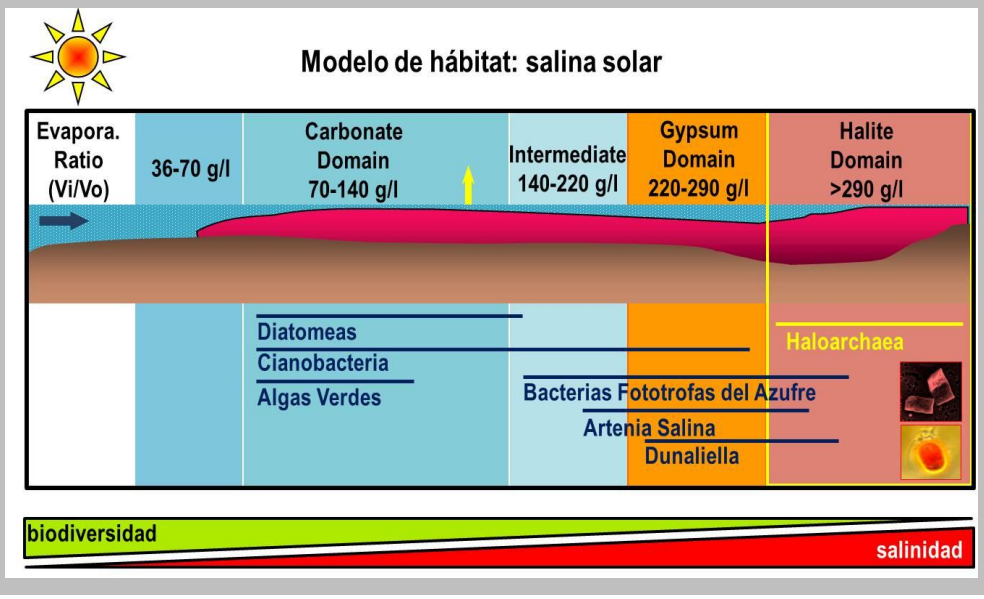
Las salinas están constituidas por una serie de estanques contiguos, donde se produce un gradiente de salinidad discontinuo y la precipitación fraccionada de la sal mediante la evaporación del agua de mar. En el primer tipo de estanques, denominados "preparadores", precipita el carbonato cálcico. El siguiente tipo de estanques son los "concentradores" donde precipita el sulfato cálcico. Finalmente, el tercer tipo de estanques son los "cristalizadores"

donde precipita el cloruro sódico (ver esquema Figura 2). El agua que queda después de la cristalización del cloruro sódico, la salmuera, es rica en $MgCl_2$ (Hallsworth *et al.* 2007; Oren 1994). Sólo unos pocos grupos de microorganismos son capaces de crecer a concentraciones de $MgCl_2$ superiores a 2,5 M, siempre y cuando exista una elevada concentración de cloruro sódico. Estas salmueras enriquecidas en cloruro de magnesio se pueden utilizar para lavar la sal generada en los cristalizadores o bien se recircularizan al mar.

1.2. Microbiota de las salinas.

La microbiota de las salinas solares ha sido estudiada extensamente mediante técnicas de cultivo y moleculares (Anton *et al.* 1999; Anton *et al.* 2000; Benlloch *et al.* 2001; Benlloch *et al.* 2002; Casamayor *et al.* 2002; Estrada *et al.* 2004; Litchfield & Gillevetr 2002; Maturrano *et al.* 2006; Oren 1994; Oren 2002a; Oren 2002b; Oren & Rodriguez-Valera 2001; Rodríguez-Valera *et al.* 1999). Estudios realizados en los cristalizadores de las salinas solares, mediante técnicas de análisis de pigmentos, citometría de flujo y técnicas moleculares, revelaron que conforme aumenta la salinidad se produce un descenso en la diversidad procariota indicando un efecto selectivo en los ambientes extremos (Estrada *et al.* 2004). Sin embargo, a pesar de esta baja diversidad, la densidad de células en el cristalizador puede llegar a alcanzar las 10^7-10^8 células/ml (Oren 2002b). En los estanques de alta salinidad, sólo unas pocas especies de microorganismos, halófilos extremos y muy especializados (generalmente quimio-organoheterótrofos), son capaces de crecer nutriéndose de los compuestos orgánicos procedentes del metabolismo y de la descomposición de las poblaciones de *Dunaliella* spp. (alga verde unicelular y halófila) y de otros microorganismos que se desarrollan a salinidades más bajas (Pedros-Alio *et al.* 2000). En los estanques de baja salinidad la microbiota es semejante a la que se encuentra en el agua de mar. A medida que aumenta la concentración de sales, se observa la aparición de un color rojizo en las aguas, debido a la existencia de poblaciones de archaeas halófilas, *Salinibacter ruber* y *Dunaliella* (Anton *et al.* 1999; Anton *et al.* 2000; Guixa-Boixereu 1996) (Figura 2).

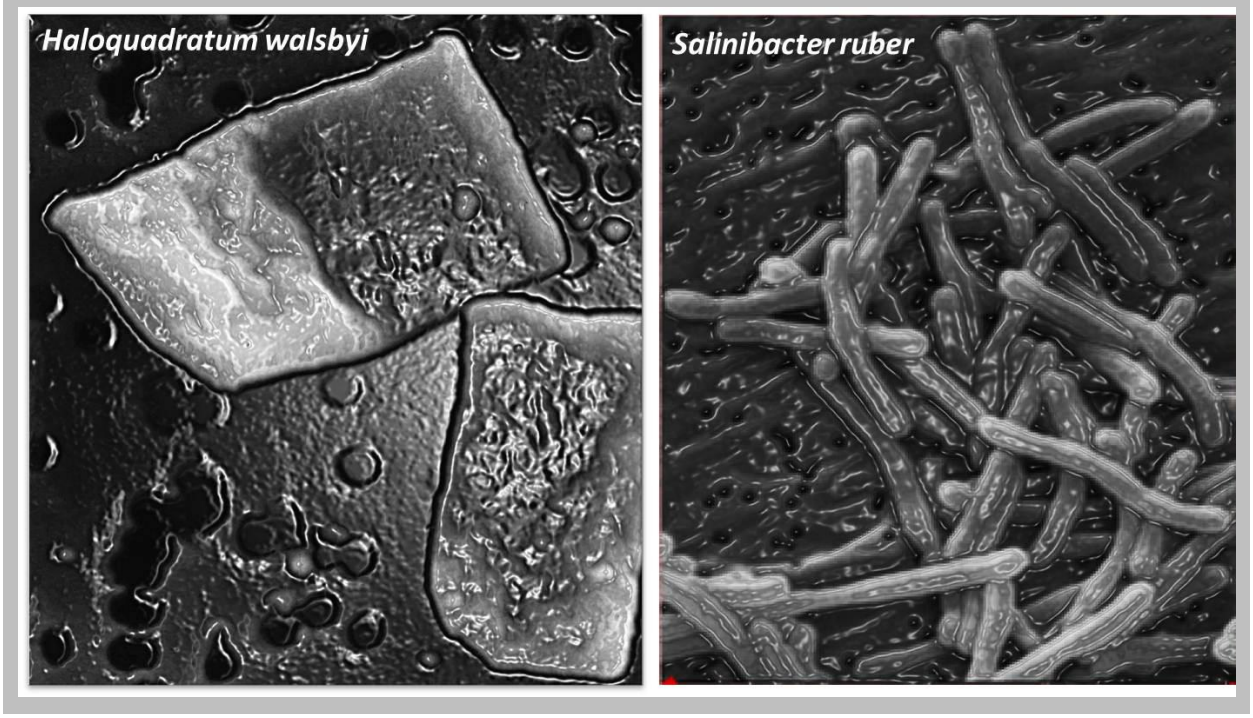
Figura 2. Modelo esquemático de la evaporación ocurrida en una salina solar y microbiota presente en cada dominio.



A concentraciones de cloruro sódico cercanas a la saturación, la mayoría de microorganismos que se encuentran son *Archaea* halófilas extremas, pertenecientes a la familia *Halobacteriaceae* (Anton *et al.* 1999; Maturrano *et al.* 2006; Mutlu *et al.* 2008; Oren 2002a; Rossello-Mora *et al.* 2008) y el pigmento carotenoide que poseen es el encargado de proporcionar la coloración rojiza de dichos estanques. También se encuentran poblaciones de *Bacteria*, tales como *S. ruber (Anton et al. 2000)* y *Salicola marasensis* (Maturrano *et al.* 2006). Generalmente, entre el 70-95% son *Archaea* y entre un 5-30% son *Bacteria (Elevi Bardavid et al. 2008)*. Las morfologías más abundantes observadas en los cristalizadores de las salinas solares de Santa Pola son formas cuadradas y bacilos largos.

A nivel molecular, la mayoría de secuencias obtenidas en los estudios de los cristalizadores se corresponden con *H. walsbyi* (para *Archaea*) y con *S. ruber* (para *Bacteria*), coincidiendo con lo observado por microscopía (Figura 3). Del total de la comunidad que existe en los cristalizadores, la mayoría se puede atribuir a representantes de estos dos grupos. Una parte minoritaria se puede asignar a otras especies de *Halobacteriaceae*, *Proteobacteria* y otros miembros del grupo *Bacteroidetes* (Baati *et al.* 2008; Benlloch *et al.* 2002; Benlloch *et al.* 1995; Estrada *et al.* 2004; Ghai *et al.* 2011). Sin embargo, existen casos en los que estos dos grupos no son los predominantes en la comunidad, como es el caso de unas salinas de Eslovenia, donde no se detectan representantes de *Haloquadratum* (Pasic *et al.* 2005) y en las salinas de Maras (Perú) (Maturrano *et al.* 2006), donde *Salinibacter* no es la bacteria mayoritaria (de hecho, ni siquiera se detecta su presencia mediante técnicas moleculares).

Figura 3. Imágenes de microscopía electrónica de barrido de *H. walsbyi* y *S. ruber.*



Los dos morfotipos más abundantes en el cristalizador CR30 de las salinas solares de Santa Pola (Alicante) se identificaron mediante técnicas moleculares y se aislaron, posteriormente, en cultivo puro. A pesar de que estos ambientes habían sido definidos clásicamente como "cultivos monoespecíficos de *Archaea*" (Guixa-Boixereu 1996), los bacilos alargados se afiliaron a un nuevo género y especie bacterianos: *S. ruber* ((Anton *et al.* 2002).

Las células cuadradas se asignaron al nuevo género y especie *H. walsbyi* (Bolhuis *et al.* 2004; Burns *et al.* 2004), dentro de la familia *Halobacteriaceae* (dominio *Archaea*). De estas dos especies mayoritarias, existen hasta la fecha dos genomas secuenciados de cada especie: *H. walsbyi* DSM 16790 (denominada también HSBQ001, aislado entre 2001-2002), *H. walsbyi* C23 (aislado entre 2002-2004), *S. ruber* M8 y *S. ruber* M31 (Bolhuis *et al.* 2006; Dyall-Smith *et al.* 2011; Mongodin *et al.* 2005; Pena *et al.* 2010). Muchas de las secuencias ambientales del marcador filogenético 16S rRNA obtenidas de cristalizadores se asocian a distintas poblaciones de *H. walsbyi* y *S. ruber*. Dentro de cada grupo, todas las secuencias presentan similitudes de entre un 97 y un 100% (Casamayor *et al.* 2002; Oh *et al.* 2010).

Ya que este trabajo se centra en secuencias que perteneces a la archaea cuadra *H. walsbyi*, se describirán a continuación las características del género y su distribución.

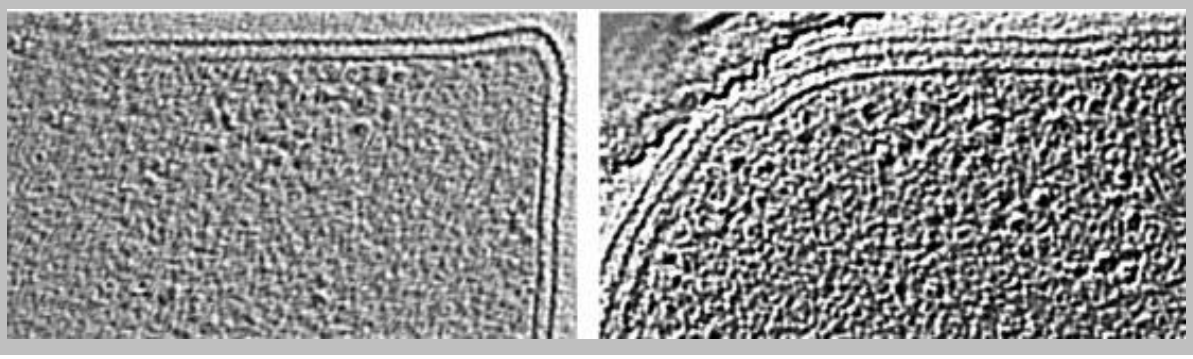
1.2.1. *Haloquadratum walsbyi*. Genomas secuenciados. Características del género y distribución.

Descrita por primera vez en 1980 (*Walsbyi*, 1980), la haloarchaea cuadrada *H. walsbyi*, es cosmopolita y habitualmente es la especie dominante en aguas hipersalinas, tales como lagos salados y los cristalizadores de las salinas (Anton *et al.* 1999; Oren 2002b). Sus células son muy características, siendo cuadrados delgados o rectángulos, con vesículas de gas y gránulos de poli-hidroxibutirato (PHA) (Kessel & Cohen 1982; Stoeckenius 1981). Crece en concentraciones de saturación de sal (requiere concentraciones de sal de al menos 14% w/v, más de cuatro veces que el agua de mar), donde puede llegar a representar un 80% de la población microbiana, y su citoplasma está completamente adaptado para funcionar óptimamente en altos niveles de cloruro potásico. También puede tolerar concentraciones molares de Mg^{2+}, algo que sólo un número muy limitado de organismos son capaces de llegar a hacer por la baja actividad de agua (Bolhuis *et al.* 2006). Una característica fundamental de *H. walsbyi* es que el genoma tiene un contenido de G+C del 48%, un valor considerablemente más bajo que el de todas las otras especies conocidas de la familia Halobacteriaceae, que tienen valores de 61-70% (Oren 2008). *H. walsbyi* es un organismo muy difícil de cultivar en el laboratorio y de crecimiento extremadamente lento. En el 2004 se publicaron paralelamente el aislamiento y cultivo de dos cepas, una de origen español (aislado HSBQ001) y otra australiano (aislado C23) (Bolhuis *et al.* 2004; Burns *et al.* 2004), formalmente se describieron como una nueva especie en el 2007 (Burns *et al.* 2007). Las cepas aisladas poseen secuencias muy similares del gen 16S rRNA y no divergen más del 1.6% (dos nucleótidos diferentes en todo el marcador ribosomal), y una similitud de hibridación cruzada DNA-DNA del 80%. Al microscopio, la diferencia más notable entre los dos aislados fue en la estructura de sus paredes celulares: la cepa HSBQ001 muestra una atípica pared celular de tres capas mientras que C23 poseía únicamente dos (ver Figura 4). En muchas halobacterias, esta capa externa formada por una sola proteína que constituye la denominada capa S (Lechner & Sumper 1987; Mengele & Sumper 1992; Sumper *et al.* 1990).

El aislado español, HSBQ001, fue secuenciado en 2006 (Bolhuis *et al.* 2006), lo que permitió la primera descripción de sus características y metabolismo general (Falb *et al.* 2008). Probablemente debido a la presencia de secuencias repetidas y pseudogenes, su densidad de genes es sólo el 76%, mucho menor que en otros haloarchaea y la mayoría de los procariotas. El aislado australiano, C23, fue secuenciado en el 2011 (Dyall-Smith *et al.* 2011) y, a pesar de la distancia geográfica entre las dos cepas, ambas cepas son mucho menos divergentes

de lo esperado. El ANI entre las dos cepas es del 98.26% y las mayores diferencias se encontraron en las denominadas islas genómicas tal y como se explica en los capítulos siguientes.

Figura 4. Micrografía de microscopía electrónica de la estructura de la capa S de *H. walsbyi* C23 (estructura bilaminar, izquierda) y de la estructura trilaminar presente en HSBQ001 (derecha).



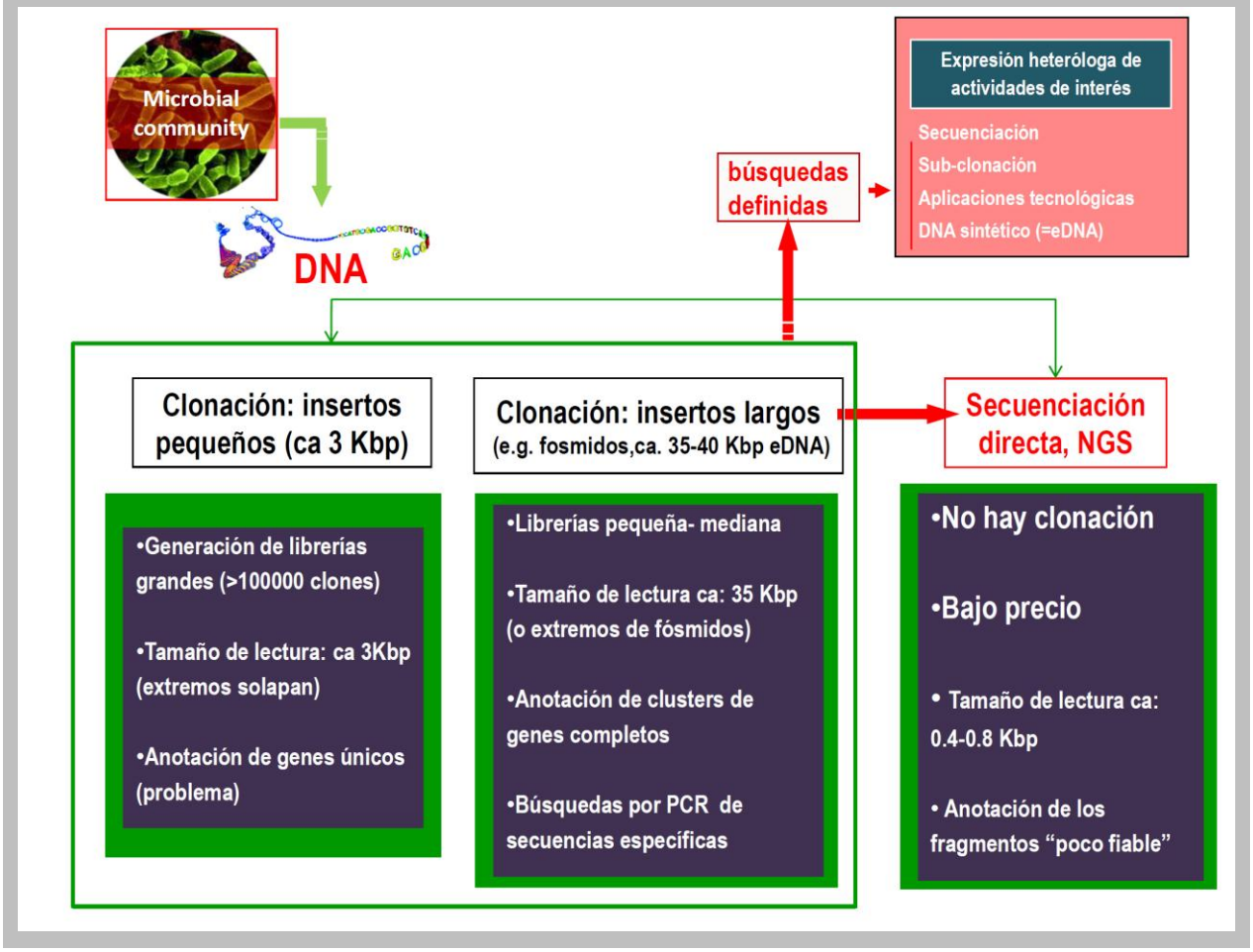
2. PAN-GENOMA, VARIABILIDAD PROCARIÓTICA DENTRO DE UNA ESPECIE

Durante los últimos años se ha producido un gran cambio en el estudio y entendimiento de la dinámica de genomas bacterianos. Tradicionalmente, el principal factor en la adaptación y evolución génica microbiana eran las mutaciones, como ocurre en el caso de metazoos y plantas. A partir de la aparición de la genómica comparativa, este punto de vista evolutivo cambió, introduciéndose la idea de que la mayor variabilidad en los genomas procarióticos se debía a la existencia de grupos de genes que variaban de una cepa a otra y que estaban relacionados con diferencias en la patogenia y otras propiedades específicas de la cepa (Bergthorsson & Ochman 1998). Además, se demostró que los cromosomas bacterianos se encuentran bajo una presión selectiva que modela profundamente la organización de los mismos, al igual que ocurre en los eucariotas. 17 años después de la aparición de la secuencia completa del primer microorganismo de vida libre, las bases de datos contienen 1945 genomas bacterianos completamente secuenciados publicados en Genebank (a fecha 18-03-2012). Sin embargo, en la mayoría de casos sólo se encuentran secuenciados 1 ó 2 genomas por especie, haciendo muy difícil estimar el número de genomas que harían falta para describir completamente una especie bacteriana (existen algunas excepciones, como el caso de *E. coli*, donde existen más de 60 cepas secuenciadas (Lukjancenko *et al.* 2010)).

La exploración de la diversidad microbiana por aproximaciones moleculares llegó a su culmen aproximadamente al finalizar el siglo pasado. En estos momentos, nos encontramos en un renacimiento del estudio de la diversidad microbiana gracias a la metagenómica, aplicando la secuenciación del DNA ambiental a gran escala. En el esquema presentado en la Figura 5, es posible encontrar varias de las vías seguidas en metagenómica: la construcción de librerías metagenómicas mediante clonación del DNA ambiental en plásmidos, fósmidos o BACs y secuenciarlos posteriormente o, secuenciar directamente el DNA extraído de la biomasa que se quiera estudiar. Obviamente, la ventaja de la primera metodología es la disposición "física" de la secuencia nucleotídica, y la gran desventaja es que sesgo que presenta todo proceso de clonación, por lo que estas libererías pueden no ser

representativas de la biodiversidad real existente (ver por ejemplo (Ghai, 2010 #680), donde se comparan los resultados de secuenciación de una librería metagenómica frente a un metagenoma pirosecuenciado directamente de una misma muestra del Mediterráneo).

Figura 5. Esquema de las vías seguidas en metagenómica: construcción de librerías metagenómicas o secuenciación directa. Una de las ventajas de la construcción de librerías metagenómicas es la posibilidad de realizar búsquedas definidas de genes o la expresión heteróloga de actividades de interés tecnológico.



Sabemos desde hace años que los genomas son entidades extraordinariamente plásticas con un enorme potencial de cambio. Al tener acceso a cientos de genomas de aislados individuales se ha podido obtener una instantánea de cómo es el "pan-genoma" estático de una especie. Los genomas de bacterias y arqueas varían ampliamente dentro de los límites de una especie bien definida, lo que ha llevado al concepto del "pan-genoma" de la especie (Tettelin *et al.* 2005) para referirse al gran reservorio genético presente en diferentes cepas o linajes. En el año 2005, Tettelin *et al.* secuenciaron ocho genomas de aislados de *Streptococcus agalactiae* concluyendo que en teoría, nunca se podrían definir completamente las especies bacterianas, ya que con cada nuevo genoma secuenciado de la especie estudiada, se añaden más genes al repertorio génico de la misma (Figura 6).

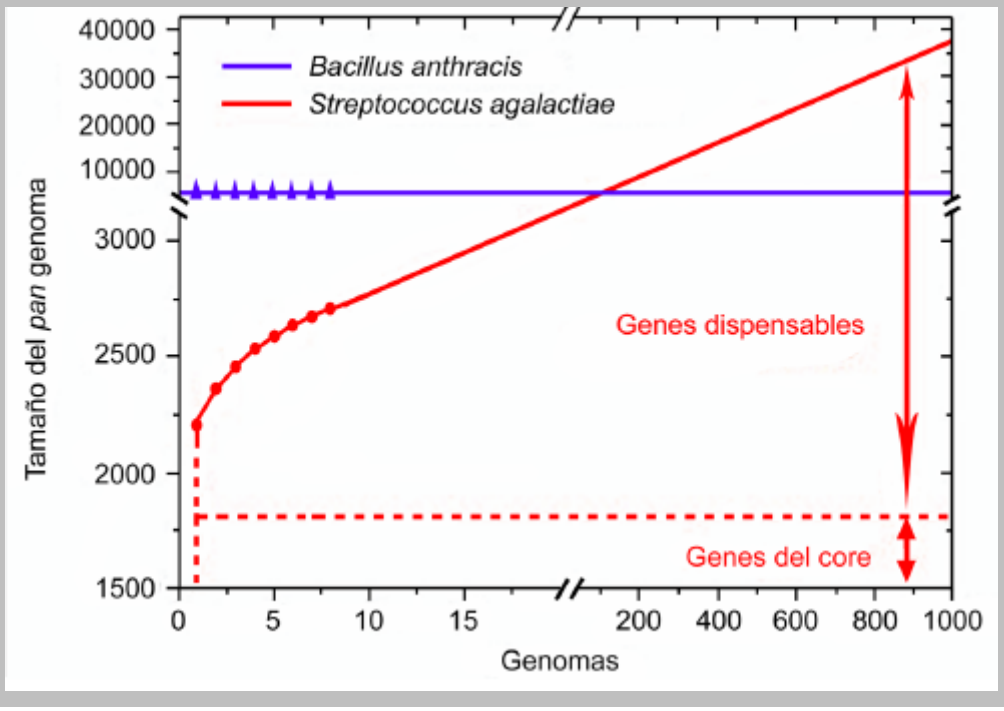
En general, el pan-genoma, definido como el repertorio génico global de una especie bacteriana, se puede dividir en 3 partes:

i) El core-genoma, que estaría formado por el conjunto de genes presentes en todos los miembros de una especie y que representa una estimación mínima del repertorio génico del ancestro común a todos ellos. Este core-genoma incluiría todos los genes responsables de los aspectos biológicos básicos de una especie y de los rasgos fenotípicos principales.

ii) El genoma adaptativo o flexible, que englobaría al conjunto de genes presente en algunas cepas de la especie pero no en todas. Este conjunto de genes contribuiría a la diversidad de las especies y podría contener genes implicados en rutas metabólicas complementarias y con funciones no esenciales para el crecimiento bacteriano pero que confieren ventajas selectivas al microorganismo que los tenga, como pueden ser la adaptación a diferentes nichos, resistencias a antibióticos o colonización de nuevos huéspedes.

iii) Los genes específicos de cepa, que son aquellos que estarían presentes en una única cepa de la especie y tendrían características similares a los genes dispensables.

Figura 6. Tamaño del pan-genoma en función del número genomas secuenciados. Las curvas muestran los valores extrapolados del pan-genoma a un mayor número de cepas secuenciadas. El tamaño del pan-genoma de las especies crece con el número de cepas secuenciadas (pan-genoma abierto, *S. agalactiae*) o se satura rápidamente (pan-genoma cerrado, *B. anthracis*). Después de secuenciar un elevado número de cepas, el número de genes dispensables en un pan-genoma abierto es mucho mayor que el tamaño del core-genoma (adaptada de Medini *et al.*, 2005).



En especies con el pan-genoma abierto cada nueva cepa secuenciada contribuye con un número determinado de genes, con lo que el número de genes específicos de cepa aumenta constantemente. Esto significaría por tanto que el pan-genoma podría tener un tamaño ilimitado (Bentley 2009). Por otro lado, las especies que contienen el pan-genoma cerrado, es decir, que no aumenta de tamaño a medida que se añaden más genomas secuenciados, como es el caso de *Bacillus anthracis,* sólo haría falta un número pequeño de cepas para explorar su diversidad génica.

2.1. VARIABILIDAD GENÉTICA DE LOS GENOMAS PROCARIÓTICOS.

Sin embargo, la dinámica de los procesos que han originado tal diversidad de genomas es más difícil de abordar. Una aproximación que está arrojando nuevo datos sobre la dinámica evolutiva de los genomas procarióticos es la metagenómica. La secuenciación de DNA extraído directamente de comunidades naturales (en lugar de cultivos), incluso cuando provienen de un solo punto en el tiempo, constituye el equivalente a una película en la que las etapas que conducen de un linaje a otro pueden analizarse. Uno de los mayores retos para entender la diversidad procariótica y su evolución es explicar la diversidad de genomas que pueden encontrarse entre una sola especie (u OTU, unidad taxonómica funcional). Hasta la fecha, los estudios llevados a cabo se concentran fundamentalmente en la secuenciación de aislados de patógenos y, en raras ocasiones, organismos no patógenos relacionados (por ejemplo: (Dempsey *et al.* 2006; Hochhut *et al.* 2006; Lukjancenko *et al.* 2010; Muzzi & Donati 2011; Petrosino *et al.* 2006; Tettelin *et al.* 2005; Willenbrock *et al.* 2006). Estos estudios han puesto de manifiesto la gran variabilidad que puede existir entre los genomas de múltiples aislados de una misma especie, con diferencias de tamaño, repertorio genético y sintenia. Sin embargo, esta aproximación tiene sus contras ya que las cepas secuenciadas deben representar la diversidad real de la especie, y esto difícilmente es garantizable ya que, por ejemplo, los aislados clínicos son representativos de linajes muy virulentos seleccionados por el sistema de defensa del hospedador o por resistencia a antibióticos. De manera similar, los aislados de vida libre son seleccionados por su adaptación a crecer en el laboratorio o por producir colonias visibles, luego el sesgo introducido por la metodología de cultivo es a menudo impredecible y no se debe obviar.

