Estudios de composición nucleotídica y uso de codones en la familia *Pelagibacteriaceae* del ubicuo clado marino SAR11.

Lic. Lucía Graña Miraglia

luciagrami@gmail.com

Orientador: Dr. Héctor Romero

eletor@gmail.com

Laboratorio de Organización y Evolución del Genoma

Departamento de Ecología y Evolución, Instituto de Biología

Facultad de Ciencias, UR

PEDECIBA

Maestría en Ciencias Biológicas

Subárea Ecología y Evolución

Marzo 2015

Tribunal

Presidenta: Cecilia Alonso

Vocales: Andrés Iriarte, Ivanna Tommasco

Índice de contenido

Resumen ... 3
INTRODUCCIÓN GENERAL ... 4
Objetivo general ... 9
Objetivos específicos y preguntas asociadas ... 9
Capítulo I: Patrones Sustitucionales en Pelagibacteriaceae ... 12
Introducción ... 12
Evolución de la composición de bases de genomas procariotas ... 12
Metodología ... 16
DATOS ... 16
DISEÑO EXPERIMENTAL ... 17
Búsqueda de genes ortólogos ... 17
Reconstrucción filogenética ... 18
Búsqueda de incongruencias filogenéticas en genes ortólogos ... 19
Detección de genes recombinantes ... 19
Congruencia en las reconstrucciones filogenéticas ... 20
Estimaciones dN y dS ... 20
Estructura Poblacional ... 21
Patrón Sustitucional ... 22
Aproximación al Estudio de Patrones Mutacionales ... 22
Reconstrucción de caracteres ancestrales ... 24
Resultados ... 28
Filogenia y genes recombinantes ... 28
Estimaciones de dS, dN y w ... 31
Patrones sustitucionales ... 32
Aproximación al Patrón Mutacional (cambios 4fold) ... 38
Discusión ... 39
Capítulo II: Uso de codones en Pelagibacteriaceae ... 42
Metodología ... 46
Datos ... 46
Identificación del Patrón de Sesgo en el Uso de Codones ... 46
Uso de codones ... 46
Análisis de Correspondencia ... 46
Identificación de Factores Asociados al Sesgo en el UCS ... 47
Fuentes de variación ... 47
Selección para la Traducción: Tasa de sustitución sinónima y estimaciones de CAI (Codon Adaptation Index) y MELP (MILC-based Expression Level Predictor) ... 47
Codones Óptimos ... 49
Resultados ... 50
Variación Inter - Genómica ... 50
- Variación Intra-genómica ... 50
- Contenido en G+C y sesgo en UC ... 59
-Selección para la traducción ... 59
- Codones Óptimos ... 63
Discusión ... 64
Conclusiones generales ... 68
Bibliografía ... 70

Resumen

La familia Pelagibacteriaceae, comprende organismos del clado SAR11 de alfa-proteobacterias oceánicas. Son organismos aerobios, heterótrofos, de los más pequeños conocidos y poseen un genoma muy pequeño y rico en A+T. En este trabajo se estudia la evolución de la composición nucleotídica y su efecto sobre el uso de codones en los genomas de *Pelagibacteriaceae.* Ambos procesos son de importancia por sus características genómicas y ecológicas: i) genoma pequeño y rico en AT ii) vida libre y gran tamaño poblacional.

Durante mucho tiempo se pensó que la mutación y la selección negativa eran los principales responsables de la composición genómica de organismos procariotas, pero cada vez se torna más evidente que la variabilidad en la composición nucleotídica entre estos organismos no se debe a un sesgo diferencial en el proceso de mutación y posterior selección purificadora. Por el contrario, se ha acumulado evidencia a favor de que procesos como la selección darwiniana o la recombinación, tienen consecuencias sobre el contenido en G+C de los organismos modificando la probabilidad de fijación de las mutaciones.

El patrón mutacional del clado, muestra un fuerte sesgo hacia A+T. La composición de bases de los genomas de Pelagibacteriaceae no es producto solamente de la mutación como se ha visto en genomas de clados emparentados, sino que hay indicios de selección sobre este factor. En principio la mutación y selección negativa tendrían un rol preponderante, acaso suficiente, en determinar el contenido en G+C de estos genomas. No obstante, existe espacio para que otros eventos selectivos puedan operar modulando suavemente a esas fuerzas.