Un acercamiento alternativo para estudiar el pan-genoma sin este sesgo es mediante metagenómica, ya que el metagenoma de un hábitat contiene secuencias de diferentes linajes de especies y esta información puede ser usada para inferir la diversidad genómica entre los mismos. Uno de los requerimientos, por tanto, es tener al menos un genoma de una cepa de referencia para identificar fragmentos metagenómicos como pertenecientes a la especie. También, se requiere que la especie sea muy abundante en el hábitat elegido. De manera general, los ambientes extremos se caracterizan por tener una comunidad microbiana simple, luego pueden servir de excelente modelo para el desarrollo de modelos ecológicos microbianos y posteriormente, extrapolar los resultados a ecosistemas más complejos. Hasta la fecha, la complejidad de la comunidad microbiana se ha calculado en términos de diversidad de marcadores filogenéticos tales como el 16S rRNA. Sin embargo, esta aproximación no puede estimar la extensión de la variabilidad intragenómica dentro de la misma especie. De hecho, secuencias de genomas completos de muchas cepas de una misma especie (normalmente organismos patógenos cultivables) a menudo revelan un amplio contenido génico diferente (por ejemplo, (Lukjancenko *et al.* 2010). El genoma "adaptativo" presente en sólo unas cuantas cepas de una especie no parece ser esencial para la supervivencia celular, pero provee la posibilidad de adaptarse a fluctuaciones ambientales, como por ejemplo, el uso de diferentes nutrientes. El conocimiento sobre estos genes accesorios (tamaño y composición) serán esenciales en áreas desde la Ecología (cuantos fenotipos diferentes pueden existir dentro de una misma especie) a la Biotecnología (diversidad genética para la explotación tecnológica).

2.1.1. Islas Metagenómicas.

Las islas genómicas se han definido como elementos similares a las islas de patogenicidad presentes en microorganismos patógenos (Dobrindt *et al.* 2004), sin embargo, su naturlaza es muy diferente. Tienen en común el hecho de ser regiones largas del cromosoma que forman parte del conjunto de genes accesorios, es decir, genes que no están presentes en todos los representantes de la especie y reflejan una adaptación específica al ambiente en el que habita el microorganismo. Muchas de las islas de patogenicidad representan regiones del genoma que han sido previamente transferidas por un elemento genético móvil o por mecanismos de transferencia horizontal (Kettler *et al.* 2007). Además, estas islas suelen estar asociadas a tRNAs y estar flanqueadas por estructuras repetidas (Dobrindt *et al.* 2004). Sin embargo, las islas genómicas de bacterias libres (no patógenas) no siempre comparten estas características.

En los genomas de muchos genomas de bacterias de vida libre existen también "islas genómicas" ricas en genes únicos no compartidos por el resto de las cepas aisladas ni tampoco por el resto de las cepas o linajes presentes en el metagenoma (Coleman *et al.* 2006; Kettler *et al.* 2007; Rodriguez-Valera *et al.* 2009; Ting *et al.* 2002). Coleman *et al.* (2006) utilizó este enfoque para estudiar la diversidad genómica en *Prochlorococcus marinus*, una cianobacteria marina predominante en aguas oceánicas oligotróficas. Mediante la comparación del genoma de una cepa de referencia frente al metagenoma del Mar de los Sargazos (Venter *et al.* 2004), se identificaron regiones bien definidas del genoma de referencia que tenían secuencias de muy poca o ninguna homología en el metagenoma. Estas regiones, llamadas islas genómicas (GI), se consideran hipervariables y de hecho pueden ser únicas para la cepa de referencia. Muchos de los genes de la isla metagenómica estaban relacionados con el estrés de nutrientes y la adaptación distinta intensidad de la luz. Estos linajes pueden coexistir por especialización en micro-nichos gracias a ese contenido diferencial de genes (por ejemplo mediante la utilización heterótrofa de diferentes compuestos orgánicos). Curiosamente, en algunas de las islas metagenómicas se encuentran genes que codifican proteínas que participan de alguna manera en estructuras celulares extracelulares (como el exopolisacárido, pilis y componentes flagelares), componentes de la superficie celular (por ejemplo, glicoproteínas (CSGs)), proteínas de gran tamaño que son probablemente extracelulares y factores implicados en glicosilación de componentes de la superficie (Avrani *et al.* 2011; Coleman *et al.* 2006; Wilhelm *et al.* 2007). Todos estos genes son posibles sitios de reconocimiento de fagos, lo que podría indicar que esta esta variabilidad tendría un papel en la prevención del reconocimiento de los fagos (Rodriguez-Valera *et al.* 2009). Por tanto, los fagos pueden desempeñar un papel fundamental como garantes de la micro-diversidad necesaria para aprovechar los recursos ecológicos de manera eficiente (Thingstad 2000).

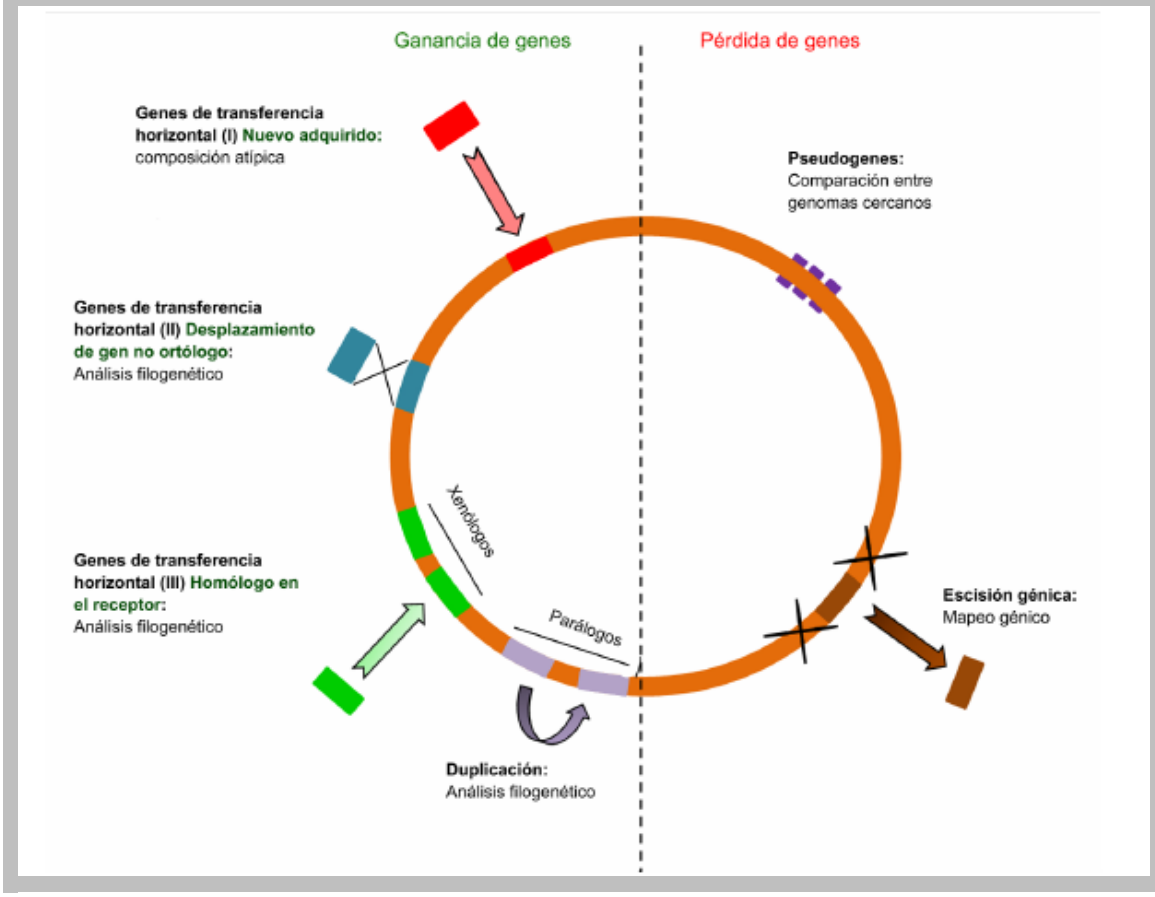
2.1.2. Mecanismos de variabilidad genética.

Los procesos por los cuales se ganan o pierden genes en un genoma, contribuyendo a la diversidad del repertorio génico de un microorganismo, han sido ampliamente estudiados (Abby & Daubin 2007). Los procariotas poseen genomas altamente dinámicos que adquieren, pierden y reordenan información génica relevante (Figura 7), lo que hace que sean ecológica y fenotípicamente muy diversos (Dutta & Pan 2002). Para explicar esta gran diversidad génica presente en los procariotas, se postulan dos mecanismos:

i) modificaciones internas de la información genética, es decir, variaciones que derivan de la divergencia (Milkman 1997). Las mutaciones puntuales producidas en los genes pre-existentes de un organismo pueden llevar a la modificación lenta pero continua de los mismos, permitiendo así una expansión gradual de su nicho. Pueden ser responsables de la diversificación y la especiación de los microorganismos en una escala evolutiva a largo plazo (Dutta & Pan 2002).

ii) pérdida o adquisición de un conjunto de genes de otras especies por procesos de transferencia horizontal de genes (HGT) (Syvanen 1998). Un genoma puede perder genes o bien por escisión, cuando el gen desaparece completamente del genoma del microorganismo, o por formación de pseudogenes (resto de un gen funcional antiguo que ya no es funcional), lo que ocurre cuando se acumulan mutaciones (puntuales y/o inserciones o deleciones) dando como resultado la pérdida de función de esos genes.

Figura 7. Dinámica y técnicas de detección de los mecanismos por los que se produce la variabilidad en el repertorio génico de un microorganismo (adaptada de Abby y Daubin, 2007).



La constante pérdida de genes y/o función de los mismos en un genoma puede estar compensada por la adquisición de nuevo material genético. En procariotas se pueden producir duplicaciones y pérdida de genes pre-existentes (Lerat *et al.* 2005) o se pueden integrar genes con procedencias distintas dentro de un genoma mediante HGT. La transferencia horizontal hace referencia a la adquisición de genes foráneos por parte de un organismo y parece ser un mecanismo importante en la especiación y adaptación de los procariotas a nuevos

ambientes (Ochman 2001). Los procesos de HGT pueden resultar en una alteración a gran escala de la estructura y organización de los genomas y, por tanto, serían capaces de generar nuevas variantes de cepas procariotas. Este fenómeno es bastante frecuente entre procariotas, especialmente como respuesta a cambios en el ambiente, y hace que la diversidad génica aumente proporcionando acceso a nuevos genes además de los heredados (Ochman *et al.* 2000; Syvanen 1998). La mayoría de genomas de microorganismos secuenciados muestran algún ejemplo de posibles casos de HGT (Ragan 2001). En estudios realizados con genomas completos bacterianos y de archaeas, se ha detectado que el porcentaje de genes de HGT varía entre 1,56 y 14,47 % del total de genes (Garcia-Vallve *et al.* 2000).

Para llevar a cabo una transferencia efectiva del material genético entre especies, se necesitan tres pasos: la cesión del gen desde el donador hasta el receptor, la incorporación de la nueva secuencia en el genoma del receptor y, finalmente, la expresión del gen. Los procariotas han desarrollado tres mecanismos principales para la realización de esta transferencia de material genético: transducción (proceso de transferencia genética mediada por bacteriófagos que contienen el DNA de la célula donadora, conjugación (transmisión del DNA mediada por un plásmido o mediante transposones conjugativos) y transformación (DNA desnudo se incorpora a la célula receptora). Una vez que un gen se ha transferido de una célula donador a una receptora, tienen lugar la estabilización del nuevo material genético en el receptor. La integración del nuevo gen en el cromosoma del receptor puede producirse por recombinación homóloga mediada por la proteína *recA*, o por recombinación no homóloga que puede ser específica de sitio (requiere la existencia de pequeñas regiones de homología entre ambas moléculas de DNA y que está mediada por una integrasa) o ilegítima (que no requiere homología entre las moléculas de DNA y que está mediada por transposasas). Los genes transferidos se replican como parte del genoma del receptor y se transmiten a la descendencia de manera estable a lo largo de generaciones sucesivas.

De manera general, existen dos clases de métodos de detección de genes de transferencia horizontal: los métodos filogenéticos basados en el análisis filogenético de genes o proteínas individuales (Tamames & Moya 2008) y los basados en el análisis de la composición de nucleótidos. Estos últimos están basados en el análisis de la composición de DNA y en la asunción de que un genoma individual tiene características propias asociadas a la composición de su DNA (Jordan & Koonin 2004). En general, estas aproximaciones se basan en la idea de que cada genoma tiene un patrón característico de contenido en G+C, uso de codones y frecuencia de di-, tri- y tretranucleótidos. Los genes que difieren significativamente del resto del genoma en estos parámetros se consideran posibles genes de transferencia horizontal. Sin embargo, la composición anómala de nucleótidos sólo se puede aplicar a fenómenos de transferencia reciente puesto que cuando un gen se incorpora a un genoma nuevo se produce un proceso conocido como proceso de mejora (Lawrence & Hendrickson 2005), mediante el cual este gen ajusta su composición en bases y uso de codones a los de la célula hospedadora. Los genes que estén bajo el proceso de mejora mostrarán características intermedias entre las originales (características del organismo donante) y las de su nuevo entorno (organismo receptor).

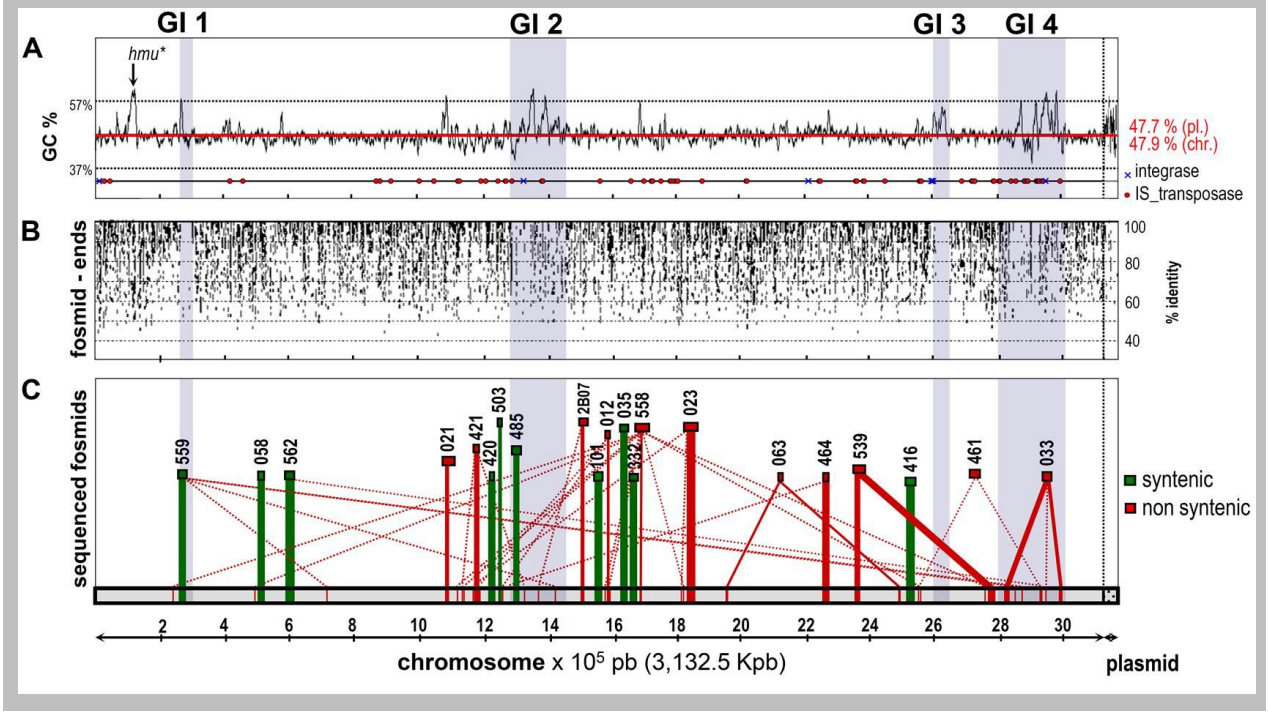
2.2. VARIABILIDAD GENÓMICA en el CRISTALIZADOR de las SALINAS SOLARES. Islas Genómicas de *H. walsbyi* DSM 16790.

El cristalizador CR30 de la salina solar de Santa Pola (Alicante), donde se alcanza la saturación de NaCl, por su simplicidad y por la cantidad de datos experimentales que se poseen (Anton *et al.* 1999; Benlloch *et al.* 2001; Benlloch *et al.* 2002; Benlloch *et al.* 1995), es uno de los pocos hábitats donde se puede estudiar mejor la variabilidad genómica de una especie. Las aguas hipersalinas con más del 30% w/v de sal suelen tener altas concentraciones de células procariotas (10^7 ml^{-1}), y las células de *Haloquadratum* pueden llegar a alcanzar hasta un 80% de la comunidad total del cristalizador. Además, se han descritos los niveles más altos de partículas de virus descritos hasta la fecha: 10^9 ml^{-1} (hay pocos o ningún ciliados y, tampoco flagelados depredadores (Pedros-Alio *et al.* 2000)). La diversidad a nivel mundial de este organismo ha sido examinada utilizando secuencias del gen 16S rRNA, y estos datos indicaron que las poblaciones de *H. walsbyi* son altamente coherentes (Oh *et al.* 2010), con un nivel de variación menor del 2%, un valor muy cercano a la divergencia del 1,6% observado entre los dos genomas genomas de *H. walsbyi* (Bolhuis *et al.* 2006; Dyall-Smith *et al.* 2011). No sólo fue *H. walsbyi* el grupo dominante microbiana en estos sitios, sino que también parecía ser la única especie de este género, siendo común encontrar secuencias idénticas en sitios distantes geográficamente. Esta imagen de conservación global contrasta con otros géneros también frecuentes de haloarchaea, tales como *Halorubrum* y *Haloarcula*, los cuales muestran niveles mucho más altos y mayor divergencia en los genes rRNA (7%).

En el momento en que este trabajo se realizó, además de los dos genomas de *H. walsbyi*, también se disponía de una librería metagenómica de fósmidos construida a partir de biomasa recogida en el cristalizador de la salina donde la cepa *H. walsbyi* DSM 16790 fue aislada. De esta librería, se secuenciaron aproximadamente los extremos de unos 1500 fósmidos, dando a conocer la existencia por primera vez las cuatro islas metagenómicas existentes en esta cepa (definidas como regiones genómicas de más de 20 Kpb con un reclutamiento de secuencias ambientales significativamente menor) (Cuadros-Orellana *et al.* 2007; Legault *et al.* 2006) (Figura 8B). Posteriormente, se secuenciaron aleatoriamente 23 fósmidos completos de esta librería (Figura 8C) (Cuadros-Orellana *et al.* 2007). Más tarde, en el 2011, se publicó un segundo metagenoma de biomasa del CR30, SS37, esta vez obtenido mediante pirosecuenciación (Ghai *et al.* 2011). Éste último ha sido el usado en este trabajo, permitiendo delimitar mucho mejor las islas metagenómicas de *H. walsbyi* DSM 16790 y C23.

En efecto, la clonación y secuenciación del DNA ambiental procedente de las poblaciones naturales de *Haloquadratum* demostró que el pan-genoma de la especie contendría, al menos, otro genoma equivalente (3 Mpb) al genoma de la cepa secuenciada (Legault *et al.* 2006). Tras el análisis de las secuencias de *Haloquadratum* que no formarían parte del core-genoma de la especie (fragmentos genómicos comunes en todas las cepas de la especie) se comprobó que las diferencias encontradas tienen implicaciones eco-fisiológicas para *H. walsbyi* y se postuló que las diferentes cepas o linajes estarían especializadas en la explotación de diferentes compuestos orgánicos, evitando la competencia directa por el recurso o sustrato natural. Muchos de estos genes, específicos de los diferentes linajes y localizados en regiones hipervariables del genoma o islas genómicas, estarían relacionadas con proteínas de superficie que darían a las distintas cepas susceptibilidades diferentes al ataque de virus.

Figura 8. Islas genómicas de *H. walsbyi* DSM 16790 (Cuadros-Orellana *et al.* 2007). (A) Contenido G+C a lo largo del genoma. (B) Reclutamiento genómico usando los extremos de los fósmidos. (C) Representación de los fósmidos secuenciados donde se indica la sintenia (verde), o la ausencia de ella (rojo), respecto del genoma de referencia.



La imagen de la enorme diversidad y plasticidad presente en el cristalizador de la salina, un sistema casi mono-específico, contrasta con la simplicidad existente en otros ambientes extremos como el biofilm de una mina ácida analizado también mediante metagenómica (Tyson *et al.* 2004). En el biofilm, la diversidad genética de los quimiolitotrofos presentes ha demostrado ser bastante restringida incluso a nivel de sustituciones de nucleótidos. En contraste, los organismos dominantes en las salmueras saturadas son heterótrofos, y requieren una amplia variedad de carbono y otros nutrientes (durante el día las haloarchaeas puede obtener energía a partir de la rodopsina, pero no pueden fijar CO_2 o N_2). Estos nutrientes se obtienen de un conjunto muy diverso de compuestos orgánicos liberados por el alga *Dunaliella* sp., y otros microbios presentes a bajas salinidades (Gasol *et al.* 2004; Pedros-Alio *et al.* 2000). Por lo tanto, los resultados sugieren que diferentes clones o linajes dentro de *H. walsbyi* se especializan en la explotación de los diferentes compuestos orgánicos y todos ellos conviven en un conjunto diverso de recursos.

2.2.1. Isla genómica I de *H. walsbyi.*

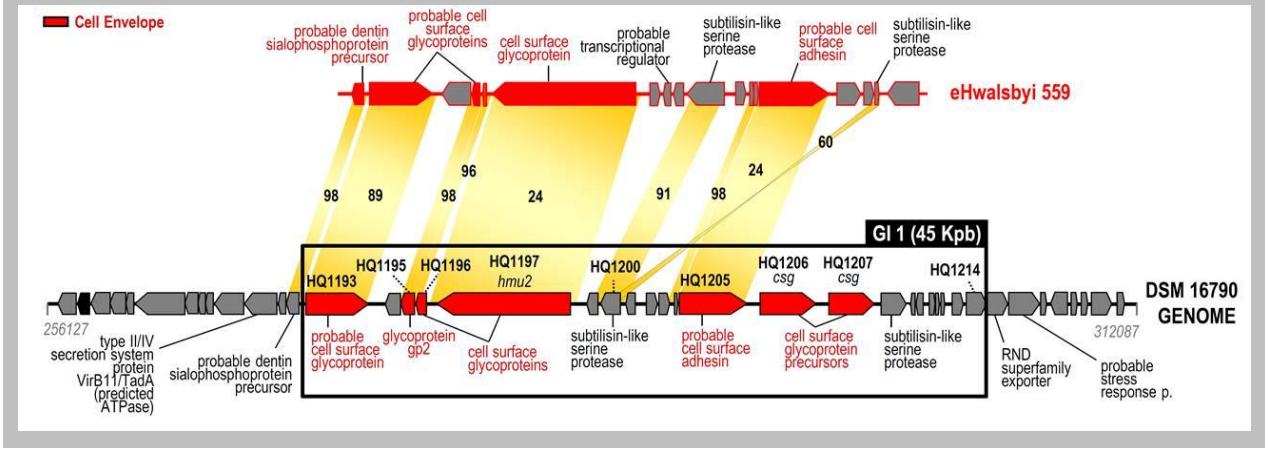
La isla genómica 1, (GI1), está situada entre los nucleótidos 257397-302834 (45,4 Kpb) del genoma de la cepa HSBQ001 (Figura 8 y 9). Esta isla es atípica en comparación con otras islas existentes en el genoma por dos razones importantes: en primer lugar, el contenido promedio de GC (47,86%) es similar a la media del genoma (47,90%) y en segundo lugar, no posee elementos IS o genes putativos de fagos (HQ_1109, la transposasa en el interior GI1 no es funcional). La característica principal de GI1 es poseer genes que codifican gran parte de las glucoproteínas que tiene el genoma, las cuales están glicosiladas en su mayoría (CSG, Cell Surface glycoprotein) (Schaffer *et al.* 2001; Schaffer & Messner 2001). Se ha sugerido que alguna de las cuales podrían ser los

componentes de la capa S, la cual proporciona rigidez a la envoltura celular de haloarchaea (Mengele & Sumper 1992; Schaffer & Messner 2001). La comparación con otras proteínas que forman la capa S de otras haloarchaeas, como las de *Halobacterium salinarum* y *Haloarcula marismortui* ATCC 43049, permitió identificar a la ORF HQ_1207 como la posible candidata para desempeñar un papel similar en *H. walsbyi* (Blaurock *et al.* 1976; Trachtenberg *et al.* 2000). A diferencia de otros genomas de haloarchaeas secuenciadas, donde el número de CSGs es dos o menor, *H. walsbyi* DSM 16790 tiene un total de 18. Esta gran diferencia se ha atribuido a la arquitectura compleja asociada con la morfología de las células cuadradas (Walsby 2005). Sin embargo, una explicación alternativa estaría relacionada con la alta densidad de población que este organismo alcanza en el cristalizador (hasta 10^8 células/ml). Esta alta densidad de población la hace un blanco ideal para la depredación de fagos con consecuencias drásticas si todas las células de *H. walsbyi* compartiesen las mismas CSGs y ésta pudiese ser reconocida por un tipo de fago. De esta manera, esta variabilidad estaría protegiendo a *H. waslbyi* de alguna manera. (Esta idea se desarrolla en el apartado de Discusión).

2.2.1.1. Variabilidad de la isla genómica I de *H. walsbyi.*

La primera vez que se observó la variabilidad existente en la IG1 fue a través de la secuenciación de los 23 fósmidos publicados en Cuadros-Orellana *et al.* (2007). Por azar, uno de los fósmidos secuenciados, eHw-559, mapeaba en esta zona del genoma de *H. walsbyi* DSM 16790, la cual contenía genes necesarios para la formación de la envuelta celular.

Figura 9. Variabilidad de la isla genómica I de *H. walsbyi* DSM 16790 comparada con eHw-559 (Cuadros-Orellana *et al.* 2007).



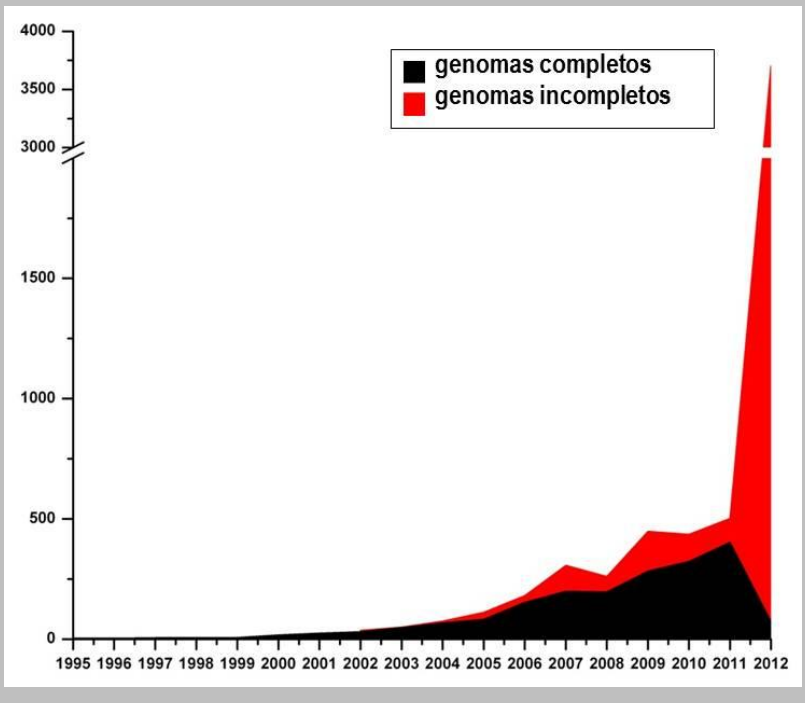
La hipótesis anterior está apoyada por la secuencia que se encuentra en eHw-559 y cuya comparación con el genoma de *H. waslbyi* DSM 16790 se encuentra en la Figura 9. Este fósmido tiene un conjunto alternativo de CSGs con dos regiones principales de variación: i) uno de ellos es el gen *hmu2* gen (HQ_1197) que codifica para una CSG con la similitud con el gen halomucin *hmu* (HQ_1081) y podría ser parte de la capa de envoltura celular externa (Bolhuis *et al.* 2006). ii) La otra ORF es HQ_1205, anotada como putativa adhesina y con las características de una glicoproteína secretada. La identidad entre las copias de las proteínas Hmu2 y HQ_1205 y sus homólogos ambientales no sobrepasa el 24% de similitud.