A diferencia de la mayoría de los endosimbiontes, con los cuales comparte la composición nucleotídica y el pequeño tamaño del genoma, Pelagibacteriaceae, mostró variación intragenómica en el uso de codones, asociada incluso al nivel de expresión de los genes. Esta diferencia puede tener su origen en presiones selectivas sobre Pelagibacteriaceae que aunque débiles, se evidencian gracias a su gran tamaño poblacional efectivo estimado.

INTRODUCCIÓN GENERAL

La familia Pelagibacteriaceae, comprende organismos del clado SAR11 de alfa-proteobacterias oceánicas (Morris et al., 2002). El clado SAR11 fue descubierto en 1990 a través de la amplificación por PCR de muestras de ARNr 16S obtenidas del mar de los Sargazos (Giovannoni et al., 1990). Estudios posteriores encontraron miembros del clado a lo largo de todas las aguas oceánicas, en abundancias de hasta un 50% de la comunidad de bacterias heterótrofas, convirtiéndolo en un grupo clave en el ciclo del carbono en el océano (Brown et al., 2012; Morris et al., 2002). También se encontró variación en la abundancia de estos organismos de forma periódica (Morris et al., 2002), y estacional (Carlson et al., 2009). Un muestreo temporal de 16S por qPCR en aguas oceánicas del Pacífico subtropical, demostró que la abundancia de SAR11 se encuentra correlacionada de manera positiva con la abundancia de *Synechococcus* y diatomeas, y de forma negativa, con las concentraciones de fósforo y nitrato (Eiler et al., 2009; Sarmento & Gasol 2012).

Se ha sugerido que SAR11 tiene influencia a nivel global en el ciclado del C, N y S. Un estudio realizado en el mar de los Sargazos reveló que la mitad de las células de SAR11 asimilan compuestos de bajo peso molecular disueltos, aminoácidos y DMSP (dimetilsulfoniopropionato), en otros clados la proporción de células en esta actividad es mucho menor (Malmstrom et al., 2004). Otro estudio encontró que la mayoría de las células bacterianas incorporando glucosa a bajas concentraciones eran miembros de *Roseobacter* y SAR11. La contribución de ambos clados difiere en que *Roseobacter* presenta una población pequeña pero muy activa, mientras que la población de SAR11 es grande con una baja proporción de células activas. La importancia relativa de SAR11 disminuye con el aumento en la concentración de sustrato, en parte por el aumento en la proporción de células activas en otros clados, esto indica que SAR11 está particularmente adaptada a ambientes oligotróficos (Alonso & Pernthaler, 2006).

Los estudios de 16S realizados muestran a SAR11 como un grupo extremadamente diverso, globalmente distribuido y con variación periódica de abundancia (Brown et al., 2012). Dentro del clado se pueden diferenciar múltiples grupos conformados por ribotipos que muestran amplia microdiversidad (Acinas et al., 2004).

La información ecológica sugiere que el clado SAR11 consiste en por lo menos 12 ecotipos (Grote et al., 2012), que pueden diferenciarse ya sea filogenéticamente o por su presencia en el ambiente en diferentes profundidades y estaciones (Brown et al., 2012). Serán necesarios más estudios para determinar las principales variables ambientales asociadas a distribución de ecotipos de SAR11 porque a pesar de estos indicios, investigaciones mas recientes y detalladas, indican que la subdivisión de nichos no coincide con un modelo en el que la variación ambiental incluyendo solamente la estacionalidad y la profundidad, dirija la evolución de SAR11, sino que otros factores que no han podido ser identificados tienen un importante rol (Vergin et al., 2013).

Los miembros del clado SAR11 son organismos aerobios, heterótrofos, de los más pequeños conocidos (Morris et al., 2002; Sun et al., 2011) y poseen el genoma de menor tamaño conocido para bacterias de vida libre (Giovannoni et al., 2005a). Su tasa de crecimiento es baja, 0.19-0.71 d^{-1}, y aparentemente característica del

clado, ya que se mantiene a lo largo de un amplio rango de ecosistemas oceánicos (Sun et al., 2011; Teira et al., 2009).