En la isla GI4, casi simétrica con respecto al origen de replicación de GI1, se encuentras tres copias parálogas de Hmu2. La recombinación entre estas dos zonas del genoma podría explicar el alto nivel de variación que se encuentra en estos genes, sin embargo esto no se ha podido comprobar.

3. NUEVAS METODOLOGÍAS DE SECUENCIACIÓN

La primera generación de técnicas para secuenciar DNA nació en 1975 con la metodología de Sanger y Coulson, la cual necesitaba clonar cada lectura inicial para producir un DNA de cadena simple. En 1977, Maxam y Gilbert publicaron la metodología de secuenciación de DNA mediante degradación química, basado en la modificación química y posterior rotura del DNA. El mismo año, Sanger publicó el método de secuenciación de DNA por síntesis química, que estableció un nuevo estándar para los próximos 30 años. Actualmente, el método de Sanger es utilizado en la mayoría de los laboratorios del mundo. Este escenario ha cambiado con el desarrollo de nuevas metodologías de secuenciación masiva (Mardis 2008a; Mardis 2008b). El aumento de la capacidad de secuenciación de los centros especializados, junto con la reducción de los costes de la misma y la aparición de estas nuevas tecnologías, ha permitido la publicación de la secuencia completa de 1925 genomas de organismos hasta Marzo 2012 y a 5230 proyectos de secuenciación de genomas en marcha (Figura 10). En estos momentos, existen muchos proyectos para secuenciar un gran número de genomas procarióticos, uno de los más ambiciosos puede encontrarse en: http://ldl.genomics.cn/page/M-research.jsp.

Figura 10. Aumento de genomas secuenciados o en proyecto de Secuenciación. (Datos tomados de Genebank, 18/03/2012, http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi)



Todos estos cambios se han producido a partir del 2003/4 gracias a la comercialización de nuevos métodos de secuenciación como son:

- La pirosecuenciación
- La secuenciación tipo Solexa (y en menor medida, Solid)
- La denominada "single-cell genomics".

Ya que el presente trabajo se basa en el estudio de varios fragmentos de DNA secuenciados mediante una de estas nuevas tecnologías, la pirosecuenciación, paso a describir los aspectos más importantes de las mismas.

3.1. Métodos de Secuenciación de Segunda Generación.

Los métodos de secuenciación de segunda generación se basan en la inmovilización del DNA directamente en soportes sin necesidad de clonarlo previamente. Tres plataformas de segunda generación han sido comercializadas por el momento: la Pirosecuenciación 454 desarrollada por Roche (454, Bradford, CT, USA; Roche Applied Science), Illumina/Solexa Genome Analyzer (Illumina, Inc.) y el sistema SOLID™ desarrollado por Applied Biosystems (Applied Biosystems, Foster City, CA, USA). Estas plataformas han aumentado la efectividad de la secuenciación del DNA órdenes de magnitud y han permitido generar lecturas de gigabases en un solo experimento (denominado *"run"*). Estos sistemas incluyen sistemas complejos de enzimas, química, ópticas de alta resolución además de "*hardware"* y "*software"* para el análisis de resultados. Difieren de los métodos tradicionales de secuenciación fundamentalmente en dos aspectos:

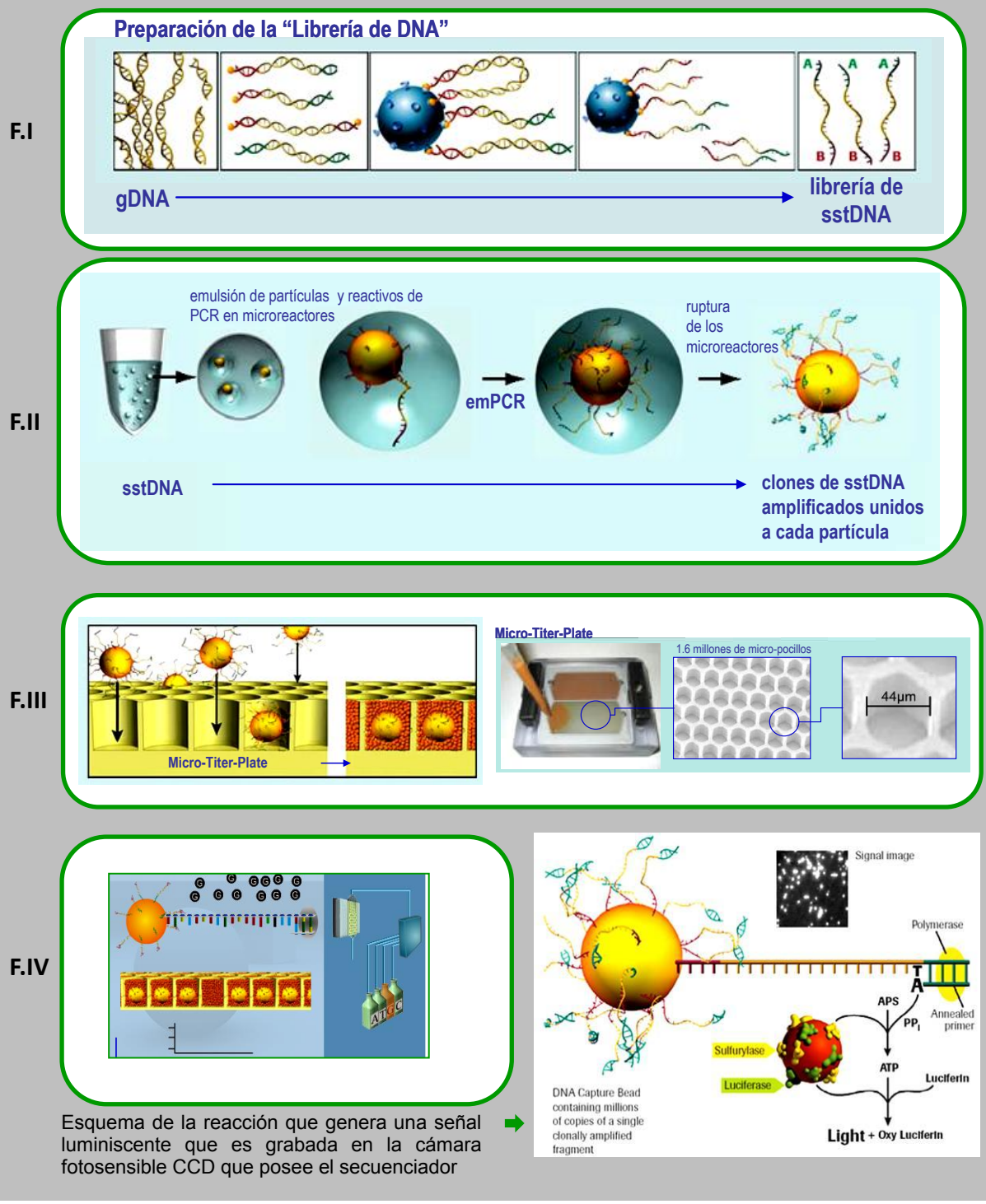
i) Primero, en vez de hacer una secuenciación de clones de DNA de algunos individuos (por ejemplo, 96 secuencias paralelas en un secuenciador capilar Sanger), cientos de miles (sistema 454) o miles de millones (Solexa y SOLiD) de moléculas de DNA son secuenciadas en paralelo, usando volúmenes de reacción menores.

ii) Segundo, la secuencias obtenidas son generalmente más cortas que las generados por secuenciación tradicional: 100-150 pb para las tecnologías Solexa y 50 pb SOLiD, aunque pueden alcanzar actualmente los 800-1000 nucleótidos para el sistema FLX plus de 454.

3.1.1. Pirosecuenciación.

El secuenciador 454 (454 Life Sciences), está basado en procesos de emulsión, secuenciación por síntesis y pirosecuenciación. Esta tecnología fue publicada en el 2005 (Margulies *et al.* 2005), y comprada por Roche Diagnostics en el 2007. La versión más actual del mismo es el "Genome Sequencer FLX Plus System" (Roche Applied Sciences) https://www.roche-appliedscience.com/sis/sequencing/index.jsp. La tecnología 454 empezó leyendo 100 pb, después de 16 meses podía leer 250 pb, más tarde, 400 pb y a partir de la segunda mitad del 2011, fragmentos de hasta 800 pb. En la Figura 11, se muestra un esquema de los pasos seguidos en este tipo de secuenciación.

Figura 11. Pasos seguidos en la pirosecuenciación 454 (Roche).



Esquema de la reacción que genera una señal luminiscente que es grabada en la cámara fotosensible CCD que posee el secuenciador

Fase I: Preparación del DNA. El objeto es generar una cadena de DNA de una sola hebra unida a los adaptadores A y B, denominada sstDNA. El proceso de preparación comprende una serie de pasos enzimáticos para producir una cadena de DNA de cadena sencilla (sstDNA) que tenga incorporada los iniciadores, o *primers,* y los adaptadores de unión. Por ejemplo, cuando se trata de DNA genómico (DNAg), éste es fraccionado en trozos de menor tamaño (300-800 pb) que son subsecuentemente truncadas en los extremos para hacerlos romos. En estos extremos se ligarán unos adaptadores cortos, denominados en la figura A y B. Estos adaptadores son las secuencias iniciadoras o "*primers*", para la amplificación mediante PCR de la Fase II (emPCR). También serán

útiles para la secuenciación de fragmentos de librerías de muestras. Cuando partimos de DNA de bajo peso molecular, éste es usado directamente sin fragmentar y la preparación de la muestra comienza con la ligación de los adaptadores A y B. La librería de sstDNA producida al final de esta etapa es analizada para conocer tanto su calidad como la cantidad óptima (número de cadenas de DNA copia por partícula) necesaria para la emPCR, el paso siguiente.

Fase II: PCR en emulsión. En esta etapa tiene lugar la inmovilización del DNA en partículas microscópicas y la producción de clones de la hebra sstDNA unidas a la misma microesfera. Cada una de ellas transporta un único fragmento de sstDNA de la librería genómica. El conjunto global de estas partículas es emulsionado con los reactivos usados en la reacción de amplificación de DNA (o comúnmente llamada PCR) en una mezcla que contiene básicamente agua y aceite. Posteriormente, cada una de estas partículas es separada en un micro-reactor donde se lleva a cabo la reacción de amplificación de DNA. Esta amplificación es llevada a cabo en un conjunto, por lo que el resultado final son cientos de miles de partículas en las que cada una de ellas posee unida miles de fragmentos iguales entre sí de DNA y diferentes entre una partícula y otra.

Fase III: En esta fase tiene lugar la inclusión de cada una de las partículas preparadas en la Fase II en una matriz sólida llamada "*PicoTiterPlate"* y la secuenciación en paralelo a la síntesis de la cadena complementaria de las sstDNA. Ayudado con un sistema de bombas, el sistema hace pasar los reactivos a través de los pocillos del "*PicoTiterPlate".* Estos reactivos contienen los cuatro nucleótidos que constituyen el DNA, Adenina, Timina, Citosina y Guanina. Durante el flujo de nucleótidos, cada una de las cientos de miles de partículas, con millones de copias de DNA cada una, es secuenciada en paralelo. Primero, se hace fluir un nucleótido (G en la figura), si éste es complementario al nucleótido correspondiente de la cadena de DNA unida a la partícula, la enzima DNA-polimerasa (añadida en la mezcla) extenderá la cadena de DNA unida a esa partícula en cuestión mediante la adicción de ese nucleótido. Cada ciclo de secuenciación consiste en el flujo de nucleótidos individuales en un orden previamente fijado (TACG) a través del "*PicoTiterPlate"*.
La adicción de uno (o más) nucleótidos transcurre a través de una reacción que genera una señal luminiscente que es grabada en la cámara fotosensible CCD que posee el secuenciador. La fuerza de la señal es proporcional al número de nucleótidos incorporados en cada flujo de carrera individual de cada nucleótido. La luz resultante de la reacción de secuenciación viaja atravesando la placa base del "*PicoTiterPlate"* y es capturada por la cámara CCD. La combinación de la intensidad de señal con la posición del origen de esa señal permite al software determinar la secuencia de cientos de miles de reacciones individuales simultáneas, produciendo por tanto, millones de bases de secuencia por hora en una sola carrera.

Fase IV: Por último, tiene lugar el ensamblaje de las secuencias generadas y análisis de los datos. El programa que usan por defecto (Roche) los pirosecuenciadores 454 es el ensamblador "Newbler", pero existen otros ensambladores en el mercado actual gratuitos y además, no siempre es necesario disponer de las secuencias ensambladas.

3.2. Métodos de Secuenciación de Tercera Generación.

A partir del 2010/2011 comenzaron a desarrollarse las metodologías de secuenciación denominadas de "tercera generación", basadas en su mayoría en la secuenciación de una sola molécula y en la minimización de la química y aparataje usados en el proceso de secuenciación (Munroe & Harris 2010). Estos avances permitirán en un futuro muy cercano reducir el precio de la secuenciación de uno a dos órdenes de magnitud, lo cual propiciará el desarrollo del "la genómica personal": hacer la secuenciación de todo el genoma humano de cualquier persona en menos de un día, por 1.000 dólares o menos (Mardis 2008a; Mardis 2008b). Por lo tanto, estos avances permitirán la secuenciación de miles de procariotas (y virus), cuyos genomas son de un tamaño muy inferior al de células eucarióticas.

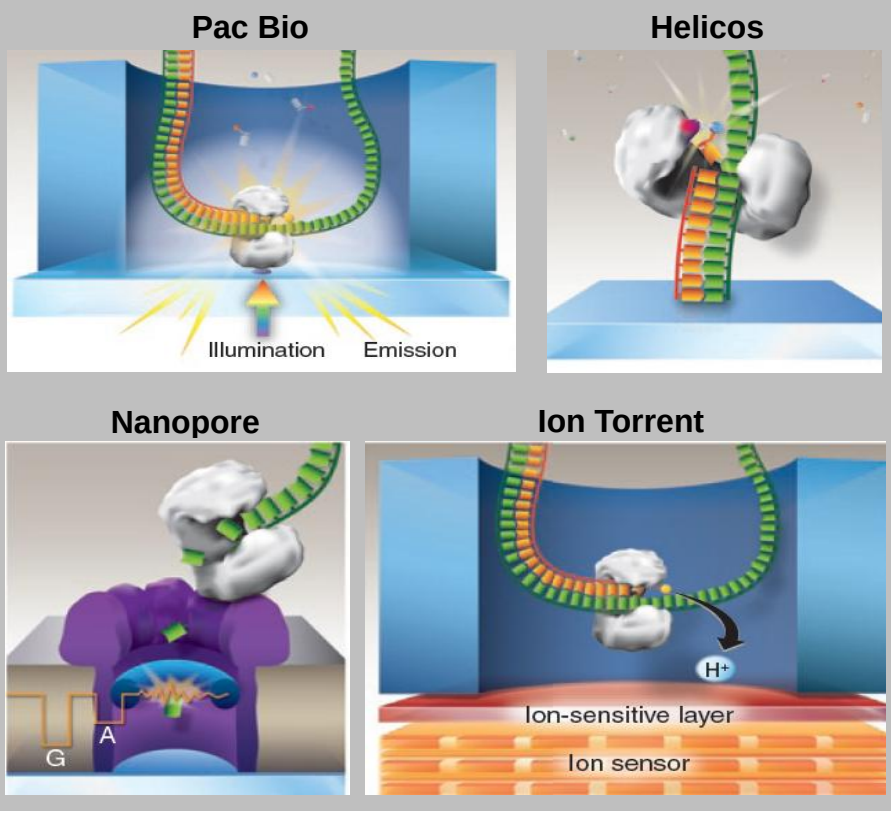
Alguna de estas metodologías se muestras esquemáticamente en la Figura 12:

1. *"Helioscope(TM) single molecule sequencing"* (http://www.helicosbio.com). El secuenciador de Helicos BioSciences se basa en la secuenciación de una sola molécula de DNA y utiliza adaptadores poli-A para fijar los fragmentos de DNA sobre la superficie de la célula de flujo. Mediante lavados cíclicos de la celda de flujo con nucleótidos marcados con fluorescencia (un tipo de nucleótidos a la vez, como con el método de Sanger), se sintetiza y se secuencia a la vez cada una de las cadenas de DNA. Las lecturas son cortas, de hasta 55 pb, y, supuestamente, este sistema es capaz de leer homopolímeros de DNA con una tasa de error baja. Esta tecnología es capaz de llegar a producir miles de millones de cadenas en un solo experimento (produciendo más que 2 Gpb de datos de secuenciación por día).
2. *"Nanopore".* Este método se basa en la lectura de la señal eléctrica que ocurre al pasar los nucleótidos de una cadena de DNA por un canal constituido por la proteína alfa-hemolisina unida covalentemente con ciclodextrina. Estos componentes serán sustituidos en el futuro por polímeros que desempeñen la misma función. El DNA, al pasar por este nanoporo cambia su corriente iónica, y este cambio, depende de la forma, tamaño y longitud de la secuencia de DNA. En teoría, cada tipo nucleótido provoca un flujo de iones con un período de tiempo diferente, de manera que cada uno de ellos puede ser identificado (en la realidad son tres nucleótidos los que están en contacto con el poro y por tanto su señal electroquímica). El método tiene un enorme potencial ya que no requiere nucleótidos modificados y teóricamente, no tendría límite en la longitud de la secuencia. En estos momentos, se ha demostrado que pueden llegar a generar lecturas de hasta 50 Kpb y generar 1.3 Gpb por run en 6 horas. En la segunda mitad del 2012, se espera la comercialización de su tecnología: MinION, un mini-secuenciador con conexión USB a un precio de 900 dólares y hasta 512 nanoporos (150 Mb/h), y GridION: mayor en tamaño que el anterior y con capacidad para 2000 nanoporos (en el 2013, llegarán a los 8000 nanoporos). Ete dispositivo podría trabajar en paralelo y 20 de ellos, por ejemplo, secuenciarían un genoma humano en 15 minutos. El error de secuenciación asociado es alto, 96%, pero supuestamente, podría ser minimizado a través de software.
3. *"Ion Torrent Systems Inc."* Es un sistema basado en la secuenciación estándar Sanger, pero con un novodoso sistema de detección basado en semiconductores. A diferencia de los métodos ópticos utilizados en los otros

sistemas de secuenciación, este método se basa en la detección de iones de hidrógeno liberados durante la polimerización del DNA. Cada una de las hebras de DNA de cadena sencilla es inmovilizado junto con una polimerasa en micropocillos, los cuales se inundan con un solo tipo de nucleótido. Si el nucleótido introducido es complementario a la cadena de DNA, éste se incorporará en la misma gracias a la DNA-polimerasa presente. Esto provoca la liberación de un ion de hidrógeno que dispara un sensor de iones hipersensible, lo que indica que una reacción se ha producido y por lo tanto se ha incorporado un nucleótido en la hebra de DNA sintetizándose. En el caso de homopolímeros, múltiples nucleótidos serán incorporados en un solo ciclo con la correspondiente liberación de protones y una señal electrónica proporcional. La longitud acutal de las lecturas de DNA con este sistema es de unas 100-200 bp, pero prometen alcanzar muy pronto las 400 pb.

4. *"PacBio".* Se basa en el anclaje de una DNA-polimerasa en micropocillos los cuales son inundados por nucleótidos marcados con fluoróforos de manera similar que en la tecnología Sanger. Las cadenas DNA se secuenciarían por síntesis de la cadena complementaria mediante la incorporación de un nucleótido complementario con la consiguiente liberación de fluorescencia. El error de secuenciación asociado es muy alto (hasta una 10%), sin embargo, a través de software y de grandes volúmenes de datos éstos se llegarían a minimizar hasta un rango aceptable.

Figura 12. Alguna de las tecnologías de secuenciación de Tercera Generación (adaptación de (Munroe & Harris 2010)).



En el Anexo III, Tabla B y Tabla C, es posible encontrar más información sobre otras tecnologías de tercera generación (Información obtenida de Tavis *et al.* 2011).

2.- PLANTEAMIENTO DEL TRABAJO Y OBJETIVOS

2.- PLANTEAMIENTO DEL TRABAJO Y OBJETIVOS

Los cristalizadores de las salinas solares están poblados por densas comunidades de *Archaea* halófilas, siendo *H. walsbyi* la más abundante. Tradicionalmente los ambientes hipersalinos se han descrito como sistemas muy simples dominados por unas pocas especies de microorganismos. Se han realizado numerosos estudios basados en el gen del 16S rRNA; sin embargo, el análisis este gen no refleja necesariamente la diversidad existente en la comunidad procariota de un determinado ambiente. De hecho, estudios genómicos y metagenómicos llevados a cabo con microorganismos de distintos orígenes han puesto de manifiesto la existencia de una gran diversidad dentro de muchas especies.

El objetivo principal de este trabajo es el estudio de la microdiversidad existente en una zona muy concreta del genoma de *H. walsbyi* DSM 16790 (HSBQ001), la denominada Isla Genómica I descrita por Cuadros-Orellana *et al.* en el 2007 por primera vez. Para ello se plantearon los siguientes objetivos específicos:

- Estudio de la diversidad de GI1 mediante secuenciación de fósmidos ambientales que codifiquen esa zona del genoma de *H. walsbyi*.
- Descripción de los genes presentes en GI1
- Comparación genómica entre las secuencias obtenidas y los genomas existentes de *H. walsbyi*.
- Estudiar los mecanismos de variabilidad genética de GI1 y la posible transferencia horizontal de los genes. Estudiar posibles recombinaciones y los hot-spots existentes en esta zona hipervariable.

3.- MATERIALES Y MÉTODOS

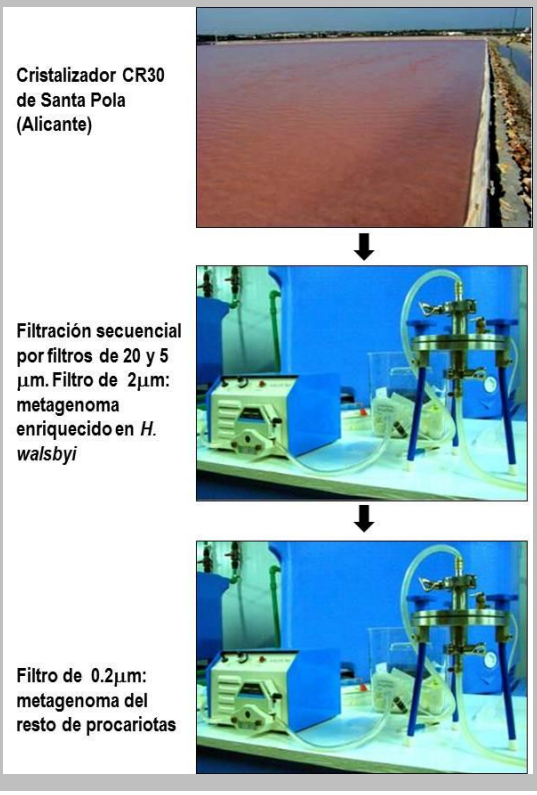
3.- MATERIALES Y MÉTODOS

Los experimentos se realizaron en el laboratorio de Microbiología de la Universidad Miguel Hernández de Alicante bajo la supervisión del Dr. Francisco Rodríguez Valera. La librería de fósmidos del cristalizador CR30 de las salinas de Santa Pola fue construida en el 2004, y la secuenciación de los extremos de los fósmidos fue publicada en el 2005-2006 por Legault *et al.* Esta colección de secuencias se encuentran en Genebank: GSS, 2947 secuencias totales, números de acceso DU826964-DU824018 (colección FLAS). Estas secuencias fueron los datos de partida en la búsqueda de fósmidos que contuviesen la isla GI1 (o parte de ella) usando como referencia el genoma de *H. walsbyi* DSM 16790 (HSBQ001). Aunque este trabajo se centra en el análisis bioinformático de los fósmidos secuenciados, se explicará en esta sección el origen de las mismas.

3.1. Construcción de la librería ambiental de fósmidos.

La biblioteca genómica ambiental utilizada en este trabajo se construyó como se describe en Legault *et al.* (2006). Aproximadamente 30 litros de agua se filtró secuencialmente por filtros de 20 y 5 micras de diámetro para eliminar las partículas de mayor tamaño y los posibles eucariotas presentes. Aprovechando la peculiar forma de *H. walsbyi*, células cuadradas, se pudo enriquecer la biomasa recogida del cristalizador CR30 en esta haloarchaea usando un filtro de 2 micras. El resto de la microbiota pasaría a través de este filtro y quedaría retenida en el filtro de 0.22 micras (Figura 13).

Figure 13. Enriquecimiento del metagenoma en *H. walsbyi*.



El DNA fue extraído de los filtros usando fenol-cloroformo y precipitándolo con etanol en presencia de sales. Previamente a la clonación, los extremos del DNA ambiental, fragmentos genómicos de 35 a 40 Kpb, se repararon según el protocolo indicado en el *CopyControlTM Fosmid Library Production Kit* (Epicentre). Se procedió a su ligación con el vector de clonación pCC1FOSTM (Epicentre) siguiendo una ratio molar de 10:1 entre el vector y el DNA empleado como inserto. Las reacciones de empaquetamiento y transición se llevaron a cabo según las recomendaciones del fabricante del *kit*. Se empleó la cepa EPI300-T1R de *E. coli* (Epicentre) como hospedador en la transformación. La eficiencia de la genoteca fue de aproximadamente 2000 clones. Las secuencias terminales de unos 1500 fósmidos se secuenciaron generando aproximadamente 2,4 Mpb de secuencia (2948 extremos) (colección FLAS).

3.2. Selección de los fósmidos a secuenciar.

A partir de los extremos de los fósmidos de la colección FLAS, se realizó una comparación mediante BLASTN para ver cuáles de ellos tenían similitud en el interior, en la frontera o en las cercanías de GI1 del genoma de *H. walsbyi* DSM 16790. El cut-off usado fue de un e-value 10^{-5}. Un total de nueve fósmidos fueron elegidos en función de la mayor similitud de al menos alguno de los extremos del fósmido con el genoma de *H. walsbyi* DCM 16790 (ver Figura 17). Únicamente, 8 de éstos contenían una nueva versión de la GI1 o parte ella.

3.3. Extracción de DNA.

El DNA de los fósmidos elegidos fue extraído para ser secuenciado. Para ello, cada clon se inoculó en 40 ml de LB con cloranfenicol (concentración final 12.5 µg/ml) y se incubó a 37ºC durante 16 horas. Las células se recogieron por centrifugación (durante 10 minutos a 3900 g) y el DNA se extrajo con el kit QIAprep Spin Miniprep kit (QIAGEN), siguiendo las indicaciones del fabricante y eluyendo el DNA en 40 µl de agua libre de nucleasas precalentada a 70ºC, tras una incubación de 3 minutos a 70ºC y 3000 rpm (*Thermomixer compact*, Eppendorf). Finalmente, los fósmidos extraídos se sometieron a electroforesis en geles de agarosa al 1% y su concentración se midió utilizando Quant-it ™ PicoGreen ® dsDNA reactivo (Invitrogen).

3.4. Pirosecuenciación.

Se procedió posteriormente a su pirosecuenciación (Roche 454, GS-FLX. Compañía: GATC, Konstanz, Alemania) marcando cada uno de los fósmidos individualmente mediante una adaptador MID (múltiplex identifier). Estos MID son secuencias únicas de 10 nucleótidos reconocidas por el software de análisis de secuenciación, lo que permite la clasificación automatizada de las lecturas que contienen los marcadores MID. La longitud promedio de cada una de las secuencias fue de 230 pb y el número promedio de lecturas fue de aproximadamente 4500 por fósmido (aproximadamente 20% de ellos pertenecían a *E. coli* EPI300 y el vector de clonación).

3.5. Ensamblaje de los fósmidos eHw.

El ensamblaje de cada uno de los fósmidos se realizó con tres programas diferentes para tratar así de disminuir la probabilidad de obtención de quimeras. El primero de ellos, fue el programa SeqMan (DNAstar) utilizando los siguientes parámetros: 50 pb de la secuencia de la superposición y el 95% de similitud. El segundo de ellos fue el

programa MIRA (http://www.chevreux.org/projects_mira.html) y el tercero de ellos CLC Genomics WorKpbench 3 usando un cut-off del 95% de identidad en al menos 50 pb.Todos los fragmentos ensamblados igualmente por las tres metodologías fueron los usados en este trabajo. Todos los fósmidos, con la excepción de eHw-5 y eHw-7, fueron reunidos en un solo contig. Los que no, fueron finalmente ensamblados en un solo contig y se procedió a su comprobación por PCR. Las características generales de los fósmidos se encuentran en la Tabla 1.

3.6. Anotación de los fósmidos eHw.

La predicción de las fases de lectura abierta (ORFs) se realizó usando el programa Glimmer v. 2.0 (Delcher *et al.* 1999) y posteriormente, fue repasada manualmente para eliminar los posibles errores característicos de las anotaciones automáticas. Todas los espaciadores intergénicos fueron comparados frente la base de datos nr ("no redundante", http://www.ncbi.nlm.nih.gov/) utilizando BLASTX (http://www.ncbi.nlm.nih.gov/BLAST/, (Altschul *et al.* 1990; Altschul *et al.* 1997) para asegurarse de no perder ninguna ORF posible. La búsqueda de homólogos de las proteínas codificadas en las secuencias de los fósmidos se realizó mediante BLASTP usando la base de datos nr. El cut-off usado fue: e-value 10^{-5}. No se tuvieron en cuenta aquellas ORFs menores de 30 codones y sin similitud alguna con otras proteínas conocidas. Adicionalmente, las secuencias de los fósmidos se anotaron automáticamente en la plataforma ISGA (http://isga.cgb.indiana.edu/Home) para su comparación con la anotación manual realizada. Se buscaron genes de RNA de transferencia mediante el programa tRNAScan-SE (http://lowelab.ucsc.edu/tRNAscan-SE) (Lowe & Eddy 1997).