Estos organismos pueden usar inusuales fuentes de energía. Poseen proteorhodopsina, una bomba de protones que depende de la luz, cuyo rol en la ecología de los océanos, suministrando energía para el metabolismo microbiano, es muy importante (Giovannoni et al., 2005a). Pueden usar también compuestos C1 (formato, metanol, metilaminas, formaldehído) (Tripp, 2013) lo que les adjudica un rol principal en la mineralización del carbono orgánico disuelto en el mar (Sun et al., 2011).

La reconstrucción metabólica realizada a partir del genoma de *Ca.* P. *ubique* (Fig. 1.1) evidenció la falta de muchos genes comunes entre los quemorganohetrótrofos (Carini et al., 2013). Se observó que los genes que codifican para enzimas involucradas en el metabolismo de carbohidratos, no son conservados a lo largo del clado, diferencias como estas se asocian a un proceso de partición de nicho, o existencia de los ecotipos observados (Grote et al., 2012). Entre otras particularidades, el análisis de múltiples genomas de SAR11, determinó que el uso de glucosa como fuente de C no es conservado en el clado, no todas las cepas presentan la vía ED (Enter-Doudoroff), y por lo tanto no todas usan glucosa sino piruvato como fuente principal de C (Tripp, 2013).

Una peculiaridad importante que presenta el grupo es la inusual ausencia de genes comunes para la biosíntesis de serina y glicina (Tripp et al., 2009), lo que las hace auxótrofas para estos metabolitos.

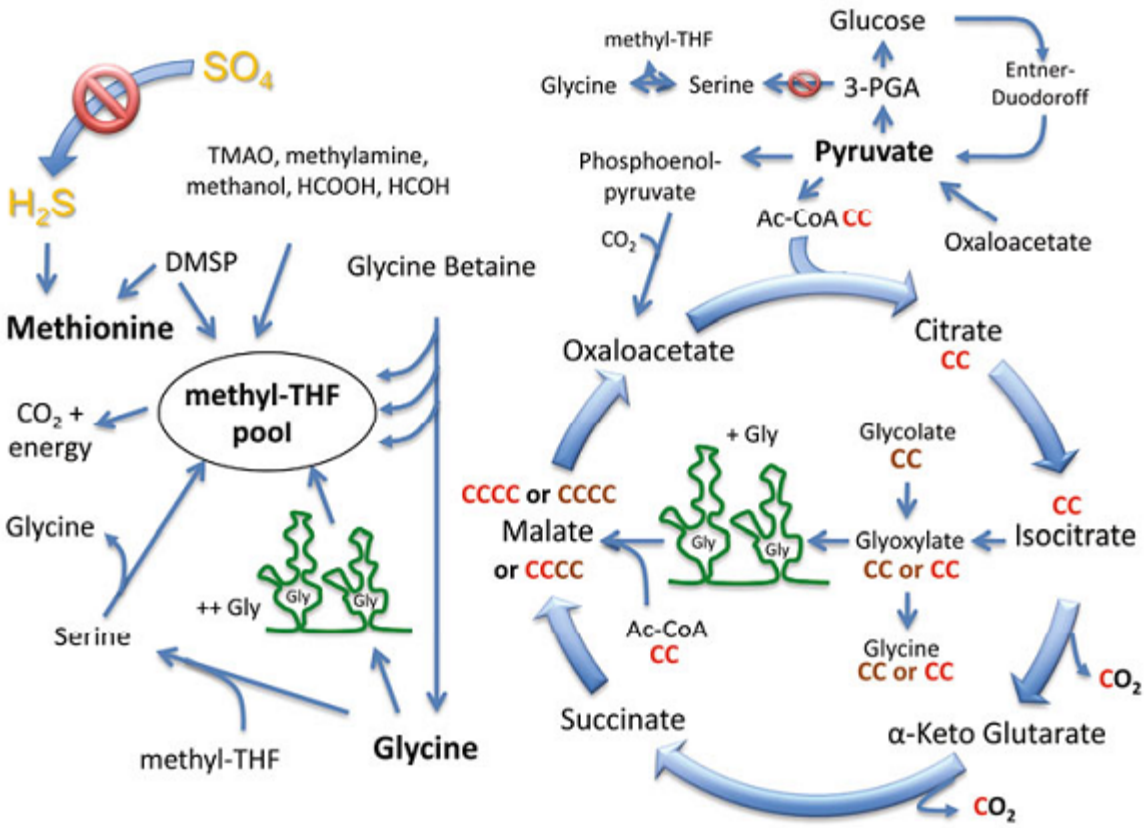


Figura 1.1: Esquema de las vías metabólicas principales reconstruídas para Ca. P ubique y sus particularidades. (Tripp et al. 2008)

Aunque las reconstrucciones metabólicas sugieren que existe potencial para la biosíntesis de glicina, hasta ahora las condiciones experimentales no han sido capaces de encontrar una sola fuente de carbono que pueda sustentar el crecimiento sin glicina en el medio (Tripp, 2013).

Ca. P. ubique también carece de los genes necesarios para la reducción y asimilación de sulfato, como resultado, requiere compuestos como metionina o 3-dimetilsulfoniopropionato para su crecimiento (Tripp et al., 2008), la dependencia de metabolitos presentes a muy bajas concentraciones en el ambiente, por parte de una bacteria de vida libre y muy abundante, es un hecho completamente inesperado. La adaptación a la vida en ambientes oligotróficos hizo que fuera un gran desafío el cultivo de estos organismos. Luego de muchos intentos se pudo cultivar la cepa HTCC1062 en el laboratorio, el medio artificial de cultivo que tuvo éxito, contiene sólo metionina, glicina, piruvato, vitaminas y sales inorgánicas (Rappé et al., 2002).

Como se dijo anteriormente, la reconstrucción del metabolismo de *Ca. P. ubique* demuestra como inusual requerimiento, el suministro de materia orgánica, esto se ha visto que es una consecuencia de la selección en favor de un genoma mínimo y simplicidad metabólica (Giovannoni et al., 2005b). La pérdida adaptativa de genes, o sea la pérdida de funciones por un organismo de vida libre bajo presión selectiva, es explicado por la hipótesis BQH (*Black Queen Hypothesis*) (Morris et al., 2012). La BQH dice que si un recurso costoso pero universalmente requerido fuga al ambiente, pueden desarrollarse presiones selectivas en subpoblaciones que lleven finalmente a la auxotrofía para el mismo (Tripp, 2013). Cualquier función que sea energéticamente costosa y su producto fugue al ambiente es un blanco potencial. Por ejemplo la fijación de N, la asimilación de nutrientes inorgánicos, o la biosíntesis de glicina y serina para las que Pelagibacteriaceae es auxótrofa, que cumplen los siguientes criterios: (i) el producto de la función es energética o nutricionalmente costoso, (ii) la función es realizada y por lo tanto el producto generado solo por una fracción de la comunidad, (iii) el producto es lo suficientemente permeable y por lo tanto se encuentra disponible en el ambiente para su uso por el resto de la comunidad, (iv) el producto es vital para la comunidad. La evolución de esta dependencia funcional genera por definición relaciones de comensalismo o mutualismo (Morris et al., 2012).

De acuerdo a esto, se plantea que la conformación metabólica en *Ca. P ubique* es adaptativa, no sólo porque es simple y requiere menor cantidad de nutrientes para la replicación, sino porque se ajusta a los ambientes planctónicos donde existe un flujo de materia orgánica bajo pero continuo (Carini et al., 2013). Esta hipótesis también se apoya en el hecho de que el mismo proceso puede explicar el proceso de reducción genómica sufrido por *Prochlorococcus*, otra especie de bacterioplancton marino.

A diferencia de la reducción genómica de los parásitos, debida a la presión de selección relajada y deriva génica, la selección en favor de minimizar el tamaño del genoma con la presión de conseguir un modo de vida más económico en un ambiente oligotrófico, es un factor importante en la evolución del genoma en SAR11. Este proceso se conoce como *genome streamlining* (Giovannoni et al. 2005; Grote et al. 2012;Viklund, Ettema, y Andersson 2012).