Para cada una de las ORFs se analizaron también la presencia de motivos y/o dominios conservados mediante el programa HmmPfam, del paquete informático HMMER (Eddy 2008) usando los modelos hmm para los dominios proteicos de la base de datos Pfam (http://pfam.sanger.ac.uk). Otras bases de datos usadas fueron: Pfam (Finn *et al.* 2010) (http://pfam.sanger.ac.uk/), InterProscan (Quevillon *et al.* 2005) http://www.ebi.ac.uk/Tools/InterProscan/), PROSITE (Sigrist *et al.* 2010), http://www.expasy.ch/prosite/), Motif Scan (http://myhits.isb-sib.ch/cgi-bin/motif_scan), y SMART (Letunic *et al.* 2009) (http://smart.embl-heidelberg.de/). Para el análisis de la presencia de dominios transmembrana en las proteínas codificadas se usó TMHMM 2.0 (Krogh *et al.* 2001). Para la predicción de la estructura secundaria de las proteínas así como las regiones de baja complejidad se usó Medor (http://www.vazymolo.org/MeDor/index.html). La predicción de la estructura terciaria de las proteínas se realizó con CPHmodels-3.0 server (http://www.cbs.dtu.dk/services/) y Swiss-model (http://swissmodel.expasy.org/). También se analizó la presencia de péptidos señal en los extremos de las secuencias proteicas mediante el programa SignalP 3.0 (Bendtsen et al. 2004) (http://www.cbs.dtu.dk/services/). Para la detección de la posible localización celular, se usó Targetp v1.1 (http://www.cbs.dtu.dk/services/).

3.7. Análisis bio-informáticos de los fósmidos eHw.

Comparaciones genómicas. Las comparaciones entre los fósmidos secuenciados, los extremos de los fósmidos y el genoma HSBQ001 se realizaron con BLASTN. Para la identificación de regiones de similitud se llevaron a cabo búsquedas recíprocas BLASTN y TBLASTX entre los fósmidos y los genomas de *H. walsbyi* disponibles. Para permitir la visualización interactiva de estas comparaciones se usó el programa Artemis Comparison Tool ACTv.8 ACT (The Sanger Institute, http://www.sanger.ac.uk/Software/ACT/).

N-glicosilación y O-glicosilación. La recombinación se detectó usando los programas NetNGlyc y NetOGlyc de http://www.cbs.dtu.dk/services/.

Contenido en GC: se calculó usando el programa "geecee" del paquete EMBOSS (Rice *et al.* 2000).

Análisis de frecuencia de tetranucleótidos. Se calculó la frecuencia de tetranucleótidos de los fósmidos secuenciados mediante el programa "wordfreq" del paquete EMBOSS (Rice *et al.* 2000).

Uso de codones: se determinó con la aplicación CODONTREE (http://bioweb.pasteur.fr/seqanal/interfaces/codontree.html).

Análisis de multicomponentes. "R" se usó para llevar a cabo el cálculo de multicomponentes de las frecuencias de tetranucleótidos y del uso de codones (R package FactoMineR, Le *et al.* (2008)).

Masas y puntos isoeléctricos de los péptidos: se determinaron *in silico* con la aplicación ProtParam en http://www.expasy.ch/tools/protparam.html (Gasteiger *et al.*, 2005).

Repeticiones: las repeticiones en tándem se analizaron *online* en la dirección http://tandem.bu.edu/trf/trf.html (Benson, 1999).

Alineamientos y análisis filogenéticos: Los alineamientos se realizaron con MUSCLE v. 3.6 (Edgar 2004) y editados manualmente según fue necesario. Los análisis filogenéticos de las proteínas se realizaron utilizando el software de análisis filogenético MEGA4 usando "*Neighbour-Joining"* (http://www.megasoftware.net) y Phylip usando *"maximum-likelihood"* (seqboot, 100 árboles FastTree, Consense program).

Detección de recombinaciones. Para detectar posibles recombinaciones entre los fósmidos secuenciados y los genomas de *H. walsbyi*, se usó el programa RDP3 ("Recombination Detection Program" (Martin *et al.* 2010)) que usa varias metodologías diferentes (ver apartado 4.7) integradas en el mismo *software para* detectar fenómenos de recombinación en un alineamiento de al menos de más de tres secuencias introducido. Al menos, el fenómeno de recombinación ha de ser detectado por tres métodos diferentes con valores mayores de e-value 10^{-5}.

(Los programas desarrollados en Perl para este trabajo se encuentran en el ANEXO II).

3.8. Reclutamientos metagenómicos.

Las secuencias del metagenoma SS37 pertenecen a una muestra recogida en CR30 en Junio del 2008 en las salinas "Bras del Port", Santa Pola, Alicante (38° 12' N, 0° 36' W), y ha sido recientemente descrito en Ghai *et al.* 2011 (el nombre hace referencia a la salinidad medida en el momento de muestreo). Usando el metagenoma SS37 del cristalizador CR30 como *"Query"*, se llevaron a cabo búsquedas mediante BLASTN frente a una base de datos formada por cada uno de los fósmidos secuenciados o cada uno de los genomas completos de *H. walsbyi*. El cut-off usado para considerar la secuencia metagenómica o *"hit"* fue, como mínimo, una identidad del 70% en un 50% de su longitud. OriginPro v. 8.0724 fue usado para la representación gráfica de los reclutamientos genómicos. El ratio para cada uno de los genes de la isla se calculó teniendo en cuenta el número de secuencias que daban un hit al gen dividido por la longitud del mismo en pares de bases.

3.9. Número de acceso de las secuencias públicas usadas en este trabajo.

Las secuencias de los fósmidos no han sido publicadas previamente y son material inédito del presente trabajo. Éstas estarán a disposición de quien las solicite. Tal como se citado anteriormente, los números de acceso para la

colección FLAS es: DU826964-DU824018. El número de acceso para eHw-559 es: EF584001.1. El metagenoma SS37 ha sido depositado en el archivo SRP007685. La secuencia de la cepa *H. walsbyi* HSBQ001 está disponible bajo los números de adhesión GenBank AM180088.1 (cromosoma) y AM180089.1 (plásmido pL47). Las secuencias de cepa *H. walsbyi* C23 están disponibles en el BioProject: Accession: PRJEA49335ID: 49335.

Para la gráfica de las curvas de GC% de la Figura 14: *S. ruber* DSM 13855 (M31): BioProject: PRJNA16159, *S. ruber* M8: GenBank: FP565814.1 y *Halorubrum lacusprofundi*: BioProject PRJNA18455.

4.- RESULTADOS

4.- RESULTADOS

Uno de los retos más importantes para la comprensión de la diversidad procariótica y la evolución es la notable heterogeneidad de los genomas que se pueden encontrar dentro de una sola especie u OTU. Los estudios llevados a cabo principalmente con cepas patógenas han puesto de manifiesto que los genomas son muy dinámicos y que éstos pueden cambiar drásticamente en el repertorio de genes tamaño y sintenia entre las diferentes cepas o linajes ambientales pertenecientes a una misma especie. Un caso muy particular de esta variabilidad es la isla genómica GI1* de *H. walsbyi,* donde se encuentran proteínas implicadas en la construcción de la envoltura celular (capa S). Antes de este trabajo se conocían tres versiones diferentes de GI1, la de las cepas de los genomas secuenciados HSBQ001 y C23 y la encontrada por casualidad en el fósmido eHw-559. Para conocer la existencia de otras posibles variaciones en éste área del genoma, se secuenciaron 9 fósmidos más que provenían de la misma librería metagenómica construida en el 2004 y de la cual se derivaron los trabajos de Legault *et al.* (2006) y Cuadros-Orellana *et al.* (2007) (ver Introducción).

***Nota:** el término GI1 usado en este texto hace referencia a la isla genómica rica en glicosiltransferasas como la encontrada en la cepa HSBQ001. Con este mismo nombre haremos referencia a la zona equivalente en C23, aunque esta no sea la primera isla genómica en orden y no cumpliría tampoco los criterios de tamaño establecidos por Legault *et al.* (2006) de tener al menos 20 Kpb.

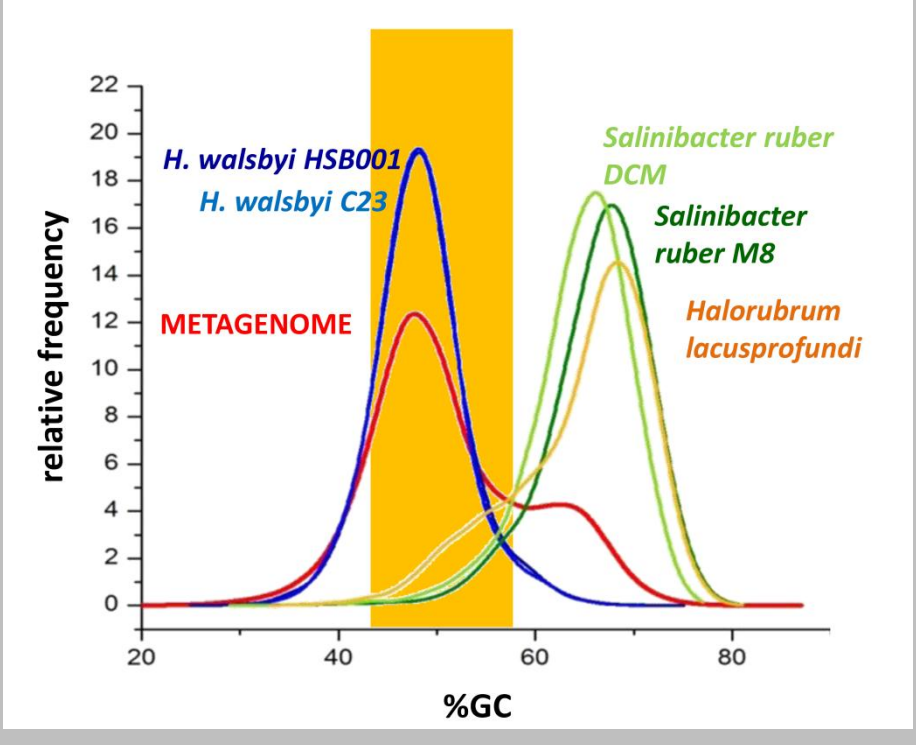
4.1. *Haloquadratum*, una sola especie: *H. walsbyi.*

En nuestro caso, antes de comenzar con el análisis de los fósmidos secuenciados, era de especial importancia saber si bajo el género de *Haloquadratum* existen más especies además de *H. walsbyi*, ya que entonces, cabe la posibilidad de que alguno de los fósmidos secuenciados no pertenezcan a la especie *H. walsbyi*, sino a otro *Haloquadratum* aún no aislado. El criterio tradicional generalizado para definir "especie" en procariotas viene dado por el marcador filogenético 16S rRNA: por encima del 97% de identidad en este gen se considera misma especie (Konstantinidis & Tiedje 2005). Como se menciona anteriormente, el análisis de la población de *Haloquadratum* a través de secuencias del marcador 16S rRNA en varias salinas de Australia (Oh *et al.* 2010) reveló una divergencia dentro de esta especie muy baja, mostrándose que todas las secuencias ribosomales se parecen más del 98%, en concordancia también con la diferencia del 1.6% encontradas entre los 16S de los genomas HSBQ001 y C23 (Burns *et al.* 2007). Así, bajo el criterio de este marcador filogenético, únicamente existiría una sola especie de *Haloquadratum* dentro del género, *H. waslbyi*. Sin embargo, tal como se ha mencionado en la Introducción del presente trabajo, existen numerosos casos en los que la conservación en la similitud del 16S rRNA, no es indicativo de la conservación en el resto del genoma. Se considera que la media de identidad nucleotídica (ANI) entre micoroorganismos de una misma especie ha de ser mayor del 94%. Sin embrago, es posible encontrar casos en los que cepas de una misma especie que poseen un 16S rRNA muy similar, la media de identidad

nucleotídica (ANI) tiene valores muy bajos, como por ejemplo, entre cepas de *Prochlorococcus* sp. (Coleman *et al.* 2006), donde la ientidad del 16S rRNA es del 99.2% pero comparten un ANI del 78%), o *Alteromonas macleodii,* donde los aislados tienen un 99% de identidad en el 16S rRNA pero un ANI del 81.24% (Ivars-Martinez *et al.* 2008).

Aunque el ANI de los dos genomas de *H. walsbyi* es del 98.6% (Dyall-Smith *et al.* 2011), los datos revelados por la variabilidad encontrada en los fósmidos secuenciados publicados por (Cuadros-Orellana *et al.* 2007), sugerían una variabilidad mayor. Ya que no existen más genomas disponibles de *Haloquadratum*, realizamos una aproximación usando el metagenoma SS37 para tratar de averiguar si dentro de este género, podrían existir otras especies diferentes a *H. walsbyi*. Para ello, en primer lugar, se construyó una curva del contenido G+C de las secuencias del metagenoma SS37 (Figura 14), ya que por el bajo contenido en G+C característico de *Haloquadratum*, se podrían reconocer las secuencias en el metagenoma pertenecientes a esta especie. Se comprobó que, al igual que el primer metagenoma publicado en Legault *et al.* (2006), existe un máximo que coincide con el observado para los dos genomas de *H. walsbyi* (aproximadamente un 70% son secuencias con un G+C comprendido entre el 40-55%, Figura 15). Sólo un bajo porcentaje de secuencias ambientales pertenecerían a *S. ruber* o *Halorubrum lacusprofundi* (detectados también por 16S rRNA en el cristalizador, (Ghai *et al.* 2011)).

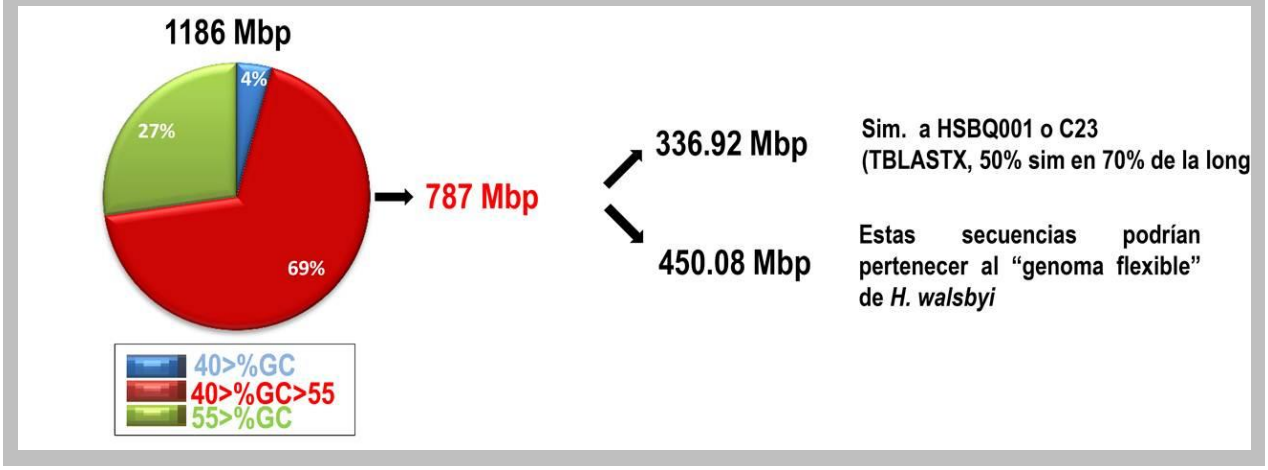
Figura 14. Contenido en G+C del metagenoma SS37 y de los microorganismos más abundantes en CR30 según Ghai *et al.* (2011): *H. walsbyi*, S. ruber y *H. lacusprofundi*.



Mediante comparaciones TBLASTX y usando como cut-off del 50% en similitud en un 70% de la longitud de las secuencia ambiental, se determinó que el 57.2% de las secuencias del metagenoma podrían pertenecer al genoma flexible de *H. walsbyi* (Figura 15) ya que carecían de similitud con ninguno de los genomas secuenciados de *H. walsbyi.* Este dato, se ha de tomar con precaución ya que parte de estas secuencias pueden pertenecer a virus

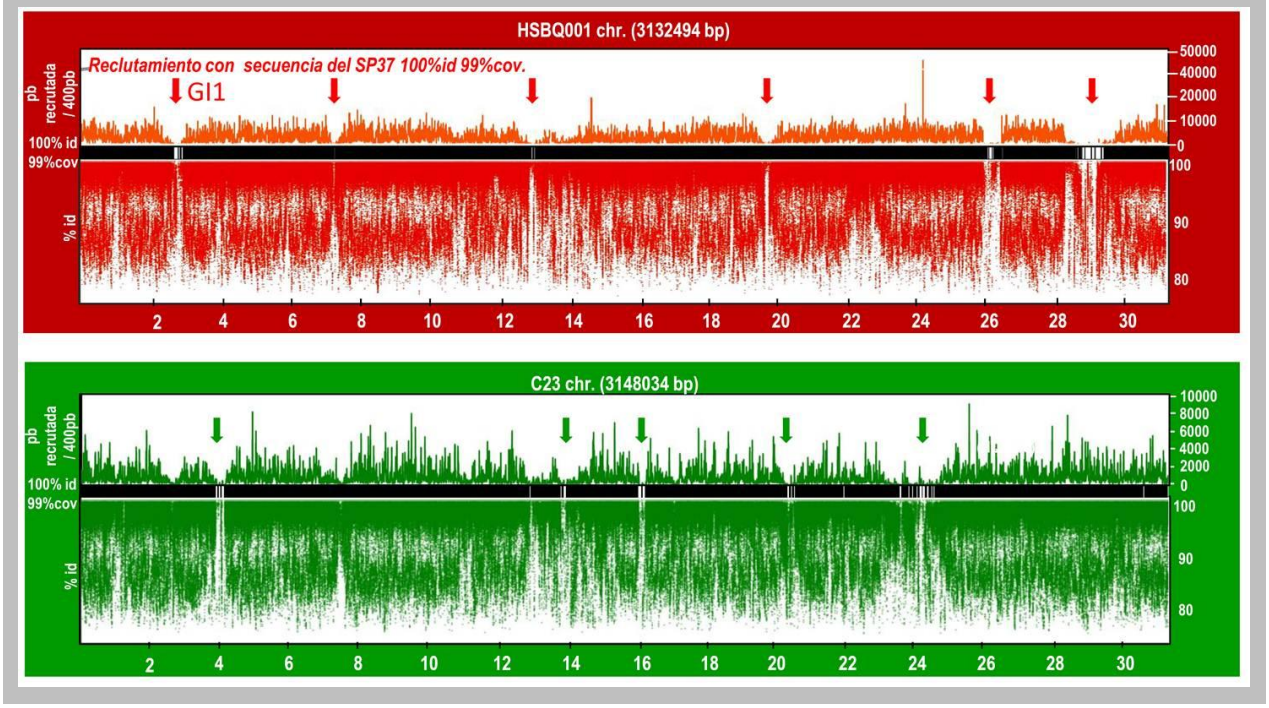
presentes en las células en el momento de la toma de muestra (y que pueden tener muy probablemente el mismo G+C, (Garcia-Heredia, 2012 #6911)(Santos *et al.* 2010)). También, estas secuencias podrían pertenecer a otros microorganismos de bajo G+C aún no descritos en los cristalizadores, como las nanoarchaeas presentes a bajas salinidades (Ghai *et al.* 2011; Narasingarao *et al.* 2011). Aun tomando estas consideraciones, la gran mayoría de estas secuencias pertenecerán a *H. walsbyi*, el organismo mayoritario, y por tanto, la variabilidad existente dentro de esta especie podría ser enorme.

Figura 15. Clasificación de las secuencias del metagenoma SS37 en función de contenido en G+C y similitud con *H. walsbyi*.



Esta variabilidad queda reflejada también en los reclutamientos realizados de los dos genomas disponibles, HSBQ001 y C23. A partir de los datos de un BLASTN (cut-off: 70% de la longitud de la secuencia ambiental alineable con el genoma) se construyó la gráfica de la Figura 16.

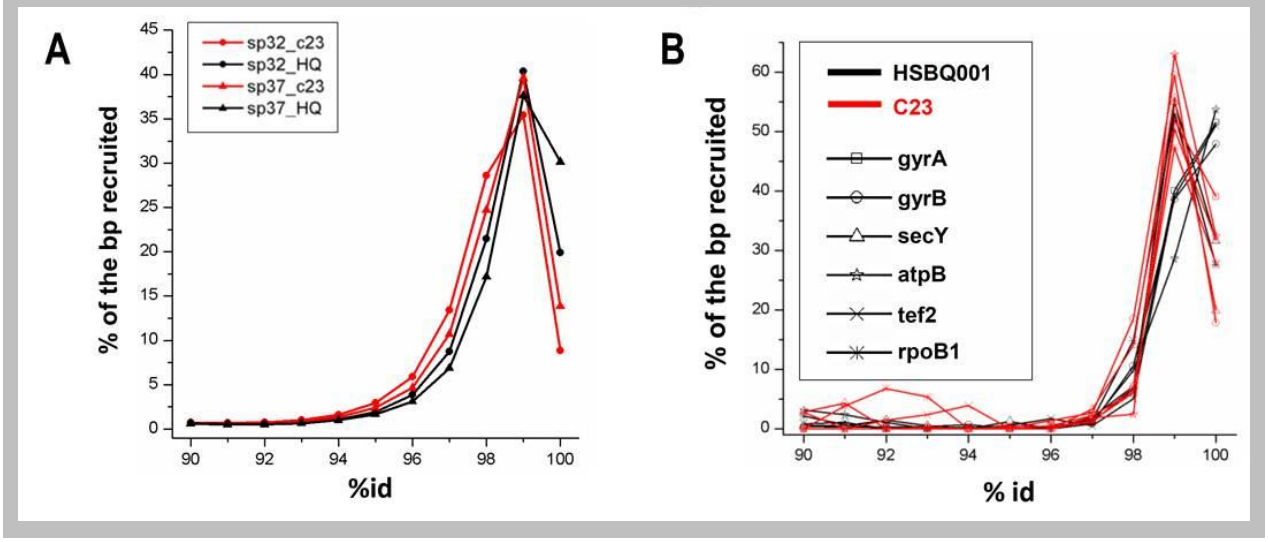
Figura 16. Reclutamiento de los genomas de los aislados *H. walsbyi* HSBQ001 y C23 en SS37 (Santa Pola, Alicante).



La presencia de ambos clones de *H. walsbyi* en las salinas de Santa Pola se demuestra por el alto número de secuencias de por encima del 95% de identidad, con un total de 123 Mpb encontradas en SS37 (39,79% del número total de lecturas, ver Figura 15). Sin embargo, a pesar de que el metagenoma usado es de un tamaño considerable, más de 1 Gpb y que el grado de cobertura de ambos genomas es muy alto (39x HSBQ001 y 28x C23), es posible aún seguir identificando islas genómicas. Cuando únicamente se tiene en cuenta el reclutamiento de las secuencias 100% idénticas, estas islas son más fácilmente identificables (ver panel superior de cada uno de los gráficos de los reclutamientos de la Figura 16). La existencia de islas genómicas en ambos aislados a pesar de la alta cobertura, sugiere la co-existencia de varios clones o linajes cuyas únicas diferencias son estas zonas variables del genoma. Además, cuando se analizó el número de secuencias 100% idénticas a cada uno de los clones, se comprobó que HSBQ001 es 3 veces más abundante que C23, por lo que la distribución de estos clones no es equitativa.

En la gráfica del reclutamiento puede observarse que existe una discontinuidad de variación genética entre el 90 y el 95%. Esta discontinuidad se observa mejor en la Figura 17A. Este hecho se puede asociar a la existencia de una sola especie de *H. walsbyi.* Por debajo del 90% se encontrarían secuencias pertenecientes a otras especies de haloarchaeas presentes en el cristalizador. Este mismo hecho se observa cuando se hacen reclutamientos para genes individuales, como los genes "*housekeeping" gyrA, gyrB, secY, atpB, tef2* y *rpoB1* (Figura 17B). Los resultados muestran que, en efecto, independientemente del gen, la mayoría de las secuencias se encuentran dentro del rango 95-100%, el intervalo de variación esperable dentro de secuencias de una misma especie (Konstantinidis & Tiedje 2005).

Figura 17. Número de secuencias del metagenoma SS37 que reclutan entre el 90-100% con (A) los genomas de *H. walsbyi* HSBQ001 y C23, (B) diferentes genes "housekeeping".

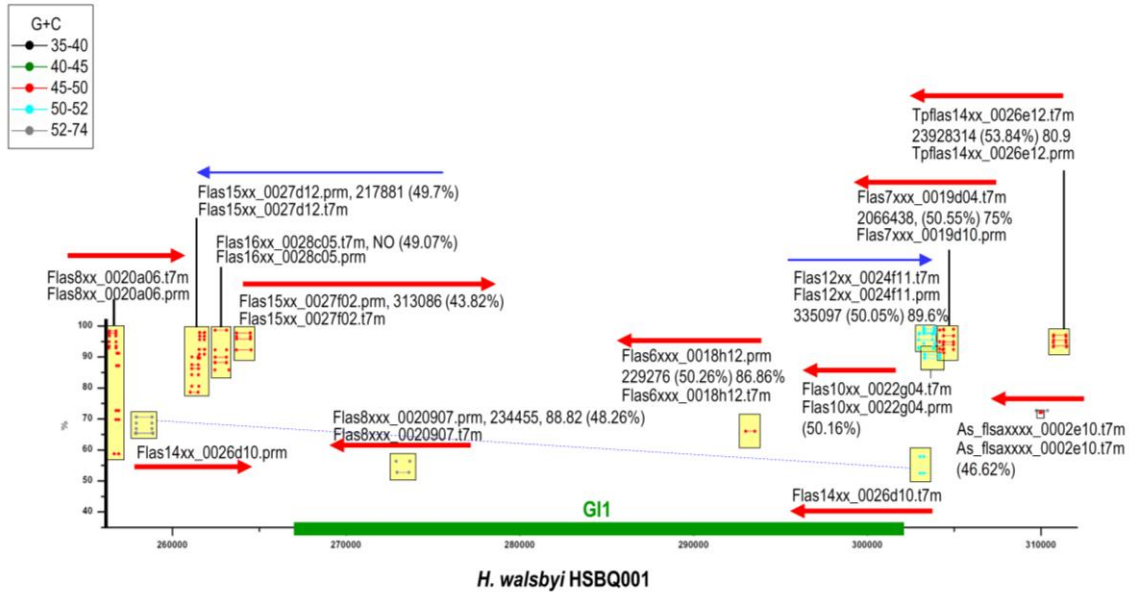


Con estos resultados se concluye que existe una sola especie dentro del género de *Haloquadratum* y por lo tanto, todas las GI1 secuenciadas pertenecerían a diferentes clones de *H. walsbyi* presentes en el momento de muestreo.

4.2. Selección de los fósmidos a secuenciar.

La probabilidad de encontrar extremos de fósmidos que mapeasen en el interior de IG1 a partir de las secuencias del primer metagenoma secuenciado (Legault *et al.* 2006) era muy baja (estas zonas no reclutaban nada, o casi nada, en el metagenoma). Sin embargo, sí era probable poder encontrar extremos de fósmidos que mapeasen en las cercanías o en los bordes de la misma. El tamaño aproximado de la IG1 de *H. walsbyi* DSM 16790 es de 45 Kpb, luego, un único fósmido sería capaz de albergarla, o al menos, cubrir una parte significativa de la misma. Usando esta aproximación se eligieron un total de 9 fósmidos en función de la mayor similitud de la secuencia con *H. walsbyi* en al menos uno de los extremos y también de su contenido en G+C (Figura 18, las flechas rojas indican la posición de los extremos de los fósmidos en le genoma de referencia HSBQ001). Tan sólo en un caso, ambos extremos de los fósmidos mapearon en la GI1 (Flas14xx_0026d10, unidos por una línea discontinua en la Figura 18).

Figura 18. Extremos de los fósmidos secuenciados en las cercanías de la GI1 de *H. walsbyi* DSM 16790.



4.3. Resultados de la secuenciación. Predicción de genes y anotación de los fósmidos eHw.

Los resultados de la secuenciación indicaron que tan sólo 7 de ellos abarcaban una posible GI1 ya que los otros dos mapeaban en zonas próximas a GI1 conservadas con HSB001 y C23. Uno de los fósmidos secuenciados, era prácticamente idéntico (15 polimorfismos) al previamente descrito eHw-559 (Cuadros-Orellana *et al.* 2007). Ante la posibilidad de que se tratase de una contaminación de la librería metagenómica, esta secuencia se retiró del conjunto analizado en este trabajo. Algunas de las características principales de los fósmidos que se analizan en este trabajo, eHw-1, 4, 5, 6, 7, 9 y 12, se pueden encontrar en la Tabla 1. Todos ellos presentan un contenido de G+C típico de *H. walsbyi* que varía de 40 a 55% (el G+C de HSBQ001 es 47,9% y de C23, 47,78%). A pesar de su contenido en G+C bajo, cabía la posibilidad de que las secuencias de los fósmidos pudiesen pertenecer a otros microorganismos presentes, luego se impuso la condición de que al menos 1000 pb fuesen sinténicas y por

encima del 95% de identidad con cualquiera de los genomas de *H. walsbyi*. Todos los fragmentos secuenciados cumplían esta condición. En todas las comparaciones realizadas en este trabajo, se incluirán las otras versiones de GI1 previamente descritas, la presente en el fósmido eHw-559, y las correspondientes a los dos genomas de *H. walsbyi* disponibles, HSBQ001 y CS23.

Tabla 1. Principales características de los fósmidos secuenciados.

Fósmido	Longitud (pb)	GC(%)	Nu. ORFs	Nu. Elementos IS	Nu. CSGs/adhesinas
eHw-1	37028	49.25	29	1	2
eHw-4	41965	49.21	25	6	9
eHw-5	37502	47.06	16	1	5
eHw-6	36944	47.77	25	-	6
eHw-7	35977	52.61	23	1	8
eHw-9	36930	48.81	27	7	5
eHw-12	33841	50.98	20	6	5

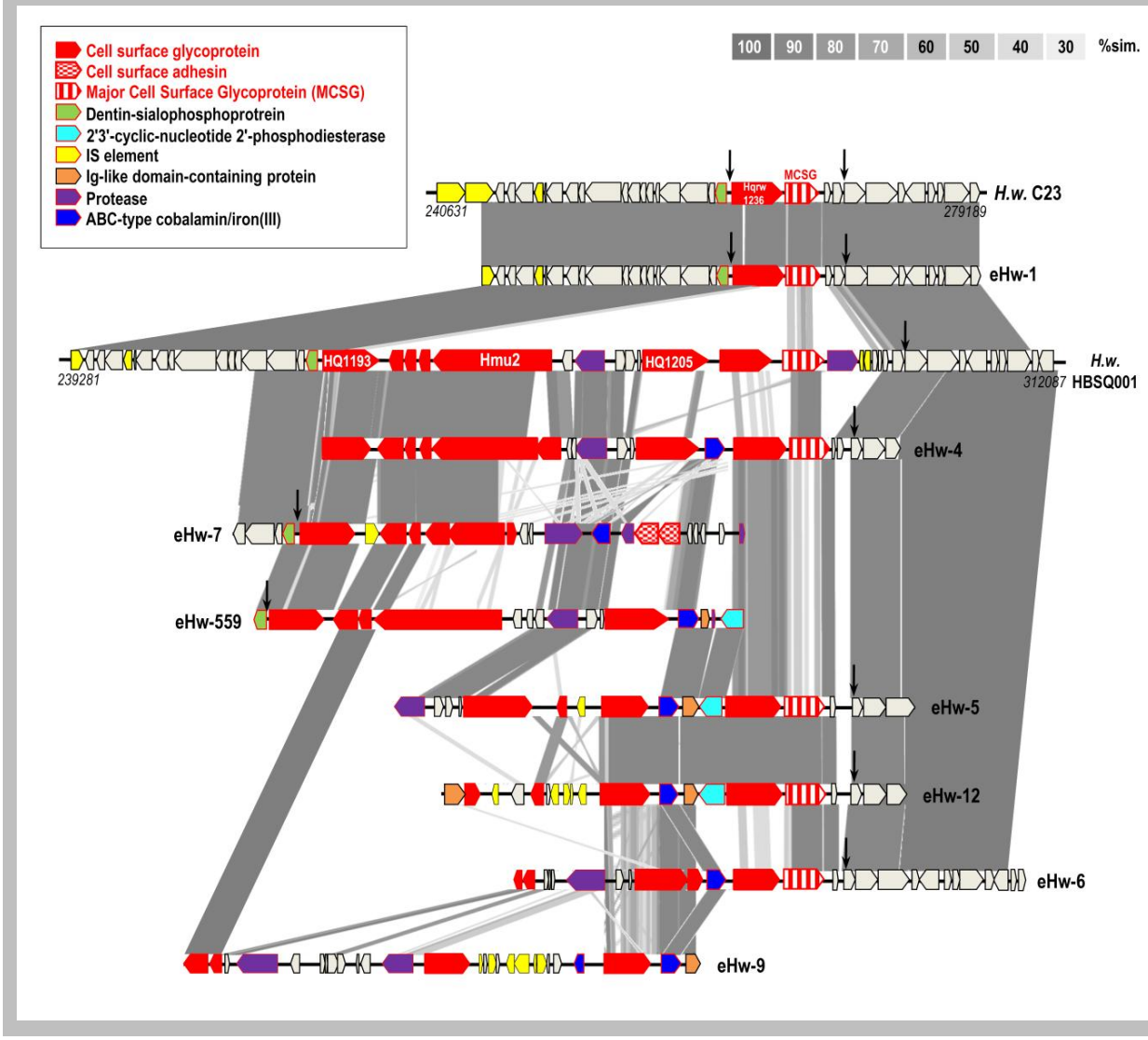
El primer paso fue realizar una anotación de todas las ORFs presentes en los fósmidos de la manera más detallada posible. Mediante BLASTP se determinó una posible función para cada proteína codificada. También se midieron los siguientes parámetros para cada uno de genes/proteína presentes (ver Materiales y Métodos): contenido en G+C, predicción de péptido señal, predicción de dominios transmembranales, predicción de dominios conservados, existencia de repeticiones y existencia de sitio de unión a ribosomas (ver Tabla Anexo I. Anotación de cada una de las ORFs detectadas en los fósmidos).

4.4. Comparaciones recíprocas y con los genomas de *H. walsbyi.*

GI1 de *H. walsbyi* contiene un número variable de genes, la mayoría de ellos son genes que codifican proteínas implicadas en la formación de la envoltura celular. Comparando los siete fragmentos de DNA secuenciados con las secuencias disponibles de la GI1 (genomas C23 y HSBQ001, y la secuencia del fósmido eHw559), se comprobó que la sintenia está parcialmente conservada excepto en el caso del fósmido eHw-1, el cual es sinténico con C23, el aislado australiano, y presenta un 96% de identidad nucleotídica con éste. La comparación de todos los fósmidos se muestra en la Figura 19.

Con los datos disponibles, podemos afirmar que existen al menos dos tipos diferentes de GI1. La versión más simple de GI1 se encuentra en C23 y eHw-1 y estaría formada por sólo dos genes. Sin embargo, el resto de las GI1 secuenciadas presentan un mayor número de genes, con una aparente mayor complejidad en la síntesis de la envuelta celular. GI1 en HSBQ001 tiene 20 ORFs, de las cuales, seis genes son glicoproteínas de la pared celular. Entre ellas, la putativa proteína de la capa S (MCSG), la denominada halomucina Hmu2, una proteína anotada como adhesina (ALP) y dos proteasas.

Figura 19. Comparación entre los fósmidos de la GI 1 secuenciados en este trabajo.



De las nueve versiones diferentes de la isla GI1 disponibles, seis de ellas contenían la MCSG (los fósmidos restantes no cubren esta parte de la isla ya que este gen se encuentra en un extremo). En algunos casos, como en eHw-6, ésta sería de las pocas proteínas en común con el resto de las islas (además de la parte conservada fuera de la GI1). Casi siempre, eHw-6 y 9 sería una excepción, los fósmidos contienen clusters alternativos de CSGs, aunque en vez de tratarse de genes diferentes, podrían considerarse homólogos con muy baja similitud. Una posible explicación de esta diferencia numérica podría ser el hecho que, a diferencia de la cepa australiana, las células de HSBQ001 tienen una doble envuelta celular (Figura 4), por lo que sería necesario un mayor número de proteínas implicadas en su síntesis. Curiosamente, Kessel & Cohen (1982), comprobaron la existencia de hasta tres tipos de subunidades diferentes en la capa S mediante difracción de Rayos X: dos de simetría hexagonal de tamaños diferentes y una tetragonal (*Haloarcula japonica* y *Halobacterium salinarum* R1 son subunidades hexagonales), luego puede ser necesario la existencia de otras CSGs en la envuelta celular si se trata de una u otra capa S.

4.4.1. Variación de la proteína mayoritaria de la capa S (MCSG, "Mayor Cell Surface Glycoprotein").

De especial interés son las proteínas detectadas como "mayor cell surface glycoproteins" (MCSG) ya que son los componentes mayoritarios de la envuelta celular (genes rallados verticalmente en la Figura 19). La descripción de alguno de los parámetros que se comentarán más adelante se puede encontrar en la Tabla 2.

Tabla 2. Características principales de las MCSG encontradas en *H. walsbyi* C23, HSBQ001, eHw-559 y los fósmidos secuenciados en este trabajo.

	CDS	Longitud (aa)	Péptido señal	Num. Sec. amb. # 99% id / 99%cov	Num. secuencias/Kpb
C23	Hqrw_1237	822	Y	8	3.24
HSBQ001	HQ1207	988	Y	53	17.86
eHw-1	eHw1-orf21	822	Y	10	4.06
eHw-4	eHw-4 orf14	982	Y	60	20.34
eHw-5	eHw-5 orf20	978	Y	49	16.68
eHw-6	eHw-6 orf16	988	Y	74	24.94
eHw-12	eHw-orf12	878	Y	46	18.11

#=Número de secuencias ambientales que reclutan a 99% identidad y 99% cobertura.

Mediante espectrofotometría de masas de las proteínas presentes en membranas celulares de *H. walsbyi* C23 (Tabla 3 suplementaria de Dyal-Smith *et al.* (2011)) se comprobó que existen dos proteínas mayoritarias: Hqrw_1237 y Hqrw_1641, presentando ambas un número de péptidos muy similar. Otras glicoproteínas detectadas, pero no tan abundantes como las anteriores fueron: Hqrw_1240, Hqrw_1641 y Hqrw_2184. Los correspondientes homólogos en el genoma de HSBQ001 son, por orden de citación: HQ_1207 (aunque su similitud es muy baja), HQ_1540, HQ_1214, HQ_1540 y HQ_2017. De todas ellas, Hqrw_1237 es la única que se encuentra en la GI1 de C23, mientras las demás están en zonas del genoma conservadas. La similitud de las dos proteínas mayoritarias, Hqrw_1237 y Hqrw_1641, es muy baja, alcanzando como máximo un 32% en el extremo amino y carboxilo (un 29% de la longitud total). Ambas se parecen a otras proteínas de la capa S de otros archaeas, Hqrw_1641 es un 42% similar a la de *Archaeglobus profundus* DSM 5631, y Hqrw_1237, a la de *H. japonica* TR-1. Ya que Hqrw_1237 mapeaba en GI1, la región que estamos analizando en este trabajo, pasaré a detallar más aspectos de la misma y de sus homólogas presentes en los fósmidos secuenciados.

Todas las MCSG encontradas en *H. walsbyi*, tanto en los genomas como en los fósmidos, tienen todos los dominios y características confirmadas en otras proteínas mayoritarias de la capa S (Figura 19 y 20): dos dominios transmembranales (N y C-terminal), varios sitios de N y O-glicosilación, y un tamaño y estructura secundaria muy parecida para la MCSG que forma la capa S en *H. japonica* TR-1 de patrón hexagonal (Nakamura *et al.* 1995). Con ésta comparte una similitud del 46% (65% en la parte C-terminal, zona que se encuentra mejor conservada entre todas ellas) (Figura 21). En el resto de los fósmidos secuenciados, las putativas MCGGs difieren en tamaño, hasta 166 aminoácidos más y, también, en estructura secundaria (Figura 22) (excepto el extremo C-terminal, mejor

conservado). Todas ellas presentan una similitud entre el 40-48% con la proteína de la capa S de *H. japonica* TR-1 (62 y 71% en la parte más conservada) y entre el 38-45% con la de *Halobacterium salinarum* R1 (o NRC-1). La mayor similitud de Hqrw_1237 además de su homóloga del fósmido eHw-1, es con la presente en el fósmido eHw-12, con una similitud del 49% (45% con la de HSBQ001, HQ_1207). Entre las proteínas de la capa S detectadas en los fósmidos, la mayor similitud la presentan entre las versiones presentes en eHw-12 y eHw-5, con una similitud del 86%, seguido del par presente en eHw-4 y HQ_1207 con un 85%. El alineamiento de estas proteínas se encuentra en la Figura 21. En él se han marcado con asteriscos los sitios conservados de O-glicosilación con la MCSG de *Haloferax volcanii* (Jarrell *et al.* 2010a; Jarrell *et al.* 2010b). También se ha remarcado el dominio transmembranal conservado, el péptido señal y una zona de baja complejidad de aminoácidos repetidos (PTPTPT…).

Los resultados obtenidos de la predicción de la estructura secundaria mostraron que en *H. walsbyi* existen al menos tres topologías diferentes para las MCSG detectadas (Figura 22). La primera sería el par formado por Hqrw_1237 y la de eHw-1, la segunda estaría formada por HQ-1207 y la de eHw-4 y finalmente, las presentes en eHw-5, eHw-6 y eHw-12, siendo esta última de menor tamaño ya que carece de un fragmento hacia el extremo amino rico en láminas beta. Esta relación estructural también se mostró al estudiar su relación filogenética (Figura 23). También se llevó a cabo la predicción tridimensional de las MCSG de *H. walsbyi* y todas ellas poseían una homología remota con auto-transportadores con un dominio pertactina en su estructura: las MCSG de eHw-5 y 6 son similares a la estructura del cristal 1WXR.A, y las presentes en eHw-1, 4, 12 y HSBQ001 al cristal 2IOU.G. 1WXR es una proteína de unión al grupo hemo presente en *E. coli* y 2IOU es un complejo del bacteriófago de *Bordetella* DGR unido a una proteína variable determinante del tropismo de esta bacteria (la cual tiene un dominio pertactina) (Liu *et al.* 2002; Miller *et al.* 2008). Interesantemente, la subunidad G del cristal 2IOU es la parte del complejo reconocida por el fago. Este resultado sugiere por tanto que estas MCSGs de *H. walsbyi* podrían ser las dianas reconocidas por fagos presenten en el cristalizador.

Figura 20. Dominios estrucuturales encontrados en las putativas MCSG encontradas en los fósmidos secuenciados (eMCSG) y en los genomas de *H. walsbyi*. (SP: péptido señal, TM: dominio transmembranal, N: asparraginas glicosiladas, S y T: serinas y treoninas glicosiladas. Los números indican la posición del SP y de la posible glicosilación). Debajo de cada MCSG, se encuentran mapeadas las secuencias ambientales del metagenoma SS37 en función del número de polimorfismos respecto a la MCSG de referencia.

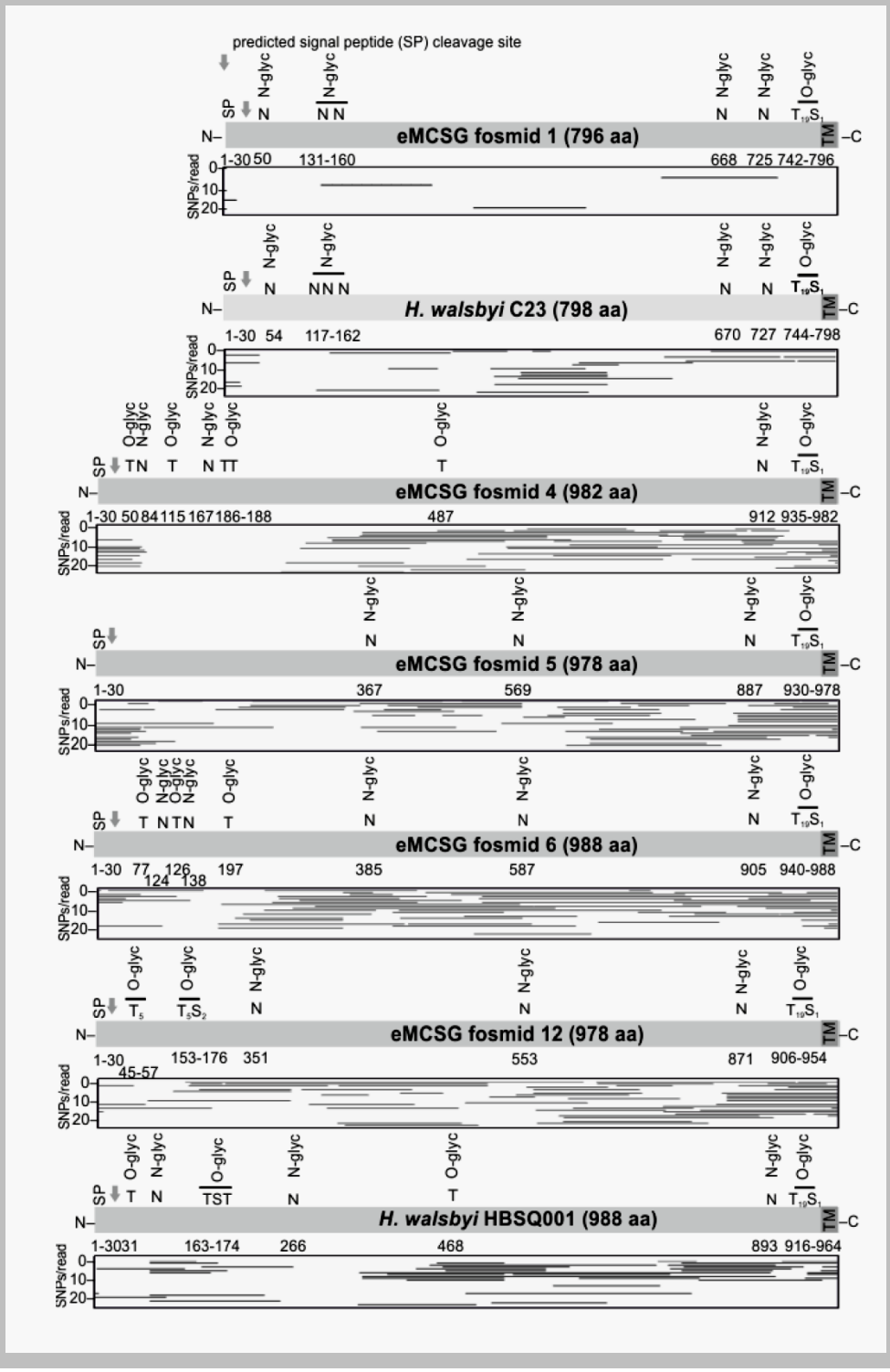


Figura 21. Alineamiento de las proteínas de la capa S (MCSG) presentes en los genomas de *H. walsbyi* C23, HSBQ001 y en los fósmidos secuenciados en este trabajo. Se han incluido en el alineamiento las proteínas de la capa S de *Haloferax volcanii, Halobacterium salinarum, Haloarcula marismortui, y Haloarcula japonica*

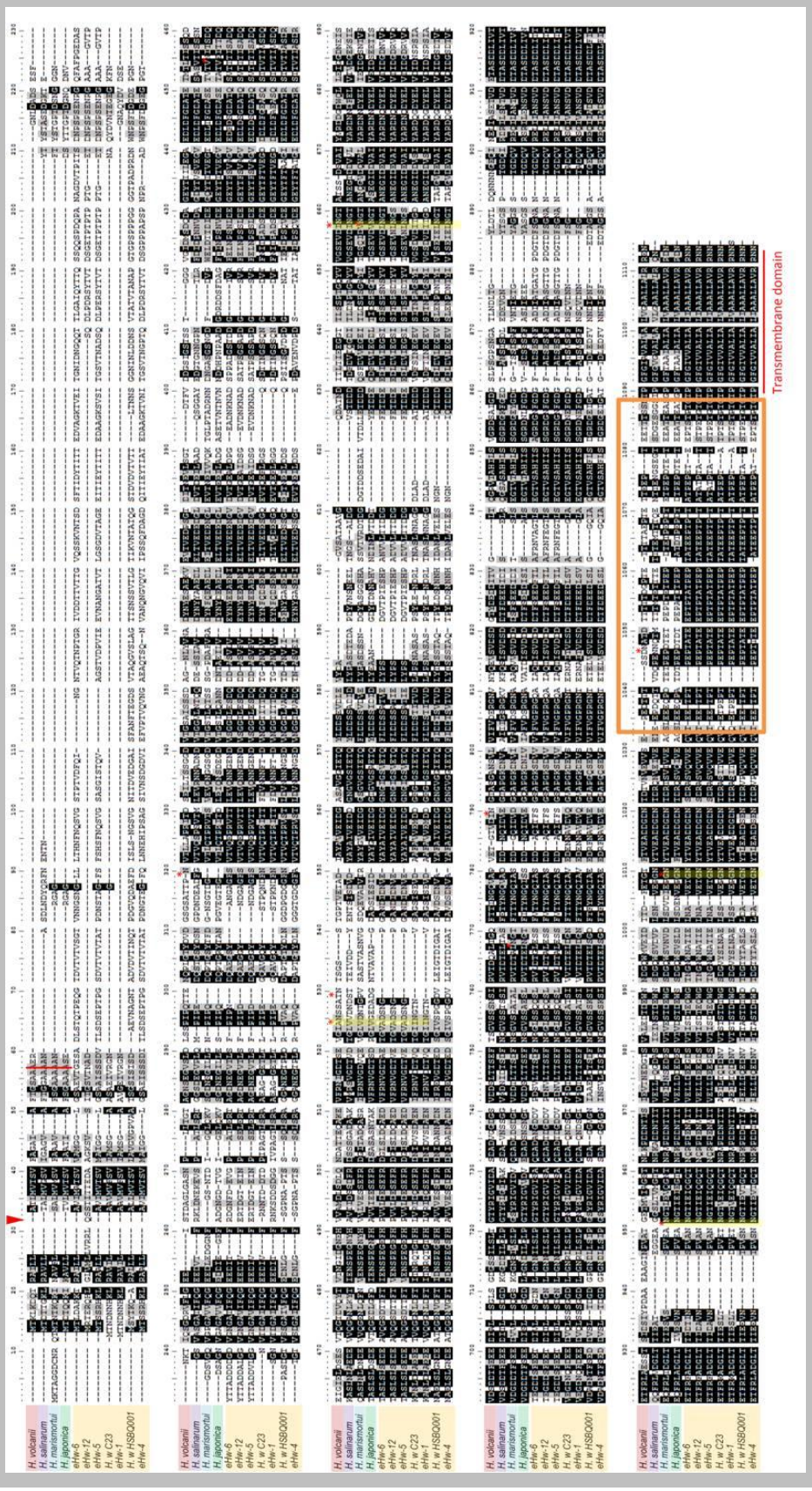


Figura 22. Estructura secundaria de las MCSG de *H. japonica, H. salinarum* comparada con las de *H. walsbyi* C23, HSBQ001 y las encontradas en los fósmidos.

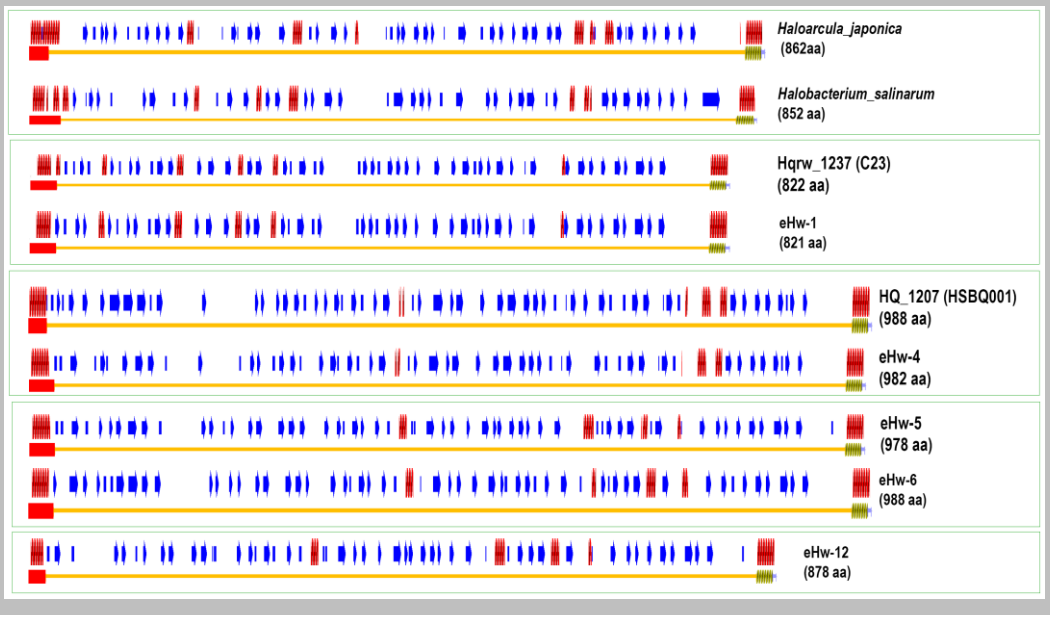
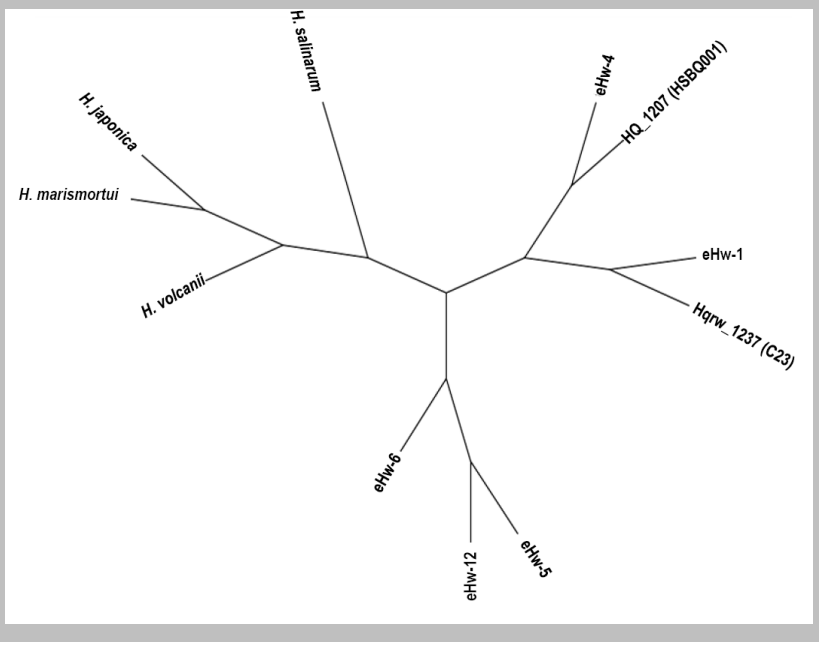


Figura 23. Relación filogenética de las MSCG de distintas haloarchaeas junto con las de los genomas C23 y HSBQ001 de *H. walsbyi* y las MSCG encontradas en los fósmidos.



4.4.2. Otros genes presentes en los fósmidos.

La isla GI1 comienza en todos los casos justo después de una proteína anotada como “dentin-sialofosfoproteína”, la cual está conservada entre todos los fragmentos que tienen secuenciada esa parte. Tras la re-anotación de los genes realizada en este trabajo, esta proteína podría tratarse de otra glicoproteína (CSG) con motivos tipo LPXTG para unirse a la pared celular. Se desconoce su función específica.

En C23, que poseía únicamente dos genes en la GI1, además de la MCSG (Hqrw_1237), se encuentra otro gen, Hqwr_1236, que codifica también para una glicoproteína de pared celular. Ésta está conservada parcialmente en el extremo 5' de las versiones de la GI1 de HSBQ001 (HQ_1193, ver Figura 19), eHw-559, eHw-4 y eHw-7; es decir, está presente en todos los fósmidos en los que está secuenciado este extremo. Excepto en la parte N-terminal donde se llegan a parecer hasta un 87%, su similitud disminuye drásticamente después de los primeros 230 aminoácidos en una región de baja complejidad, rica en repeticiones PTPT. Esta disposición de los genes sugiere un mecanismo sobre cómo podrían haberse originado las versiones más largas de la GI1: el cluster de CSG extras presentes en estas islas de mayor tamaño podría insertarse entre los dos genes presentes de la versión más simple, C23 y eHw1. También, es posible que no trate de una inserción, sino que las versiones cortas de GI1 sean el producto de una reducción de las versiones más largas. Tras un análisis exhaustivo de la zona intergénica entre Hqrw_1236 y Hqrw_1237 (y la región equivalente en eHw-1) no se encontró ninguna repetición o secuencia que pudiese explicar este tipo de proceso génico.