El tamaño del genoma en Pelagibacteriaceae es de aproximadamente 1.3 Mb, el más pequeño conocido para bacterias de vida libre, los genomas de tamaño por debajo de los 1.5 Mb son típicos de linajes asociados a un huésped (Viklund et al., 2012). A su vez, el contenido G+C de *Ca. P. ubique* se encuentra entre 0.29 y 0.30, esta

característica se observa en otras especies de bacterioplancton, *Prochlorococcus* y OM43, así como en organismos patógenos endosimbiontes y de genomas reducidos (Giovannoni et al., 2005b; Grote et al., 2012).

Por lo tanto, *Ca. P. ubique* comparte características con bacterias parásitos intracelulares obligatorios. De hecho, características como las que presenta la familia Pelagibacteriaceae no se observan comúnmente en organismos de vida libre, sino que son particulares de estos parásitos, como *Rickettsiales* y organismos simbiontes (Foerstner et al., 2005; Rocha & Danchin, 2002).

Estudios recientes en filogenómica ubican a SAR11 compartiendo un clado con *Rickettsiales* y las alfaproteobacterias ancestrales a las mitocondrias (Thrash et al., 2011).

De acuerdo a Rodríguez-Ezpeleta et al. 2012, el uso de modelos de sustitución poco sofisticados, que no tienen en cuenta el sesgo en la composición nucleotídica (la frecuencia de bases), generaron artefactos en la reconstrucción filogenética y llevaron a adjudicar como grupo hermano de Pelagibacteriaceae al clado *mitocondria/Rickettsiales* (Thrash et al. 2011). Los últimos resultados indican que el grupo SAR11 no está especialmente emparentado con el clado mitocondria/*Rickettsiales,* y que la reducción genómica y sesgo hacia un alto contenido en A+T, evolucionaron de forma independiente en ambos linajes (Rodríguez-Ezpeleta & Embley, 2012). El patrón de pérdida de genes difiere en ambos grupos, y en contraste con los genomas de *Rickettsiales* que contienen fagos, plásmidos, hasta un 30% de sus genes fragmentados, pseudogenes y ADN no codificante, el genoma de *Ca. Pelagibacter ubique* prácticamente no posee ADN no funcional y presenta la regiones intergénicas más pequeñas (3bp) de los genomas hasta ahora secuenciados (Viklund et al., 2012).

Otros argumentos resaltan que los nichos de ambos grupos, no podrían ser más distintos, así como la dinámica y tamaño poblacional que han evolucionado hacia polos opuestos. Mientras que los parásitos presentan especialización al huésped, tamaños de población pequeños y cuellos de botella poblacionales frecuentes (Toft and Andersson, 2010), SAR11 es de vida libre y tiene un tamaño poblacional estimado como uno de los más grandes entre las bacterias (Morris et al., 2002). Se puede distinguir entre dos tipos de estimaciones de tamaño poblacional. Nc que refiere al tamaño censal de la población, estimado a partir del conteo de organismos, y el Ne, tamaño poblacional efectivo, que es un concepto evolutivo y refiere a al tamaño poblacional necesario para generar la diversidad genética neutral observada (Luo et al., 2014). El producto de Ne y el coeficiente de selección, determina el destino de las mutaciones favorables y deletéras, y es por lo tanto un parámetro fundamental para determinar el potencial adaptativo de una población. El producto de Ne y la tasa de mutación determina la diversidad genética neutral de una población (Charlesworth, 2009).

Los miembros de la familia Pelagibacteriaceae, presentan un nivel de conservación de genes del *core* altísimo, es uno de los grupos de bacterias de vida libre más conservados. Los caracteres compartidos por el grupo incluyen sintenia de genes muy alta, una gran región hipervariable unida por genes rRNA, y pocos parálogos. La variación entre los genomas se debe principalmente a genes del metabolismo del fósforo, glicolisis y C1, lo que sugiere especialización adaptativa en el uso de los recursos, importante en el proceso de partición de nicho y divergencia de ecotipos dentro del clado (Grote et al., 2012). Otra importante característica de los genomas de Pelagibacteriaceae, que además la diferencia de clados de alfaproteobacterias parásitas, es el nivel de recombinación encontrado. Se han detectado altos grados de recombinación en poblaciones de *Pelagibacter*. La tasa de cambio nucleotídico como resultado de recombinación en relación al resultado de la mutación