Existen proteínas conservadas en muchas de las GI1, como la subtilisin-proteasa (presente en todas las versiones de la isla menos en C23), el transportador tipo ABC de hierro (III) (presente en todas las versiones de GI1 menos en las de los genomas), proteínas que contienen dominios de inmunoglobulina y las que tienen cierta similitud con la anotada como adhesina en HSBQ001 que se explica más adelante. El orden de los genes e incluso el tamaño de ellos se conservan en muchas ocasiones y es sólo interrumpido por la inserción de elementos IS (como en eHw-12 o eHw-9) o de otras ORFs cortas mencionados anteriormente. Esta conservación apoya la teoría propuesta sobre el origen de las GI1 de mayor tamaño: la inserción en el medio de la GI1 tipo C23 un fragmento que contenía las CSG extras, las proteasas y los componentes necesarios para construir, muy probablemente, una envoltura diferente que proteja la célula. La divergencia observada puede ser producto de la presión selectiva a la que están sometidos estos genes y haber evolucionado rápidamente, especialmente los que pueden ser objetivos de fagos. No obstante, esta variabilidad tendría un estricto límite: la necesidad de construir una pared celular funcional que proteja a las células (esta idea se discute más adelante).

Es interesante que, en todos los fragmentos, se encuentre una CSG larga precediendo siempre la MCSG (Figura 19). En el caso de C23 se trata de Hqrw_1236, cuya homóloga en HSBQ001 es la mencionada anteriormente HQ_1193 (se encuentra en el extremo 5' de la GI1). Sin embargo, en HSBQ001 sigue existiendo otra CSG que supliría esta función y que además está conservada con las otras versiones encontradas en las islas largas (homólogos que comparten hasta un 98% como es el caso de eHw-12 y eHw-5). Esto sugeriría que es probable que ambas proteínas, éstas y la MCSGs, puedan ser esenciales en la formación de la capa S externa (esta disposición de genes no está conservada en ningún otro genoma de archaeas halófilas).

La mayoría de los genes que se encuentran en las IG1 de mayor tamaño son diferentes glicoproteínas, proteínas inicialmente anotadas como adhesinas y proteasas. Entre ellas se encuentra una proteína de gran tamaño, Hmu2 (Halomucina 2) en HSBQ001. Esta proteína comparte dominios con la halomucina gigante Hmu que se encuentra 162 Kpb "aguas arriba" del genoma HSBQ001. Hmu se ha propuesto como un componente estructural en forma de cápsula que ayuda a las células contra la deshidratación y sobrevivir en la extrema salinidad y la alta concentración de magnesio de su ambiente natural (Burns *et al.* 2007; Legault *et al.* 2006). Proteínas similares a Hmu2 se observa también en eHw-4, eHw-559 y en eHw-7 (una versión truncada). Su funcionalidad es desconocida.

Hacia el extremo 3' de las las proteínas similares a Hmu2 (aunque HSBQ001, eHw-4 y eHw-7 tienen partes en común), todas la GI divergen presentando diferentes conjuntos de proteínas pequeñas, la mayoría de ellas hipotéticas, y elementos móviles IS como en eHw-5, eHw-9 y eHw-12. En eHw-7 la sintenia se pierde después de la proteína truncada similar a Hmu2 y continua con dos genes pequeños que tienen una alta similitud con un terminasa (e-value 10^{-17}) y una proteína hipotética (e-value 10^{-42}) de fago aislada en el mismo cristalizador CR30 seis años más tarde mediante una aproximación metagenómica (Santos, Yarza et al. 2010). Este es un indicador de que esta región del genoma podría servir como un objetivo de fago y facilitar la adquisición de nuevos genes. Otras versiones de las islas comparten sólo un pequeño conjunto de genes con el resto, como eHw-9 y eHw-6, la cual sólo comparte la MCSG y la zona conservada fuera de la isla.

Otros genes presentes en las islas son aún más variables. Dos de los genes menos conservados se encuentran en eHw-7 y son dos putativas adhesinas con un dominio PKD implicado en unión celular (genes con trama roja en la Figura 19). Este dominio se identificó por primera vez en la proteína Policistina-1, la cual es una glicoproteína de superficie celular de gran tamaño involucrada en la interacción proteína-proteína y proteína-carbohidrato. Los dominios PKD es normal encontrarlos en proteínas extracelulares implicadas en la interacción con otras proteínas y también en proteínas de superficie de arqueas de ambientes extremos (Jing, Takagi et al. 2002). Es posible que todas ellas seas proteínas implicadas en la síntesis de la capa S o cualquier otro componente de la envoltura celular (Dyall-Smith *et al.* 2011; Lechner & Sumper 1987; Sumper *et al.* 1990).

Otras proteínas muy poco conservadas son aquellas anotadas como adhesinas por una baja similitud con la anotada como adhesina en la isla de HSBQ001 (HQ_1205) (Figura 19). La GI1 de C23 carece de esta proteína y la única proteína adhesina del genoma está muy lejos de esta región del genoma. Esto sugiere que esta proteína, junto con la CSG presentes, pueden no ser necesarias para la construcción de la capa S, pero sí quizá son necesarias para construir la estructura trilaminar observada en HSBQ001 (Burns, Camakaris et al. 2004). Con la excepción de las adhesinas de eHw-5 y 12, que comparten un 79% de similitud (eHw-12 y 9 sólo el 40%), el resto de ellas son muy diferentes a HQ_1205. Todos ellas tienen un dominio de unión de calcio en el C-terminal (dominios EF) y un péptido señal en el extremo N-terminal, por lo que se supone han de ser proteínas secretadas (Figura 24). Se sabe que para estabilizar la capa S, son necesarios cationes divalentes como puede ser el calcio, luego este dominio puede jugar un aspecto importante en este aspecto (Lechner & Sumper 1987). Las versiones presentes en eHw-559, eHw-6 y eHw-4 tienen un dominio tipo-1 inmunoglobulina (Figura 24) presente en intiminas, proteínas de superficie bacterianas y de origen vírico. En general, las intiminas son moléculas de adherencia celular que median la interacción hospedador-célula y pueden contener tres dominios, dos tipo inmunoglobulina y uno tipo lectina implicado en el reconocimiento de carbohidratos. También relacionado con el reconocimiento de carbohidratos, HQ_1205 y las versiones de esta proteína en eHw-4 y eHw-6 poseen repeticiones beta-hélices paralelas típicas de enzimas cuyos sustratos son polisacáridos. HQ_1205 y la presente en eHw-9 tienen también un dominio carboxipeptidasa presente en otras proteínas de superficie. De nuevo, una posible explicación de por qué estas putativas adhesinas se encuentran tan poco conservadas podría ser debido a que éstas son dianas comunes de los fagos. Aumentando la diversidad de estas dianas, la probabilidad de ser detectado por un fago disminuiría, y en el caso de que uno de estos clones pudiese ser reconocido e infectado, se aseguraría la continuidad de la especie tal como se observa en el cristalizador.

Otro de los componentes abundantes en la GI1 son las proteasas. Las que se encuentran en los fragmentos secuenciados son muy variables (hasta 7 versiones diferentes). Éstas podrían estar jugando un papel crucial en la eliminación del péptido señal de las CGS secretadas al exterior. Sin embargo, puesto que no están presentes en el genoma de C23, su papel no debe ser esencial para la biosíntesis de la capa S. Un dato que apoya la idea de la mayor complejidad de la pared celular de HSBQ001, es también el mayor número de proteasas implicadas en la degradación de proteínas de pared que se encuentran en el genoma de HSBQ001. De las 12 proteasas descritas, 4 de ellas no están conservadas en C23. HQ_1200 y HQ_1208 son dos subtilisin-serin proteasas que se encuentran en la GI1 y se parecen un 67%. La primera de ellas no tiene péptido señal (por lo que puede no ser funcional) y están conservadas en 5 de los fósmidos secuenciados. Las otras dos proteasas no conservadas en C23 son HQ_2153 y HQ_3284. La primera de ellas está anotada como una proteasa del tipo "Lon", implicada en la expresión de genes relacionados con la biogénesis de la pared celular de *Pseudomonas syringae*. La segunda, HQ_3284, codifica para la proteína HtpX, una Zinc metalo-proteasa que participa en la degradación de proteínas de membrana en *E. coli* (Putaporntip *et al.* 2009). En el genoma de C23 hay 7 proteasas y tan sólo Hqrw_2698, una subtilisin-serin proteasa es única respecto a HSBQ001. Justamente, esta proteasa está en una zona de bajo reclutamiento genómico del genoma en C23 (datos no mostrados), lo que podría sugerir que podría estar realizando el mismo tipo de función que la subtilisinas que se encuentran en la GI1 de HSBQ001, endopeptidasas que eliminarían el péptido señal de la MCSG y de otras CSGs que participan en la formación de la envoltura celular. En otros genomas de haloarchaeas, como *H. volcanii* o *H. salinarum*, no se encuentran proteasas cerca de la proteína de la capa S. También en relación con la actividad proteolítica, en la isla se encuentra algunas proteínas con un dominio "Glug" (motivo conservado G-L-X-G), presente en las metalo-peptidasas de inmunoglobulina A (IgA), las cuales se unen al peptidoglicano de la pared celular mediante un enlace amida.

De manera general, estos resultados indican la presencia de varias versiones de la GI1 en la población de *H. walsbyi* coexistiendo en un mismo punto temporal y espacial. Esta divergencia puede ser debida a una fuerte selección por tratarse de putativas dianas de reconocimiento de fagos. Además, que estos clones son ubicuos y muy probablemente están extendidos en todos los hábitats en los que puede crecer *H. walsbyi*. Prueba de ello es la presencia de una secuencia prácticamente idéntica de la GI1 en Santa Pola que a la GI1 de C23, aislado varios años después en Australia.

Figura 24. Dominios encontrados en las proteínas anotadas como adhesinas en los genomas de *H. walsbyi* y en los fósmidos secuenciados.

4.5. Reclutamientos metagenómicos de los fósmidos eHw.

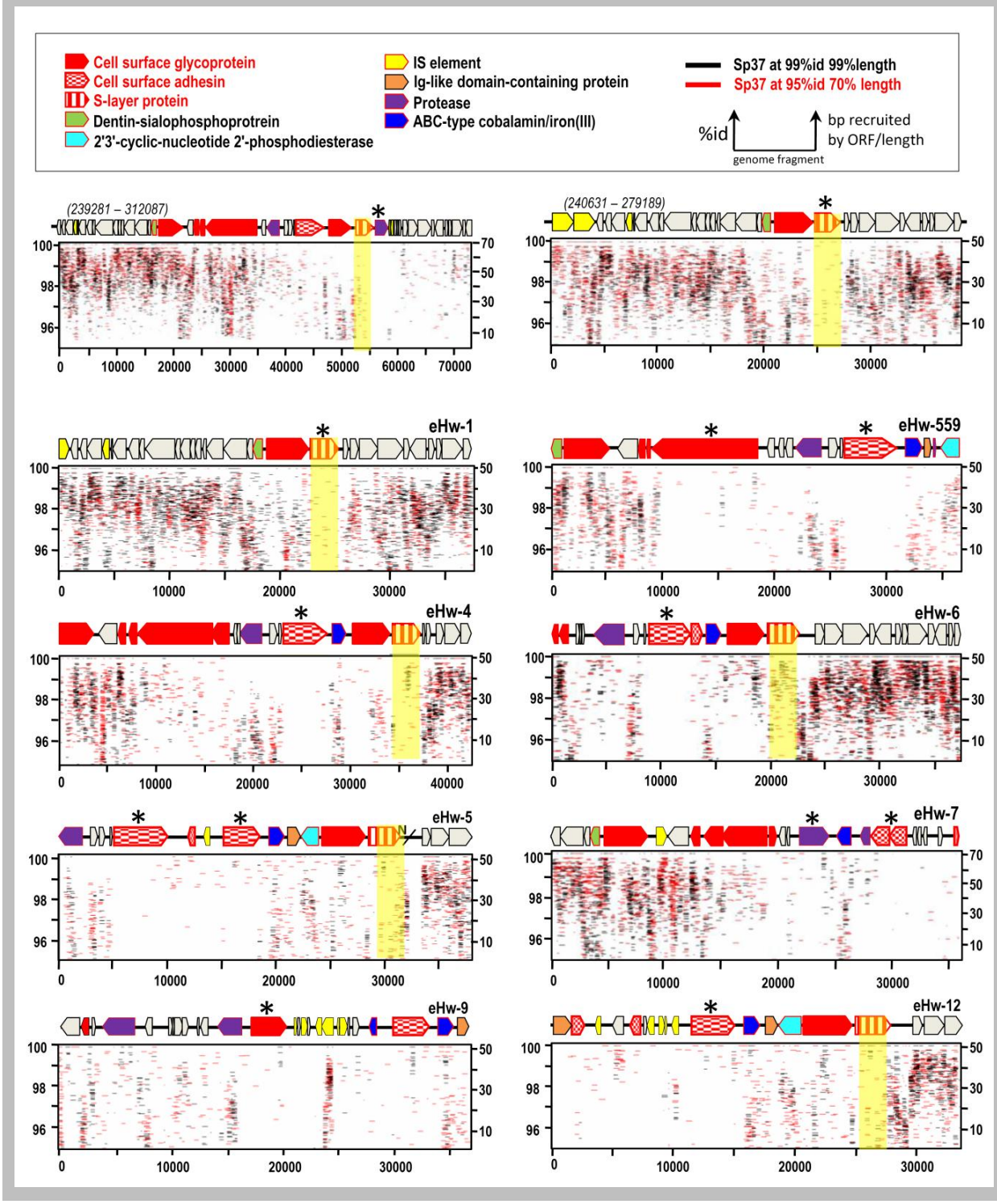
Para averiguar la presencia de diferentes linajes de *H. walsbyi* portadores de las diferentes versiones encontradas de GI1, llevamos a cabo ensayos de reclutamiento usando el metagenoma SS37 (Ghai *et al.* 2011), obtenido del mismo sitio que el genoma HSBQ001 y de la biomasa con la que se construyó la primera librería de fósmidos. SS37 es un metagenoma de más de 1 Gb de secuencias obtenidas por pirosecuenciación (Roche) de una muestra recogida en el 2007.

Los resultados muestran que, a pesar de que en un número muy bajo, todos los genes de las diferentes GI1 secuenciadas reclutaban un número significante de secuencias ambientales, incluso al 100% de identidad en un 100% de su longitud, indicando por tanto su persistencia en la comunidad del cristalizador (Figura 25). Aunque el reclutamiento era significativamente menor que los genes que pertenecían al core-genoma (las proteínas ribosomales de HSBQ001 reclutaban de 5 a 20 veces más (cobertura de 40.47x)), es indicativo de que una parte significativa de los clones presentes en este ambiente son portadores de genes casi idénticos a los encontrados en las GI1 secuenciadas (95% identidad en un 70% de la longitud de la secuencia ambiental). Sin embargo, este reclutamiento no es constante a lo largo de las islas y algunos genes reclutan más que otros (ver por ejemplo los datos de reclutamiento referidos a las MCSG en la Tabla 2). Este hecho se puede también en la Figura 25, donde se han marcado con una sombra amarilla aquellos genes que carecían de homólogos en el metagenoma SS37 y cuya cobertura era nula o casi nula. En todas las versiones de la GI1, hay un gen (o dos) que recluta muy poco o nada. 8 de éstos 10 genes que no reclutan codifican las posibles adhesinas antes descritas. Estos resultados indican por tanto que la mayor parte de genes están presentes en la comunidad del cristalizador, pero que existen algunos que o bien ya no existen o que están muy modificados con respecto a las versiones secuenciadas en los fósmidos y en los genomas disponibles. También puede ocurrir que esa versión del gen esté presente pero de manera muy diluida y que sería necesario seguir secuenciando más para poder detectarla.

Curiosamente, uno de los genes que reclutan menos a alta similitud es la MCSG de C23 (Hqrw_1237) y eHw-1 (la cobertura de estos genes es de 3.01 y 3.10 respectivamente, Tabla 2). La MCSG de C23 y la del clon eHw-1 son tan similares (98% identidad) que se reclutarían una a la otra en los valores usados de similitud. Sin embargo, las otras MCSG tienen una mayor representatividad en el metagenoma con un mayor ratio: desde 7.66 la presente en eHw-4 hasta 16.20 la de eHw-6 (ver Tabla 2). Una posible explicación de esta diferencia puede venir dada del hecho de que el aislado español tiene una capa extra de la que el aislado australiano carece.

Una de nuestras hipótesis de partida es que existirían diferentes clones de *H. walsbyi* que cohabitan en el mismo ambiente con envolturas ligeramente diferentes unos de otros. Esto permitiría a la comunidad de *H. walsbyi* una continuidad en caso de que alguno de estos clones pudiese ser reconocido por un fago (Rodriguez-Valera *et al.* 2009). De acuerdo con esta idea, podría ocurrir que uno de los componentes más externos (mayor probabilidad de entrar en contacto con un fago) sería el más divergente para evitar el reconocimiento de los fagos. Ya que C23 no posee la capa extra que presenta el aislado español, la MCSG de C23 habrá evolucionado de manera más rápida, divergiendo por tanto mucho más que las otras. A cambio de esa protección extra, aquellos clones similares a HSBQ001 tendrán GI1 más largas y complejas necesarias para construir la doble capa.

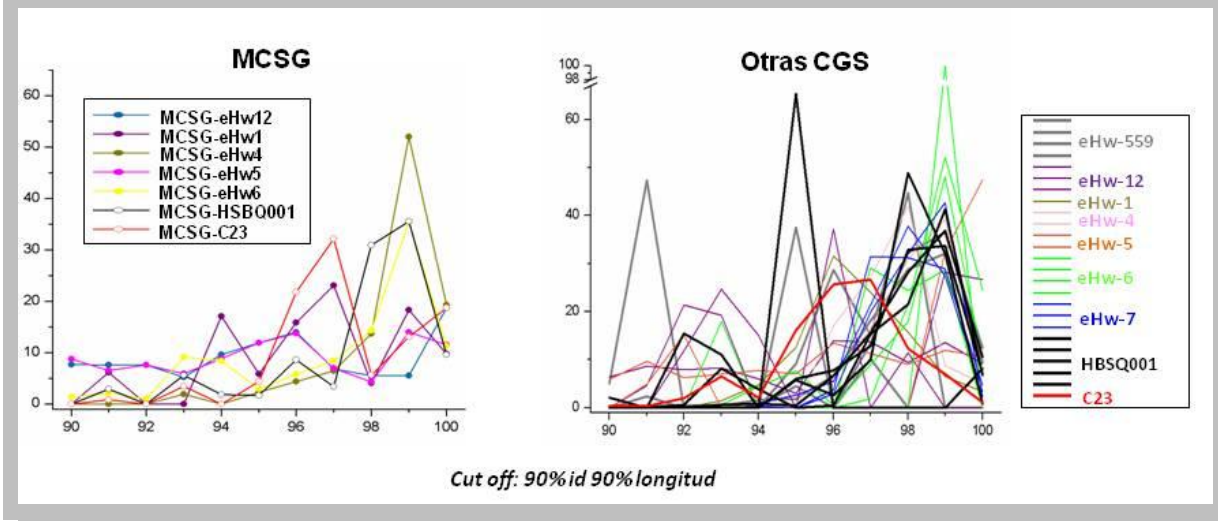
Figura 25. Reclutamientos metagenómicos de los fósmidos eHw y el metagenoma del cristalizador CR30 SS37.



Para chequear si el ratio de variación de los genes de la isla era mucho mayor que el observado en los genes del core-genoma, representamos la identidad de los fragmentos homólogos encontrados en el metagenoma frente a la frecuencia en porcentaje de bases reclutadas usando un estricto corte de 90% de identidad en un 90% de su longitud (Figura 26). En el caso de los genes housekeeping (*gyrA, gyrB, secY, atpB, tet2* y *rpoB1*) la gráfica de la Figura 17 indica que la mayor parte de los homólogos ambientales en todos los casos son 98% similares (96.12% y 91.49% para HSBQ001 y C23 respectivamente) y su número decrece rápidamente sobre el 95%. Sin embargo,

esta tendencia no se conserva para la mayoría de los genes de la isla (Figura 26). A pesar de que el número de secuencias ambientales encontradas con más de 95% es bajo, (11.34% del total de las secuencias reclutadas en las GI1), sólo 57.53% de ellas tienen una identidad mayor del 98%. En el caso de la MCSG (Figura 26, gráfica izquierda), sólo la mitad de las secuencias que reclutan a este nivel lo hacen sobre el 98% de identidad (50.24% de las secuencias). Estos resultados sugieren que el número de mutaciones acumuladas en estos genes es mayor que para los genes *"housekeeping"*. Sin embargo, si comparamos el número de secuencias que reclutan al 100%, encontramos una variación para los housekeeping entre el 17.8-39.12% para C23 y 27.5-53.75% para HSBQ001, mientras que para la MCSG, este rango varía entre el 9.6-19.27%, lo que no está lejos de los genes bien conservados del core-genoma. Así que, aunque el número de homólogos ambientales de los genes presentes en las islas es mucho menor que para los genes que pertenecen al core-genoma, el porcentaje de secuencias exactamente iguales no es muy diferente, sugiriendo una variabilidad similar y que por lo tanto, los genes de la isla estarían presentes en un número mucho menor de linajes clonales.

Figura 26. Reclutamientos metagenómicos de la MCSG y de otras CSGs presentes en los genomas y fósmidos secuenciados.



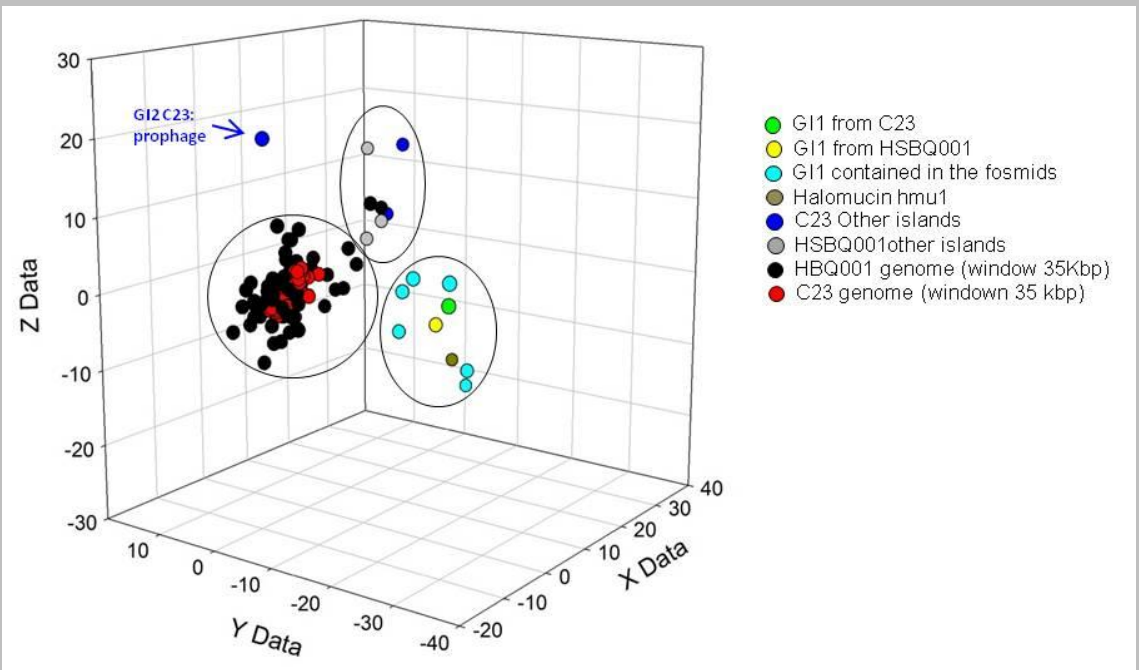
Quisimos comprobar si algo similar podía ocurrir en otras salinas, y para ello, realizamos los mismos experimentos de reclutamiento usando un metagenoma obtenido en el cristalizador de las salinas de San Diego (California, USA), en el que *H. walsbyi* está también presente pero no tan abundantemente (Rodriguez-Brito *et al.* 2010). Tan solo un 1% de las secuencias ambientales daban *"hit"* con *H. walsbyi* por encima del 95% de identidad en al menos un 70% de su longitud. Encontramos que ambos genomas reclutaban en la misma proporción al 100% de identidad (un total de 108 Kpb C23 y 115 Kpb HSBQ001). Ya que el número de secuencias es muy bajo para todo el genoma, no fue posible encontrar una sola secuencia que reclutase dentro de la isla GI1.

4.6. Frecuencia de tetranucleótidos y uso de codones de las proteínas presentes en GI1.

4.6.1. Análisis de la frecuencia de tetranucleótidos.

Se han descrito las islas genómicas como "hot-spots" o puntos calientes de los genomas donde es frecuente que ocurra una recombinación nucleotídica permitiendo la adquisición de nuevos genes y por tanto, dar lugar a las variaciones existentes entre un linaje u otro de una misma especie. Tal como se explicó en la Introducción del presente trabajo, la adquisición de genes se ve facilitada por HGT mediada por fagos y otros elementos genéticos móviles (Langile MG Detección de 2010, Kettler Patrones de GC 2007). Una de las metodologías existentes para detectar regiones transferidas mediante HGT es medir la composición inusual de nucleótidos en esa zona del genoma. Las diferencias en la frecuencia de tetranucleótidos han sido usadas previamente para identificar islas genómicas que son el resultado de una HGT (Reva & Tummler 2008). Los genomas C23 y HSBQ001 se dividieron en fragmentos no solapantes de 37 Kpb, que aproximadamente, era el tamaño medio de las GI1 secuenciadas. Previamente, se extrajeron todas las islas genómicas tal como fueron definidas en Legault *et al.* (2006). Para cada uno de estos fragmentos se calculó la frecuencia nucleotídica y se construyó una matriz para hacer un análisis de componentes principales (ver Materiales y Métodos). El resultado de este análisis se muestra en la Figura 25. Mientras que la mayoría de los fragmentos pertenecientes a los genomas se encuentran juntos, todas las islas GI1 tienden a agruparse junto con un fragmento que pertenece a la halomucina Hmu. El resto de islas de los genomas también aparecen separadas en otro grupo diferente. Esto demuestra que estas regiones, aunque tienen un contenido de GC muy similar que el genoma, poseen una marca de genómica que los hace diferentes.

Figura 27. Análisis de la frecuencia de tetranucleótidos de los genes presentes en los genomas de H. waslbyi C23 y HSBQ001 y los presentes en los fósmidos secuenciados.

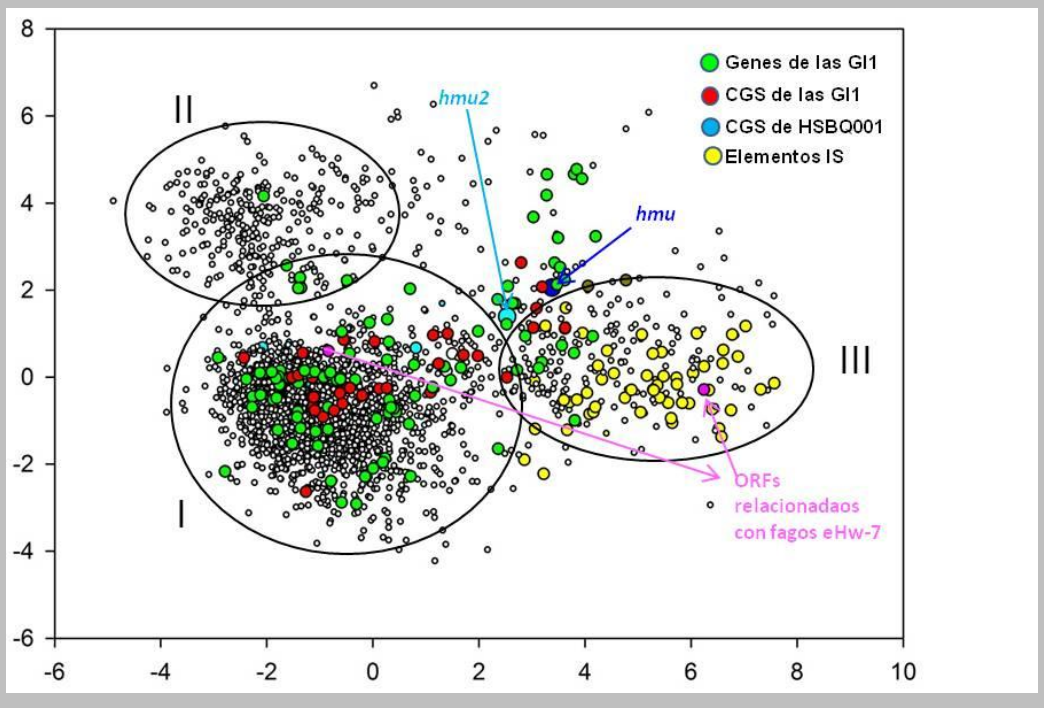


El hecho de que las GI1 posean una frecuencia tetranucleotídica diferente a la del resto del genoma, pero sin embargo un contenido en G+C semejante al resto de genes del cromosoma, podría sugerir un origen por HGT siempre que el dador del DNA tenga también un G+C muy similar al de la célula aceptora. Ya que el organismo mayoritario en el cristalizador con un G+C bajo es *H. walsbyi*, podríamos pensar que existe recombinación entre los diferentes linajes de clones coexistentes (ver más adelante).

4.6.2. Uso de codones.

Comprobamos también si el uso de codones de los genes de las islas era diferente al del resto del genoma, lo que podría dar lugar a una velocidad de traducción diferente. Se calculó el uso de codones de cada uno de los genes para los genomas y las GI1 secuenciadas y se construyó con su frecuencia relativa una matriz con la que llevó a cabo un análisis de multicomponentes. (Figura 26). En el gráfico del análisis de PCA fue posible detectar tres grandes nubes de puntos. El mayor de ellos (I) contiene a la mayoría de los genes de ambos genomas y aquí se encuentran también las proteínas ribosómicas (datos no señalados). Junto a esta nube de puntos hay un otra formada principalmente de genes que codifican proteínas con más de 3 dominios transmembrana (II). La última, está formada principalmente por elementos IS y los genes que tienen un contenido GC alto, junto muchas proteínas hipotéticas y algunas de las adjudicadas a fagos (III). La mayoría de los CSG de las GI1 (puntos rojos) se agrupan en el grupo I, junto con la mayoría de los genes del genoma. Sólo algunos de los genes presentes en las GI1 se agrupan en la nube II y otros poco aparecen dispersados entre las tres nubes. Los resultados de este experimento apoyarían la idea de que los genes de las GI1, al menos la mayor parte de las CSG presentes, podría poseer la misma tasa de transcripción que cualquiera de los otros genes del genoma (de hecho, en muchas de ellas se ha detectado el sitio de unión al ribosoma (RBS) aguas arriba del extremo 5' de gen, Figura Anexo III).

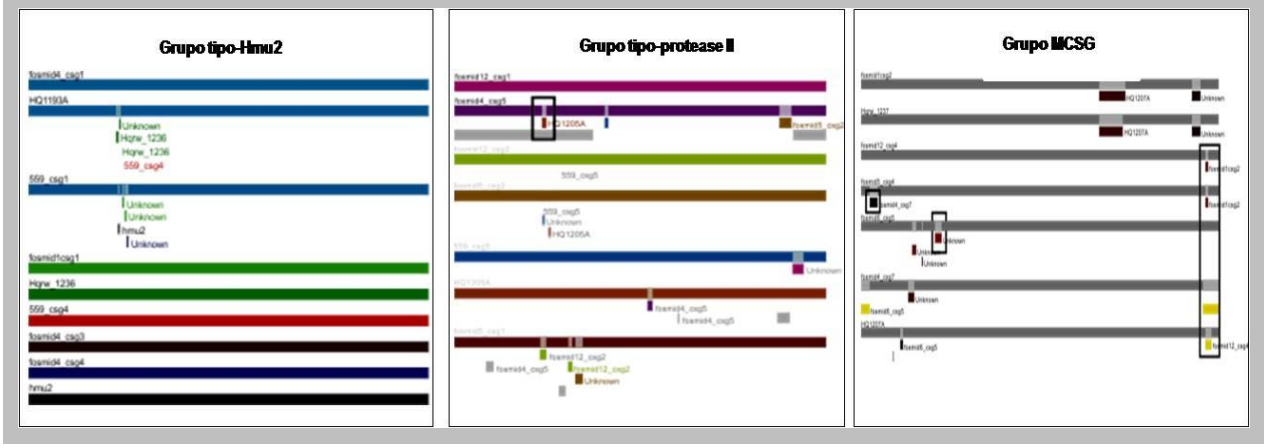
Figura 28. Uso de codones de los genes presentes en los genomas de H. waslbyi C23 y HSBQ001 y los presentes en los fósmidos secuenciados.



4.7. Estudio de posible recombinación entre las diferentes GI1 descritas.

Para estudiar posibles fenómenos de recombinación entre las islas GI1, usamos el programa "Recombination Detection Program" RDP3 (Martin *et al.* 2010) con algunos de los genes que se podían alinear de una manera coherente: MCSGs, parte de las proteínas tipo Hmu2, el transportador ABC, y una de las proteasas (Figura 29). Se eligió este programa porque usa varias metodologías diferentes integradas en el mismo *software para* detectar fenómenos de recombinación en un alineamiento de al menos de más de tres secuencias introducido. Los métodos que se inlcuyen son: BOOTSCAN (Salminien et al., 1995; Martin et al., 2005b), GENECONV (Padidam et al., 1999), Maximum Chi Square (MAXCHI; Maynard Smith, 1992; Posada and Crandall, 2001), CHIMAERA (Posada and Crandall, 2001), Sister Scanning (SISCAN; Gibbs et al., 2000), 3SEQ method (Boni et al., 2007), Reticulate compatibility matrix (Jakobsen and Easteal, 1996), VisRD (Lemey et al., 2009) y TOPAL DSS (McGuire and Wright, 1998, 2000). Para que un fenómeno de recombinación sea veraz, al menos ha de ser detectado por tres métodos diferentes con valores mayores de e-value 10^{-5}. En ninguno de los casos posibles se pudo detectar que alguna de las secuencias pudiese ser la recombinación de dos parentales con suficiente robustez. En la Figura 29 se muestran tres de estos ensayos y se han remarcado las posible zonas de los genes fruto de una muy remota recombinación (en ninguno de los caso el e-value fue mayor de 10^{-2} y nunca verificado por mas de dos métodos).

Figura 29. Estudio de recombinación entre alguno de los genes de las IG1. (Gráficas obtenidas directamente de RDP3). Remarcados en los rectángulos se muestran los posibles casos, muy remotos, de posible recombinación.



5.- DISCUSIÓN

5.- DISCUSIÓN

La alta concentración de sales presentes en los cristalizadores de las salinas solares hace imposible la vida excepto para algunos organismos hiperhalófilos y unas pocas células especializadas. Varias aproximaciones moleculares han revelado que la comunidad del cristalizador, donde precipitan estas sales, está dominado en gran parte por la archaea cuadrada *H. walsbyi*, llegando a suponer hasta el 80% de la biomasa total. El primer metagenoma de las salinas construido a partir de la secuenciación de los extremos de los fósmidos de una librería ambiental (Legault *et al.* 2006), demostró la existencia de un gran número de secuencias que aun teniendo el típico G+C bajo de *H. walsbyi,* no poseían ninguna similitud con el genoma secuenciado *H. walsbyi* DSM 16790. Una situación muy similar se ha encontrado al comparar un segundo metagenoma, SS37, mucho mayor que el anterior (Ghai *et al.* 2011) y obtenido mediante pirosecuenciación (454, Roche). La comparación de secuencias ambientales frente a los genomas de *H. walsbyi* como referencias han permitido definir regiones con un notable grado de conservación (100% de identidad nucleotídica) alternadas con regiones de baja o ninguna similitud con la cepa de referencia. Estas últimas se han denominado islas metagenómicas. El grado de variabilidad observado en las diferentes islas es muy alto e indica que la reserva genética de la que *H. walsbyi* puede abastecerse para cumplir con sus requerimientos ecológicos no es pequeña. Se ha propuesto que la variabilidad encontrada dentro de una especie puede reflejar la co-existencia de linajes clonales que difieren en las regiones genómicas adaptativas. Estos clones son estables en el tiempo y sobreviven durante largos períodos sin variaciones (Rodriguez-Valera *et al.* 2009). El contenido de genes de estas islas varía de un linaje a otro y permitirá, entre otros, el reconocimiento específico de un tipo de fagos, o la explotación de un determinado recurso presente en el cristalizador; por lo tanto, estos genes determinarán la abundancia de un determinado clon en un punto temporal siguiendo una dinámica tipo "*kill the winner*". Tradicionalmente se ha considerado que la variabilidad observada dentro de una misma especie se correspondía con la existencia de micronichos. Por ejemplo, la variabilidad observada en *P. marinus*, sí que se corresponde con la existencia de micronichos en la superficie del océano donde sí que co-existen diferentes ecotipos adaptados a las diferentes longitudes de onda de la luz (Coleman 2004). Sin embargo, en un hábitat tan simple como el cristalizador de la salina, donde las fuentes de carbono provienen principalmente del alga *Dunalliela salina*, no es fácil pensar en la existencia de diferentes micronichos y todo el cristalizador puede ser considerado como un conjunto simple.

A pesar de que existen muchos trabajos descritos sobre la variabilidad encontrada dentro de una especie, no existen muchos datos sobre cómo pueden variar esos linajes y muchos menos sobre cuántos pueden co-existir en un mismo punto temporal. Al menos en el cristalizador CR30 de las salinas de Santa Pola, hemos comprobado que el número de clones parece no ser infinito, y que su distribución no es homogénea. Mediante ensayos de reclutamiento con el metagenoma SS37, sabemos que el aislado australiano C23 es de 3,3 veces menos abundante que el español HSBQ001 y el mero hecho de haber secuenciado un fósmido con casi la misma versión de la GI1 que el aislado australiano sugiere que el número de clones es limitado, que son ubicuos y que son estables en el tiempo (al menos en un rango de 5 años de diferencia). La homogeneidad de *Haloquadratum* presente en todo el mundo, especialmente la referida a la conservación de los genes 16S rRNA, cuya variación

nunca es superior al 2%, puede deberse a la combinación de diversos factores. Entre ellos, se pueden citar un sistema de dispersión global, reservas naturales, y una fuerte presión de selección. Los lagos naturales de sal con frecuencia se secan por completo, y los cristalizadores de las salinas marinas solares se drenan y se rellenan cada año. La dispersión de *Haloquadratum* través de los océanos es poco probable, ya que este organismo no puede formar esporas, no crece a concentraciones de sal inferior del 14% w/v, y se lisa en bajas concentraciones de sal (2% w/v). Sin embargo, podría tener lugar la supervivencia dentro de las inclusiones fluidas de los cristales de sal. Esto se ha demostrado para otras haloarchaeas (Fendrihan *et al.* 2006; Fendrihan *et al.* 2009), y podría permitir la dispersión por el viento o por las aves migratorias que frecuentan las aguas hipersalinas. De hecho, se han encontrado haloarchaeas como flora normal de las glándulas de sal de una especie de ave migratoria (Brito-Echeverria *et al.* 2009). La capacidad para hacer frente a las condiciones extremas también pueden suponer fuertes presiones selectivas que limitan la divergencia dentro de una especie, y *Haloquadratum* pueden estar mejor adaptada a estos extremos diferentes. Por ejemplo, además de crecer bien en soluciones de saturación de NaCl, *Haloquadratum* tolera las altas concentraciones de magnesio que se producen en los cristalizadores después de la precipitación del NaCl (Bolhuis *et al.* 2004; Bolhuis *et al.* 2006). De hecho, la cepa C23 alcanza mayores densidades de células en concentraciones mayores de 1 M de $MgCl_2$ (Burns *et al.* 2007). Además de en *H. walsbyi* también existen trabajos basados en la secuenciación de genes *"house-keeping"* y 16S rRNA que describen también una alta variabilidad en otros procariotas de este hábitat de baja diversidad, como es el caso de diferentes aislados de *Halorubrum* sp. (Papke *et al.* 2004) y diferentes aislados de *S. ruber* (Pena *et al.* 2010).

Este trabajo se ha centrado en el estudio de la isla metagenómica I (GI1), enriquecida en genes que codifican componentes de la envoltura celular en HSBQ001. Desafortunadamente, en *H. walsbyi* se sabe muy poco sobre la naturaleza de la capa S que rodea a las células aunque, debido a su forma cuadrada, se ha especulado en numerosas ocasiones sobre la necesidad de un mayor número de genes implicados que para otras haloarchaeas. Tras la secuenciación del segundo genoma de *H. walsbyi*, C23, se ha visto que esto no es así ya que la GI1 de esta cepa contiene únicamente dos genes y es igualmente cuadrada. Sin embargo, sí que varía en el número de láminas exteriores observables al microscopio electrónico. Tras la selección de extremos de fósmidos que flanqueasen esta región del genoma tomando como referencia el genoma HSBQ001, se procedió a secuenciar 9 fósmidos provenientes de la librería metagenómica construida en el 2004-5 que cubriesen la zona del genoma GI1. Siete de ellos contenían nuevas versiones de genes implicados en la síntesis de la envoltura celular, presentando un cluster de CSG diferente en cada caso. Mediante ensayos de reclutamiento con el metagenoma SS37, hemos encontrado que las nuevas versiones de GI1, junto con las anteriormente descritas (las de los genomas C23 y HSBQ001, y la del fósmido eHw-559) están presentes en el metagenoma, ya que hay secuencias idénticas al 100% de identidad. Sin embargo, la cobertura a lo largo de cada uno de los fragmentos secuenciados no es homogénea, y es frecuente encontrar siempre al menos un gen, especialmente aquellos anotados como posibles adhesinas, que no reclutan ninguna secuencia ambiental, ni siquiera a baja similitud. Una de las posibles explicaciones de este hecho es que el metagenoma usado, SS37, a pesar de ser dos órdenes de magnitud mayor que el primero usado, sigue estando por debajo de las condiciones de saturación del sistema, es decir, no es representativo de toda la variabilidad presente en CR30. De hecho, y aunque no se muestra en este trabajo, ha

sido necesario ampliar el metagenoma de este cristalizador con 4 Gpb más de secuencia para poder llegar a condiciones de saturación. Aun así, determinados genes, o determinados dominios de los mismos, siguen sin reclutar secuencias ambientales en absoluto, indicando que ese clon en concreto, ha desaparecido o se ha diluido muchísimo en la comunidad procariótica en el momento de muestreo, impidiendo por tanto su detección a través de una aproximación metagenómica.

Las aguas hipersalinas (sobre un 30% w/v), *H. walsbyi* alcanza densidades de población de hasta 10^8 células /ml e incluso mayores niveles de partículas de virus (10^9 VLP/ml) (Guixa-Boixereu 1996; Pedros-Alio *et al.* 2000). Los virus se saben que ejercen una enorme presión en la evolución de los procariotas y en estos ambientes son los mayores predadores de haloarchaeas. Tras el análisis del contenido génico de cada uno de los fósmidos, los resultados indican que los componentes de la pared celular muestran un nivel muy alto de diversidad dentro de poblaciones de *H. walsbyi* concurrentes. De acuerdo con el modelo de "diversidad constante" propuesto por Rodriguez-Valera *et al.* (2009), la coexistencia de varias versiones de esta región ayudaría a distribuir entre los diferentes clones de la comunidad la probabilidad de ser reconocido e infectado por un tipo de fago que reconozca alguna de las estructuras expuestas específicas de clon o linaje. Por lo tanto, en *H. walsbyi*, esta variabilidad refleja una evasión de los fagos mediante el mantenimiento de la microdiversidad. Es fácil imaginar que los efectos de una población tan alta de fagos sobre una comunidad procariota tan poco diversa como la existente en el cristalizador, podrían ser catastróficos sin una estrategia de evasión que promueva la variación de los sitios de reconocimiento de fagos. Luego para explicar la coexistencia de tan alto número de fagos y de células de *H. walsbyi*, debe de existir una alta diversidad de genes de reconocimiento de fagos, muy probablemente las CSGs localizadas en las islas genómicas. Las descritas en el presente trabajo podrían ser una pequeña representación de cómo pueden llegar a variar estas dianas víricas.

La variabilidad de las CGS observadas también podría explicarse sin la necesidad de que co-existan múltiples linajes invariables cuya frecuencia varía en el tiempo. Puede ocurrir que la composición genómica de GI1 cambie continuamente y que las versiones secuenciadas del 2004 ya no existan más en el cristalizador, sin embargo, el hecho de haber encontrado un clon igual eHW-1 que la GI1 de C23, indicaría lo contrario. También podría ocurrir que las diferentes versiones de la GI1 que observamos en estos momentos sean el producto de la selección ocurrida una vez y que las diferencias que vemos son el acúmulo de sucesivas mutaciones puntuales en los genes de la isla. Entonces, la razón de la diferencia de reclutamiento de secuencias del metagenoma entre un gen y otro de la isla se debería a diferencias en la tasa de mutación, siendo mayor en aquellos genes que se encuentren bajo selección positiva por alguna razón, como por ejemplo: ser las dianas de fagos. Han de existir entonces mecanismos moleculares que permitan una tasa de variación extremadamente alta en esta región genómica. Cómo pueden haber variado los linajes existentes, o cómo lo pueden estar haciendo los genes concretos que no reclutan en el metagenoma es controvertido (incluso las metodologías existentes para la detección de islas genómicas es muy heterogéneo (Che *et al.* 2011; Hasan *et al.* 2012; Soares *et al.* 2012)). En las islas genómicas, sobre todo las relacionadas con patogenicidad, es frecuente encontrar cambios en el contenido G+C y fenómenos de transferencia horizontal o de recombinación mediados por la presencia de elementos móviles o tRNA en los límites de la isla u otros elementos repetidos donde tienen lugar estas recombinaciones e inserciones/deleciones.

De hecho, la recombinación del cluster de genes que codifican para el LPS (los genes *rfb*) en *Salmonella enterica* es un paradigma de esta estrategia (Wang *et al.* 2002; Xiang *et al.* 1994). Se ha mostrado en muchas ocasiones que muchos de los genes presentes en las islas genómicas son adquiridos mediante transferencia horizontal (ver por ejemplo el caso de *P. marinus* y *S. ruber* (Avrani *et al.* 2011; Pena *et al.* 2010)). Sin embargo, estos elementos que propician variabilidad no siempre aparecen en todas las islas y encontrar un mecanismo de variación no es sencillo. Tal es el caso de la GI1 en HSBQ001, que contiene un G+C muy cercano al de la media del genoma, carece de elementos transponibles o tRNAs y tampoco se han detectado repeticiones. Entre las nuevas versiones de la GI1 derivadas de este trabajo, se han encontrado elementos IS en cuatro de ellas además de dos genes pertenecientes a fagos (en Hw-7), por lo sugiere que un posible método de variación son estos elementos móviles. Ya que algunos de los genes CSGs parecen ser parálogos, y varios dominios o módulos están conservados, la recombinación entre las diferentes copias podría dar lugar a la variabilidad observada, con un menor riesgo por tanto de generar células no viables por no poder construir envolturas celulares funcionales. Sin embargo, no se ha podido detectar ninguna recombinación entre las islas disponibles (en parte porque casi no disponíamos de genes que pudiesen ser alineados, misma razón por la que fue imposible construir árboles filogenéticos que indicasen esa transferencia exógena). Cabría esperar, que con una mejor representación de esta región genómica y un mayor número de experimentos *in silico* de este tipo, se podría extraer una conclusión más robusta que la aquí presentada. Tan sólo los datos de la frecuencia de tetranucleótidos indicarían un posible origen mediante HGT; sin embargo la conservación del G+C en todas las islas y del uso de codones igual que el de los genes del resto de genoma, indicarían lo contrario. Además, algo difícil de explicar sería de qué organismo son transferidos estos genes y si esta transferencia es en bloque, es decir, la GI1 entera. o se trataría únicamente de genes individuales. Los resultados aquí presentados para intentar esclarecer el origen de la variabilidad de la GI1 en *H. walsbyi* no son concluyentes, y sería necesario disponer de mayor cantidad de datos genómicos. De hecho, este trabajo se ha continuado mediante la obtención de un metagenoma nuevo de 4 Gpb más de secuencia, y resultados preliminares parecen sugerir que, en efecto, existen genes que o han variado drásticamente en un espacio de tiempo muy corto (cuatro años) o que ese linaje en concreto se ha diluido hasta ser imperceptible con una aproximación metagenómica. Una aproximación molecular, como la ampliación de estos genes mediante PCR en varias muestras temporales del mismo cristalizador, ayudarán a comprender mejor cuál de las dos opciones es la correcta.

Como mensaje final y de carácter más general, hemos de replantear de nuevo la pregunta de cuál es realmente la unidad evolutiva en la naturaleza que obedece a la selección natural, ¿los genes, los individuos, los grupos…?. En el caso de procariotas, ahora sabemos que una especie puede ser definida por el repertorio de genes en su totalidad, es decir su pan-genoma. Éste no está incluido en una sola célula, sino repartidos en linajes independientes (o líneas clonales). Además, se debería añadir a este conjunto la comunidad de fagos que pueden reconocer estos clones y por tanto, moldear el contenido del pan-genoma de la especie procariota.

6.- BIBLIOGRAFÍA

6.- BIBLIOGRAFÍA

Abby, S. and V. Daubin (2007). "Comparative genomics and the evolution of prokaryotes." Trends Microbiol **15**(3): 135-141.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller*, et al.* (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucleic Acids Res **25**(17): 3389-3402.

Anton, J., E. Llobet-Brossa, F. Rodriguez-Valera and R. Amann (1999). "Fluorescence in situ hybridization analysis of the prokaryotic community inhabiting crystallizer ponds." Environ Microbiol **1**(6): 517-523.

Anton, J., A. Oren, S. Benlloch, F. Rodriguez-Valera, R. Amann and R. Rossello-Mora (2002). "Salinibacter ruber gen. nov., sp. nov., a novel, extremely halophilic member of the Bacteria from saltern crystallizer ponds." Int J Syst Evol Microbiol **52**(Pt 2): 485-491.

Anton, J., R. Rossello-Mora, F. Rodriguez-Valera and R. Amann (2000). "Extremely halophilic bacteria in crystallizer ponds from solar salterns." Appl Environ Microbiol **66**(7): 3052-3057.

Avrani, S., O. Wurtzel, I. Sharon, R. Sorek and D. Lindell (2011). "Genomic island variability facilitates Prochlorococcus-virus coexistence." Nature **474**(7353): 604-608.

Baati, H., S. Guermazi, R. Amdouni, N. Gharsallah, A. Sghir and E. Ammar (2008). "Prokaryotic diversity of a Tunisian multipond solar saltern." Extremophiles **12**(4): 505-518.

Bendtsen, J. D., H. Nielsen, G. von Heijne and S. Brunak (2004). "Improved prediction of signal peptides: SignalP 3.0." J Mol Biol **340**(4): 783-795.

Benlloch, S., S. G. Acinas, J. Anton, A. Lopez-Lopez, S. P. Luz and F. Rodriguez-Valera (2001). "Archaeal Biodiversity in Crystallizer Ponds from a Solar Saltern: Culture versus PCR." Microb Ecol **41**(1): 12-19.

Benlloch, S., A. Lopez-Lopez, E. O. Casamayor, L. Ovreas, V. Goddard, F. L. Daae*, et al.* (2002). "Prokaryotic genetic diversity throughout the salinity gradient of a coastal solar saltern." Environ Microbiol **4**(6): 349-360.

Benlloch, S., A. Martinez-Murcia and F. Rodriguez-Valera (1995). "Sequencing of bacterial and archaeal 16S rRNA genes directly amplified from a hypersaline environment." Syst Appl Microbiol **18**: 574-581.

Bentley, S. (2009). "Sequencing the species pan-genome." Nat Rev Microbiol **7**(4): 258-259.

Bergthorsson, U. and H. Ochman (1998). "Distribution of chromosome length variation in natural isolates of Escherichia coli." Mol Biol Evol **15**(1): 6-16.

Blaurock, A. E., W. Stoeckenius, D. Oesterhelt and G. L. Scherfhof (1976). "Structure of the cell envelope of Halobacterium halobium." J Cell Biol **71**(1): 1-22.

Bolhuis, H., E. M. Poele and F. Rodriguez-Valera (2004). "Isolation and cultivation of Walsby's square archaeon." Environ Microbiol **6**(12): 1287-1291.

Bolhuis, H. H., P. P. Palm, A. A. Wende, M. M. Falb, M. M. Rampp, F. F. Rodriguez-Valera*, et al.* (2006). "The genome of the square archaeon Haloquadratum walsbyi: life at the limits of water activity." BMC Genomics **7**(1): 169.

Brito-Echeverria, J., A. Lopez-Lopez, P. Yarza, J. Anton and R. Rossello-Mora (2009). "Occurrence of Halococcus spp. in the nostrils salt glands of the seabird Calonectris diomedea." Extremophiles **13**(3): 557-565.

Burns, D. G., H. M. Camakaris, P. H. Janssen and M. L. Dyall-Smith (2004). "Combined use of cultivation-dependent and cultivation-independent methods indicates that members of most haloarchaeal groups in an Australian crystallizer pond are cultivable." Appl Environ Microbiol **70**(9): 5258-5265.

Burns, D. G., P. H. Janssen, T. Itoh, M. Kamekura, Z. Li, G. Jensen*, et al.* (2007). "Haloquadratum walsbyi gen. nov., sp. nov., the square haloarchaeon of Walsby, isolated from saltern crystallizers in Australia and Spain." Int J Syst Evol Microbiol **57**(Pt 2): 387-392.

Casamayor, E. O., R. Massana, S. Benlloch, L. Ovreas, B. Diez, V. J. Goddard, *et al.* (2002). "Changes in archaeal, bacterial and eukaryal assemblages along a salinity gradient by comparison of genetic fingerprinting methods in a multipond solar saltern." Environ Microbiol **4**(6): 338-348.

Coleman, M. L., M. B. Sullivan, A. C. Martiny, C. Steglich, K. Barry, E. F. Delong, *et al.* (2006). "Genomic islands and the ecology and evolution of Prochlorococcus." Science **311**(5768): 1768-1770.

Cuadros-Orellana, S., A. B. Martin-Cuadrado, B. Legault, G. D'Auria, O. Zhaxybayeva, R. T. Papke, *et al.* (2007). "Genomic plasticity in prokaryotes: the case of the square haloarchaeon." Isme J **1**(3): 235-245.

Che, D., M. S. Hasan, H. Wang, J. Fazekas, J. Huang and Q. Liu (2011). "EGID: an ensemble algorithm for improved genomic island detection in genomic sequences." Bioinformation **7**(6): 311-314.

Delcher, A. L., D. Harmon, S. Kasif, O. White and S. L. Salzberg (1999). "Improved microbial gene identification with GLIMMER." Nucleic Acids Res **27**(23): 4636-4641.

Dempsey, M. P., J. Nietfeldt, J. Ravel, S. Hinrichs, R. Crawford and A. K. Benson (2006). "Paired-end sequence mapping detects extensive genomic rearrangement and translocation during divergence of Francisella tularensis subsp. tularensis and Francisella tularensis subsp. holarctica populations." J Bacteriol **188**(16): 5904-5914.

Dobrindt, U., B. Hochhut, U. Hentschel and J. Hacker (2004). "Genomic islands in pathogenic and environmental microorganisms." Nat Rev Microbiol **2**(5): 414-424.

Dutta, C. and A. Pan (2002). "Horizontal gene transfer and bacterial diversity." J Biosci **27**(1 Suppl 1): 27-33.

Dyall-Smith, M. L., F. Pfeiffer, K. Klee, P. Palm, K. Gross, S. C. Schuster, *et al.* (2011). "Haloquadratum walsbyi : Limited Diversity in a Global Pond." PLoS One **6**(6): e20968.

Eddy, S. R. (2008). "A probabilistic model of local sequence alignment that simplifies statistical significance estimation." PLoS Comput Biol **4**(5): e1000069.

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

Elevi Bardavid, R., P. Khristo and A. Oren (2008). "Interrelationships between Dunaliella and halophilic prokaryotes in saltern crystallizer ponds." Extremophiles **12**(1): 5-14.

Estrada, M., P. Henriksen, J. M. Gasol, E. O. Casamayor and C. Pedros-Alio (2004). "Diversity of planktonic photoautotrophic microorganisms along a salinity gradient as depicted by microscopy, flow cytometry, pigment analysis and DNA-based methods." FEMS Microbiol Ecol **49**(2): 281-293.

Falb, M., K. Muller, L. Konigsmaier, T. Oberwinkler, P. Horn, S. von Gronau, *et al.* (2008). "Metabolism of halophilic archaea." Extremophiles **12**(2): 177-196.

Fendrihan, S., A. Legat, M. Pfaffenhuemer, C. Gruber, G. Weidler, F. Gerbl, *et al.* (2006). "Extremely halophilic archaea and the issue of long-term microbial survival." **5**(2-3): 203-218.

Fendrihan, S., M. Musso and H. Stan-Lotter (2009). "Raman spectroscopy as a potentialmethod for the detection of extremely halophilic archaea embedded in halite in terrestrial and possibly extraterrestrial samples." J Raman Spectrosc **40**(12): 1996-2003.

Finn, R. D., J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, *et al.* (2010). "The Pfam protein families database." Nucleic Acids Res **38**(Database issue): D211-222.

Garcia-Vallve, S., A. Romeu and J. Palau (2000). "Horizontal gene transfer in bacterial and archaeal complete genomes." Genome Res **10**(11): 1719-1725.

Gasol, E., M. Jimenez-Vidal, J. Chillaron, A. Zorzano and M. Palacin (2004). "Membrane topology of system xc- light subunit reveals a re-entrant loop with substrate-restricted accessibility." J Biol Chem **279**(30): 31228-31236.

Ghai, R., L. Pašić, A. B. Fernández, A.-B. Martin-Cuadrado, C. M. Mizuno, K. D. McMahon, *et al.* (2011). "New Abundant Microbial Groups in Aquatic Hypersaline Environments." Sci. Rep. **1**.

Guixa-Boixereu, N. (1996). "Viral lysis and bacterivory as prokaryotic loss factors along a salinity gradient." Aquatic Microbial Ecology **11**: 213-227.

Hallsworth, J. E., M. M. Yakimov, P. N. Golyshin, J. L. Gillion, G. D'Auria, F. de Lima Alves, *et al.* (2007). "Limits of life in MgCl2-containing environments: chaotropicity defines the window." Environ Microbiol **9**(3): 801-813.

Hasan, M. S., Q. Liu, H. Wang, J. Fazekas, B. Chen and D. Che (2012). "GIST: Genomic island suite of tools for predicting genomic islands in genomic sequences." Bioinformation **8**(4): 203-205.

Hochhut, B., C. Wilde, G. Balling, B. Middendorf, U. Dobrindt, E. Brzuszkiewicz*, et al.* (2006). "Role of pathogenicity island-associated integrases in the genome plasticity of uropathogenic Escherichia coli strain 536." Mol Microbiol **61**(3): 584-595.

Ivars-Martinez, E., A. B. Martin-Cuadrado, G. D'Auria, A. Mira, S. Ferriera, J. Johnson*, et al.* (2008). "Comparative genomics of two ecotypes of the marine planktonic copiotroph Alteromonas macleodii suggests alternative lifestyles associated with different kinds of particulate organic matter." Isme J **2**(12): 1194-1212.

Jarrell, K. F., G. M. Jones, L. Kandiba, D. B. Nair and J. Eichler (2010a). "S-layer glycoproteins and flagellins: reporters of archaeal posttranslational modifications." Archaea **2010**.

Jarrell, K. F., G. M. Jones and D. B. Nair (2010b). "Biosynthesis and role of N-linked glycosylation in cell surface structures of archaea with a focus on flagella and s layers." Int J Microbiol **2010**: 470138.

Jordan, I. and E. Koonin (2004). "Horizontal gene transfer and prokaryotic genome evolution." Miller RV, Day MJ (eds). Microbial Evolution. Gene Establishment, Survival, and Exchange. ASM Press: Washington, DC: 319–338.

Kessel, M. and Y. Cohen (1982). "Ultrastructure of square bacteria from a brine pool in Southern Sinai." J Bacteriol **150**(2): 851-860.

Kettler, G. C., A. C. Martiny, K. Huang, J. Zucker, M. L. Coleman, S. Rodrigue*, et al.* (2007). "Patterns and Implications of Gene Gain and Loss in the Evolution of Prochlorococcus." PLoS Genet **3**(12): e231.

Konstantinidis, K. T. and J. M. Tiedje (2005). "Genomic insights that advance the species definition for prokaryotes." Proc Natl Acad Sci U S A **102**(7): 2567-2572.

Krogh, A., B. Larsson, G. von Heijne and E. L. Sonnhammer (2001). "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes." J Mol Biol **305**(3): 567-580.

Lawrence, J. G. and H. Hendrickson (2005). "Genome evolution in bacteria: order beneath chaos." Curr Opin Microbiol **8**(5): 572-578.

Lechner, J. and M. Sumper (1987). "The primary structure of a procaryotic glycoprotein. Cloning and sequencing of the cell surface glycoprotein gene of halobacteria." J Biol Chem **262**(20): 9724-9729.

Legault, B. A., A. Lopez-Lopez, J. C. Alba-Casado, W. F. Doolittle, H. Bolhuis, F. Rodriguez-Valera*, et al.* (2006). "Environmental genomics of "Haloquadratum walsbyi" in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species." BMC Genomics **7**(1): 171.

Lerat, E., V. Daubin, H. Ochman and N. A. Moran (2005). "Evolutionary origins of genomic repertoires in bacteria." PLoS Biol **3**(5): e130.

Letunic, I., T. Doerks and P. Bork (2009). "SMART 6: recent updates and new developments." Nucleic Acids Res **37**(Database issue): D229-232.

Litchfield, C. and P. Gillevetr (2002). "Microbial diversity and complexity in hypersaline environments: a preliminar assessment." J Ind Microbiol & Biotech **28**: 48-55.

Liu, M., R. Deora, S. R. Doulatov, M. Gingery, F. A. Eiserling, A. Preston*, et al.* (2002). "Reverse transcriptase-mediated tropism switching in Bordetella bacteriophage." Science **295**(5562): 2091-2094.

Lowe, T. M. and S. R. Eddy (1997). "tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence." Nucleic Acids Res **25**(5): 955-964.

Lukjancenko, O., T. M. Wassenaar and D. W. Ussery (2010). "Comparison of 61 sequenced Escherichia coli genomes." Microb Ecol **60**(4): 708-720.

Mardis, E. R. (2008a). "The impact of next-generation sequencing technology on genetics." Trends Genet **24**(3): 133-141.

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

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben*, et al.* (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-380.

Martin, D. P., P. Lemey, M. Lott, V. Moulton, D. Posada and P. Lefeuvre (2010). "RDP3: a flexible and fast computer program for analyzing recombination." Bioinformatics **26**(19): 2462-2463.

Maturrano, L., M. Valens-Vadell, R. Rossello-Mora and J. Anton (2006). "Salicola marasensis gen. nov., sp. nov., an extremely halophilic bacterium isolated from the Maras solar salterns in Peru." Int J Syst Evol Microbiol **56**(Pt 7): 1685-1691.

Mengele, R. and M. Sumper (1992). "Drastic differences in glycosylation of related S-layer glycoproteins from moderate and extreme halophiles." J Biol Chem **267**(12): 8182-8185.

Michina, Y., D. Carriere, C. Mariet, M. Moskura, P. Berthault, L. Belloni*, et al.* (2009). "Ripening of catanionic aggregates upon dialysis." Langmuir **25**(2): 698-706.

Milkman, R. (1997). "Recombination and population structure in Escherichia coli." Genetics **146**(3): 745-750.

Miller, J. L., J. Le Coq, A. Hodes, R. Barbalat, J. F. Miller and P. Ghosh (2008). "Selective ligand recognition by a diversity-generating retroelement variable protein." PLoS Biol **6**(6): e131.

Mongodin, E. F., K. E. Nelson, S. Daugherty, R. T. Deboy, J. Wister, H. Khouri*, et al.* (2005). "The genome of Salinibacter ruber: convergence and gene exchange among hyperhalophilic bacteria and archaea." Proc Natl Acad Sci U S A **102**(50): 18147-18152.

Munroe, D. J. and T. J. Harris (2010). "Third-generation sequencing fireworks at Marco Island." Nat Biotechnol **28**(5): 426-428.

Mutlu, M. B., M. Martinez-Garcia, F. Santos, A. Pena, K. Guven and J. Anton (2008). "Prokaryotic diversity in Tuz Lake, a hypersaline environment in Inland Turkey." FEMS Microbiol Ecol **65**(3): 474-483.

Muzzi, A. and C. Donati (2011). "Population genetics and evolution of the pan-genome of Streptococcus pneumoniae." Int J Med Microbiol **301**(8): 619-622.

Nakamura, S., H. Mizutani, H. Wakai, R. A. Kawasaki and K. Horikoshi (1995). "Purification and Partial Characterization of Cell-Surface Glycoprotein from Extremely Halophilic archaeon Haloarcula japonica strain TR-1." Biotechnology Letters **17** (7): 705-706.

Narasingarao, P., S. Podell, J. A. Ugalde, C. Brochier-Armanet, J. B. Emerson, J. J. Brocks*, et al.* (2011). "De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities." Isme J.

Ochman, H. (2001). "Lateral and oblique gene transfer." Curr Opin Genet Dev **11**(6): 616-619.

Ochman, H., J. G. Lawrence and E. A. Groisman (2000). "Lateral gene transfer and the nature of bacterial innovation." Nature **405**(6784): 299-304.

Oh, D., K. Porter, B. Russ, D. Burns and M. Dyall-Smith (2010). "Diversity of Haloquadratum and other haloarchaea in three, geographically distant, Australian saltern crystallizer ponds." Extremophiles **14**(2): 161-169.

Oren, A. (1994). "Enzyme diversity in halophilic archaea." Microbiologia **10**(3): 217-228.

Oren, A. (2002a). "Diversity of halophilic microorganisms: environments, phylogeny, physiology, and applications." J Ind Microbiol Biotechnol **28**(1): 56-63.

Oren, A. (2002b). "Molecular ecology of extremely halophilic Archaea and Bacteria." FEMS Microbiol Ecol **39**(1): 1-7.

Oren, A. (2005). "A hundred years of Dunaliella research: 1905-2005." Saline Systems **1**: 2.

Oren, A. (2008). "Microbial life at high salt concentrations: phylogenetic and metabolic diversity." Saline Systems **4**: 2.

Oren, A. and F. Rodriguez-Valera (2001). "The contribution of halophilic Bacteria to the red coloration of saltern crystallizer ponds(1)." FEMS Microbiol Ecol **36**(2-3): 123-130.

Papke, R. T., J. E. Koenig, F. Rodriguez-Valera and W. F. Doolittle (2004). "Frequent recombination in a saltern population of Halorubrum." Science **306**(5703): 1928-1929.

Pasic, L., S. G. Bartual, N. P. Ulrih, M. Grabnar and B. H. Velikonja (2005). "Diversity of halophilic archaea in the crystallizers of an Adriatic solar saltern." FEMS Microbiol Ecol **54**(3): 491-498.

Pedros-Alio, C., J. I. Calderon-Paz, M. H. MacLean, G. Medina, C. Marrase, J. M. Gasol*, et al.* (2000). "The microbial food web along salinity gradients." FEMS Microbiol Ecol **32**(2): 143-155.

Pena, A., H. Teeling, J. Huerta-Cepas, F. Santos, P. Yarza, J. Brito-Echeverria*, et al.* (2010). "Fine-scale evolution: genomic, phenotypic and ecological differentiation in two coexisting Salinibacter ruber strains." Isme J **4**(7): 882-895.

Petrosino, J. F., Q. Xiang, S. E. Karpathy, H. Jiang, S. Yerrapragada, Y. Liu*, et al.* (2006). "Chromosome rearrangement and diversification of Francisella tularensis revealed by the type B (OSU18) genome sequence." J Bacteriol **188**(19): 6977-6985.

Putaporntip, C., S. Jongwutiwes, M. U. Ferreira, H. Kanbara, R. Udomsangpetch and L. Cui (2009). "Limited global diversity of the Plasmodium vivax merozoite surface protein 4 gene." Infect Genet Evol **9**(5): 821-826.

Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler*, et al.* (2005). "InterProScan: protein domains identifier." Nucleic Acids Res **33**(Web Server issue): W116-120.

Ragan, M. A. (2001). "Detection of lateral gene transfer among microbial genomes." Curr Opin Genet Dev **11**(6): 620-626.

Reva, O. and B. Tummler (2008). "Think big--giant genes in bacteria." Environ Microbiol **10**(3): 768-777.

Rice, P., I. Longden and A. Bleasby (2000). "EMBOSS: the European Molecular Biology Open Software Suite." Trends Genet **16**(6): 276-277.

Rodriguez-Brito, B., L. Li, L. Wegley, M. Furlan, F. Angly, M. Breitbart*, et al.* (2010). "Viral and microbial community dynamics in four aquatic environments." Isme J **4**(6): 739-751.

Rodríguez-Valera, F. (1988). "Characteristics and microbial ecology of hypersaline environments.

." Editado por F. Rodríguez-Valera. Boca-Raton, FL: CRC Press. **1**: 3-30.

Rodríguez-Valera, F., S. Acinas and J. Antón (1999). "Contribution of molecular techniques to the study of microbial diversity in hipersaline environments. ." En Oren A (editor)

Microbioloyg and Biogeochemistry of hypersaline environments. CRC Press, Boca-Raton, FL: 27-38.

Rodriguez-Valera, F., A. B. Martin-Cuadrado, B. Rodriguez-Brito, L. Pasic, T. F. Thingstad, F. Rohwer*, et al.* (2009). "Explaining microbial population genomics through phage predation." Nat Rev Microbiol **7**(11): 828-836.

Rodriguez-Valera, F., J. J. Nieto and F. Ruiz-Berraquero (1983). "Light as an Energy Source in Continuous Cultures of Bacteriorhodopsin-Containing Halobacteria." Appl Environ Microbiol **45**(3): 868-871.

Rossello-Mora, R., M. Lucio, A. Pena, J. Brito-Echeverria, A. Lopez-Lopez, M. Valens-Vadell*, et al.* (2008). "Metabolic evidence for biogeographic isolation of the extremophilic bacterium Salinibacter ruber." ISME J **2**(3): 242-253.

Santos, F., P. Yarza, V. Parro, C. Briones and J. Anton (2010). "The metavirome of a hypersaline environment." Environ Microbiol.

Schaffer, C., M. Graninger and P. Messner (2001). "Prokaryotic glycosylation." Proteomics **1**(2): 248-261.

Schaffer, C. and P. Messner (2001). "Glycobiology of surface layer proteins." Biochimie **83**(7): 591-599.

Sigrist, C. J., L. Cerutti, E. de Castro, P. S. Langendijk-Genevaux, V. Bulliard, A. Bairoch*, et al.* (2010). "PROSITE, a protein domain database for functional characterization and annotation." Nucleic Acids Res **38**(Database issue): D161-166.

Soares, S. C., V. A. Abreu, R. T. Ramos, L. Cerdeira, A. Silva, J. Baumbach*, et al.* (2012). "PIPS: pathogenicity island prediction software." PLoS One **7**(2): e30848.

Stoeckenius, W. (1981). "Walsby's square bacterium: fine structure of an orthogonal procaryote." J Bacteriol **148**(1): 352-360.

Sumper, M., E. Berg, R. Mengele and I. Strobel (1990). "Primary structure and glycosylation of the S-layer protein of Haloferax volcanii." J Bacteriol **172**(12): 7111-7118.

Syvanen, A. C. (1998). "Solid-phase minisequencing as a tool to detect DNA polymorphism." Methods Mol Biol **98**: 291-298.

Tamames, J. and A. Moya (2008). "Estimating the extent of horizontal gene transfer in metagenomic sequences." BMC Genomics **9**: 136.

Tettelin, H., V. Masignani, M. J. Cieslewicz, C. Donati, D. Medini, N. L. Ward*, et al.* (2005). "Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome"." Proc Natl Acad Sci U S A **102**(39): 13950-13955.

Thingstad, T. (2000). "Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic ecosystems." Limnol Oceanogr **45**: 1320-1328.

Ting, C. S., G. Rocap, J. King and S. W. Chisholm (2002). "Cyanobacterial photosynthesis in the oceans: the origins and significance of divergent light-harvesting strategies." Trends Microbiol **10**(3): 134-142.

Trachtenberg, S., B. Pinnick and M. Kessel (2000). "The cell surface glycoprotein layer of the extreme halophile Halobacterium salinarum and its relation to Haloferax volcanii: cryo-electron tomography of freeze-substituted cells and projection studies of negatively stained envelopes." J Struct Biol **130**(1): 10-26.

Tyson, G. W., J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson*, et al.* (2004). "Community structure and metabolism through reconstruction of microbial genomes from the environment." Nature **428**(6978): 37-43.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen*, et al.* (2004). "Environmental genome shotgun sequencing of the Sargasso Sea." Science **304**(5667): 66-74.

Ventosa, A. (2006). "Inusual micro-organisms from inusual habitats: hypersaline environments. SGM symposium 66: Prokariotic diversity – mechanisms and significance." N. A. Logan, H.M. Lappin-Scott y P.C.F (editores). Ovston. Cambridge University Press.

Walsby, A. E. (2005). "Archaea with square cells." Trends Microbiol **13**(5): 193-195.

Wang, L., K. Andrianopoulos, D. Liu, M. Y. Popoff and P. R. Reeves (2002). "Extensive variation in the O-antigen gene cluster within one Salmonella enterica serogroup reveals an unexpected complex history." J Bacteriol **184**(6): 1669-1677.

Wilhelm, L. J., H. J. Tripp, S. A. Givan, D. P. Smith and S. J. Giovannoni (2007). "Natural variation in SAR11 marine bacterioplankton genomes inferred from metagenomic data." Biol Direct **2**: 27.

Willenbrock, H., A. Petersen, C. Sekse, K. Kiil, Y. Wasteson and D. W. Ussery (2006). "Design of a seven-genome Escherichia coli microarray for comparative genomic profiling." J Bacteriol **188**(22): 7713-7721.

Xiang, S. H., M. Hobbs and P. R. Reeves (1994). "Molecular analysis of the rfb gene cluster of a group D2 Salmonella enterica strain: evidence for its origin from an insertion sequence-mediated recombination event between group E and D1 strains." J Bacteriol **176**(14): 4357-4365.

7.- ANEXOS

ANEXO I

Tabla A. Anotación de los fósmidos ambientales eHw de este trabajo. (Ver archivo excell anexo).

ANEXO II

Sofware desarrollado en Perl usado en este trabajo

Extraer hits del Análisis HMMERR

```
#!/usr/bin/perl

# parse multi-HMM file

use Bio::SearchIO;


print "Nombre del archivo.....  ";
chomp(my $hmm_file=<STDIN>);
print "\n";
print "Nombre del archivo de salida .....  ";
chomp(my $salida_file=<STDIN>);
print "\n";
open(INFILE, ">$salida_file");

my $in = new Bio::SearchIO(-format => 'hmmer_pull',-file => $hmm_file);

print INFILE "tQuery_name\tHitName\tHitDescription\tHitLength\te-value\tScore\n\n";

while ( $res = $in->next_result )
{
   # get a Bio::Search::Result::HMMERResult object
   print INFILE "\n",$res->query_name, "\t";
   while ( $hit = $res->next_hit )
   {
     print INFILE $hit->name, "\t",$hit->description, "\t";
     while ( $hsp = $hit->next_hsp )
     {
       print INFILE $hsp->length,"\t",$hsp->evalue,"\t",$hsp->score,  "\t";
     }
   }
}
```

Extraer secuencia de aminoácidos de los genes de un genoma (o fragmento genómico) a partir de un archivo Genbank

```
#!/usr/bin/perl

use Bio::SeqIO;
use Bio::DB::GenBank;


print "Nombre del archivo.....  ";
chomp(my $gb_file=<STDIN>);
print "\n";

print "Nombre del archivo de salida .....  ";
chomp(my $salida_file=<STDIN>);
print "\n";


my %hash;

open(INFILE, ">$salida_file");

my $seqio_object = Bio::SeqIO->new(-file => $gb_file);
while(my $seq_object = $seqio_object->next_seq)
{
 for my $feat_object($seq_object->get_SeqFeatures)
 {
                if ($feat_object->has_tag('translation'))
                {
    for my $feat($feat_object->get_tag_values ('translation'))
    {
      my $start = $feat_object->location->start;
      my $end = $feat_object->location->end;

          print $feat,"\n";
          print INFILE ">",$seq_object->display_id,"\t",$start,"@",$end,"\n", $feat,"\n";

     }

   }
 }
}

close INFILE;
```

Extraer secuencia de nucleótiods de los genes de un genoma (o fragmento genómico) a partir de un archivo Genbank

```
#!/usr/bin/perl

#
#
use Bio::SeqIO;

print "Nombre del archivo.....  ";
chomp(my $gb_file=<STDIN>);
print "\n";
print "Nombre del archivo de salida .....  ";
chomp(my $salida_file=<STDIN>);
print "\n";
my %hash;

open(INFILE, ">$salida_file");

my $seqio_object = Bio::SeqIO->new(-file => $gb_file);
while(my $seq_object = $seqio_object->next_seq)
{
 for my $feat_object($seq_object->get_SeqFeatures)
 {
                if ($feat_object->has_tag('product'))
                {
    for my $feat($feat_object->get_tag_values ('product'))
     {
     print $feat,"\t",$feat_object->seq->seq,"\n";
     print INFILE ">",$feat,"\t",$feat_object->seq->seq,"\n";
     }
   }
 }
}
```

Preparar los datos para las gráficas de reclutamientos

```
#!/usr/bin/perl -w

#
#
#
# perl program.pl  file_with_starts_ends_and_% (in 3 columns, tab delimited)
# output file is made like inputfilename.out
#



my $pattern = shift; #print "Lista Query Start: ";
chomp $pattern;

my @input_files = < $pattern >;


foreach my $file (@input_files) {

        open (INP,$file) || die;



        open(OUTFILE,">$file.out") || die "Cannot open output file $file.out";

        while (my $line = <INP>) {
                chomp $line;
                $line =~ s/^\s+|\s+$//g;
                next if $line eq "";
                my ($a,$b,$c) = split(/\t/,$line);
                print OUTFILE $a,"\t",$c,"\n",$b,"\t",$c,"\n\n";

        }

        close INP;
        close OUTFILE;

}
```

Contar secuencias de un archivo multi-fasta

```
#!/usr/bin/perl


use Bio::SeqIO;
use strict;
print "\n";
print ("nombre del archivo:\n") ;
print "\n";
chomp (my $my_file  = <STDIN>) ;
open (INP, "$my_file" ) ;

my $seq;
my $seq_obj = Bio::SeqIO->new(-file=>"$my_file");
my $count=0;


while ( my $inseq = $seq_obj->next_seq) {

  my $seq = $inseq->seq();

if ($seq ne "")    {

  $count+=1;

}

}
print "\n" ;
print ("su archivo contiene $count secuencias" )     ;
print "\n" ;
```

Filtrar los datos de BLAST

```
#!/usr/bin/perl -w
use Bio::SearchIO;
# ARCHIVO
print "\n\n";
print "ReportBlast a analizar .....  ";
chomp(my $report_blast=<STDIN>);
print "\n";
# PARAMETROS
print "Similitud mínima (%)... ";
chomp($sim=<STDIN>);
print "Longitud minima de la secuencia..... ";
chomp($lonm=<STDIN>);
print "E-value..... ";
chomp($E_value=<STDIN>);
# ARCHIVO SALIDA
print "\n";
print "Nombre archivo salida..... ";
chomp($outf=<STDIN>);

open (OUTFILE, ">$outf");

$report_obj = new Bio::SearchIO(-format => 'blast',
                        -file   => $report_blast);

print OUTFILE "Analisis de $report_blast teniendo en cuenta:\n\n";
print OUTFILE "\t\tSimilitud minima (%)... $sim\n";
print OUTFILE "\t\tLongitud minima de la secuencia.....$lonm\n";
print OUTFILE "\t\tE_value mayor.....$E_value\n\n";

print OUTFILE "tQuery_name\tQuery_description\tQuery_lenght\tHitName\tHitDescription\tSeq\tHitLength\tHspLength\tPercentID\te-
value\tScore\trank\tQueryStart\tQueryEnd\tHitStart\tHitEnd\tLocus\n\n";

while( $result = $report_obj->next_result ) {
  if( $hit = $result->next_hit ) {
    if( $hsp = $hit->next_hsp ) {
      if ( $hsp->percent_identity > $sim and $hsp->length > $lonm and $hsp->evalue < $E_value ) {

        my $hsp_percent_identity=$hsp->percent_identity;$hsp_percent_identity=~ s/\./\,/g;
        my $hsp_evalue = $hsp->evalue;$hsp_evalue=~ s/\./\,/g;
        print $result->query_name, $result->query_description, $result->query_length,"Hit\t", $hit->name, "\tHitLength ", $hit->length,"\tHspLength ",
        $hit->description,$hsp->length,"\tPercent_id ", $hsp_percent_identity, "\tE-Value ", $hsp_evalue,"\tScore ", $hsp->score,
        $hsp->start('query'),"\t",$hsp->end('query'),"\t",$hsp->start('hit'),"\t",$hsp->end('hit'),"\t", $hit->locus, "\n\n";

        print OUTFILE $result->query_name,"\t", $result->query_description,"\t",$result->query_length,"\t", $hit->name,"\t",$hit->description,"\t",$hsp->hit_string,
        "\t",$hit->length, "\t", $hsp->length, "\t",$hsp_percent_identity, "\t",$hsp_evalue,"\t",$hsp->score, "\t",$hsp->rank,
        "\t",$hsp->start('query'),"\t", $hsp->end('query'),"\t",$hsp->start('hit'),"\t", $hsp->end('hit'),"\t", $hit->locus,"\n";
      }
    }
  }
 }
close OUTFILE;
```

ANEXO III

Tabla B

Table 1 2nd and 3rd Generation DNA sequencing platforms listed in the order of commercial availability

Platform	Current company	Former company	Sequencing method	Amplification method	Claim to fame	Primary applications
454	Roche	454	Synthesis (pyrosequencing)	emPCR	First Next-Gen Sequencer, Long reads	**1***, **2**, **3***, 4, 7, 8*
Illumina	Illumina	Solexa	Synthesis	BridgePCR	First short-read sequencer; current leader in advantages†	1*, **2**, **3***, 4, 5, 6, 7, 8
SOLiD	Life Technologies	Applied Biosystems	Ligation	emPCR	Second short-read sequencer; low error rates	3*, 5, 6, 8
HeliScope	Helicos	N/A	Synthesis	None	First single-molecule sequencer	5, 8
Ion Torrent	Life Technologies	Ion Torrent	Synthesis (H^+ detection)	emPCR	First Post-light sequencer; first system <$100 000	**1**, **2**, **3**, 4, 8
PacBio	Pacific Biosciences	N/A	Synthesis	None	First real-time single-molecule sequencing	1, 2, 3, 7, 8
Starlight‡	Life Technologies	N/A	Synthesis	None	Single-molecule sequencing with quantum dots	1, 2, 7, 8

Bold indicates applications that are most often used, economical or growing.
1 = *de novo* BACs, plastids, microbial genomes.
2 = transcriptome characterization.
3 = targeted re-sequencing.
4 = *de novo* plant and animal genomes.
5 = re-sequencing and transcript counting.
6 = mutation detection.
7 = metagenomics.
8 = other (ChIP-Seq, μRNA-Seq, Methyl-Seq, etc.; see Brautigam & Gowik 2010, Shendure & Ji 2008).
*Pooling multiple samples with sequence tags (i.e. MIDs or indexes) is required for efficient use of this application
†Illumina currently leads in number and percentage of error-free reads, Illumina HiSeqs with v3 chemistry lead in reads per run, GB/run, and cost/GB.
‡A commercial launch date for the Starlight system is not yet known, but it is included here because it is in advanced development, and some information about its performance characteristics is known.

Tabla C

Table 2 Comparison of sequencing instruments, sorted by cost/Mb, with expected performance by mid 2011

Instrument	Run time[a]	Millions of reads/run	Bases/read[b]	Yield Mb/run	Reagent cost/run[c]	Reagent cost/Mb	Minimum unit cost (% run)[d]
3730xl (capillary)	2 h	0.000096	650	0.06	$96	$1500	$6 (1%)
Ion Torrent – '314'chip	2 h	0.10	100	>10	$500	<$50	~$750 (100%)
454 GS Jr. Titanium	10 h	0.10	400	50	$1100	$22	$1500 (100%)
Starlight*	†	~0.01	>1000	†	†	†	†
PacBio RS	0.5–2 h	0.01	860–1100	5–10	$110–900	$11–180	†
454 FLX Titanium	10 h	1	400	500	$6200	$12.4	$2000 (10%)
454 FLX+[e]	18–20 h	1	700	900	$6200	$7	$2000 (10%)
Ion Torrent – '316'chip*	2 h	1	>100	>100	$750	<$7.5	~$1000 (100%)
Helicos[f]	N/A	800	35	28 000	N/A	NA	$1100 (2%)
Ion Torrent – '318'chip*	2 h	4–8	>100	>1000	~$925	~$0.93	~$1200 (100%)
Illumina MiSeq*	26 h	3.4	150 + 150	1020	$750	$0.74	~$1000 (100%)
Illumina iScanSQ	8 days	250	100 + 100	50 000	$10 220	$0.20	$3000 (14%)
Illumina GAIIx	14 days	320	150 + 150	96 000	$11 524	$0.12	$3200 (14%)
SOLiD – 4	12 days	>840[g]	50 + 35	71 400	$8128	<$0.11	$2500 (12%)
Illumina HiSeq 1000	8 days	500	100 + 100	100 000	$10 220	$0.10	$3000 (12%)
Illumina HiSeq 2000	8 days	1000	100 + 100	200 000	$20 120[h]	$0.10	$3000 (6%)
SOLiD – 5500 (PI)*	8 days	>700[g]	75 + 35	77 000	$6101	<$0.08	$2000 (12%)
SOLiD – 5500xl (4hq)*	8 days	>1410[g]	75 + 35	155 100	$10 503[h]	<$0.07	$2000 (12%)
Illumina HiSeq 2000 – v3[i]*	10 days	≤3000	100 + 100	≤600 000	$23 470[h]	≥$0.04	~$3500 (6%)

[a]Instrument time for maximum read length.
[b]Average length for high-quality reads >200 bases (mode is higher); typical maximum for reads ≤150 bases (most reads reach this length).
[c]Includes all stages of sample preparation for a single sample (i.e. library preparation through sequencing; capillary = sequencing only).
[d]Typical full cost (i.e. including labour, service contract, etc.) of the smallest generally available unit of purchase at an academic core laboratory provider for the longest available read (and percentage of reads relative to a full run, rounded to the nearest whole percentage).
[e]Upgrade of the FLX instrument, due out summer 2011.
[f]Instruments and reagents are no longer sold; services are available for any organism.
[g]Mappable reads [number of raw high-quality reads (as reported for all other platforms) is higher].
[h]More reads are obtained than is needed from any single sample within most experiments, but the value illustrates the costs.
[i]AnnouncedTruSeq v3 reagents & software, reads and yield are half for HiSeq1000.
*Information based on company sources alone (independent data not yet available).
†Detail not yet available.
'~' Indicates a likely value based on unpublished information available in March 2011 (i.e. author speculation).

Figura D. RBS marcado de cada en cada uno de los fósmido secuenciados.