puntual (r/m) es una medida directa del impacto relativo de la recombinación en la diversificación de secuencias (Vos y Didelot 2009). En Pelagibacteriaceae la tasa r/m es de 63.1 en (Vergin et al. 2007), indicando un importante rol de la recombinación en la variabilidad del clado (Vos and Didelot, 2009). Se ha demostrado que las bacterias de vida libre y grandes tamaños poblacionales acumulan inserciones/deleciones y rearreglos con mayor frecuencia respecto de las mutaciones puntuales, que los endosimbiontes de pequeños tamaños poblacionales donde la influencia de la HGT (Transferencia horizontal de genes) es despreciable (Boussau et al., 2004).

En Bacteria, la composición nucleotídica de los genomas es un carácter muy estudiado. El contenido en G+C en procariotas varía entre 17% y 75% (Belozersky and Spirin, 1958; Lightfield et al., 2011; Sueoka, 1961), y se encuentra directamente relacionado con la composición de aminoácidos, el tipo de proteínas que se producen, el tamaño de los genomas y muchos aspectos potencialmente relevantes a nivel fenotípico (Rocha and Feil, 2010).

Las bacterias de vida libre presentan, en promedio, mayores valores de contenido en G+C que las bacterias parásitas o simbiontes, y los mecanismos habitualmente planteados para explicar el enriquecimiento en A+T se asocian a la forma de vida parásita (Foerstner et al., 2005; Lind and Andersson, 2008; Rocha and Danchin, 2002).

No obstante, el aumento en el contenido en A+T de los genomas de SAR11 es un dato singular. Este aumento podría atribuirse al sesgo en la frecuencia de las mutaciones, los organismos que carecen la enzima de reparación del ADN 6-0- metilguanina-DNA metiltransferasa presentan proporciones GC/AT bajas, pero *P. ubique* es capaz de codificar para esta enzima lo cual sugiere que otros factores influyen en su composición nucleotídica (Giovannoni et al., 2005b).

Alternativamente, esta característica, puede pensarse como una ventaja selectiva, dado que disminuye el requerimiento de N en la síntesis de ADN, reduciendo el requerimiento de la célula por este componente (Giovannoni et al., 2005b). N y P son dos constituyentes del ADN que frecuentemente se presentan como recursos limitantes en el océano, esto puede actuar como una fuerza selectiva importante sobre la evolución del uso de codones y la composición de proteínas en los microorganismos adaptados a este ambiente resultando en uso de aminoácidos que minimicen el costo de N y uso de codones ricos en A+T (Grzymski and Dussaq, 2012).

El uso de los codones sinónimos no es azaroso, su frecuencia de ocurrencia difiere entre organismos y entre los genes de un organismo (Gouy and Gautier, 1982; Grantham et al., 1980), fenómeno conocido como sesgo en el uso de codones sinónimos (UCS). Los motivos subyacentes a la variación en el UCS, han sido explicados tanto en contexto neutralista como seleccionista. Entre los procariotas con composición de bases extrema, grupo al que pertenecen los miembros de Pelagibacteriaceae, el uso de codones sinónimos parece estar dominado exclusivamente por el sesgo mutacional y la detección de selección sobre este rasgo se ha mantenido esquiva, los ejemplos más conocidos son *Mycoplasma capricolum* (GC% = 25), *Rickettsia prowazekii* (GC% = 29) and *Micrococcus luteus* (GC% = 72) (Andersson and Sharp, 1996; Ohama et al., 1990; Ohkubo et al., 1987). Sin embargo, en algunos casos como el de *Chlamydya trchomatis* y *Clostridium*, especies con genomas pobres en G+C, se detectaron indicios de selección operando sobre el uso de codones (Musto et al., 2003; Romero et al.,

2000).

El estudio de la evolución de la composición nucleotídica, así como su efecto sobre el uso de codones del genoma cobra importancia en Pelagibacteriaceae dadas sus características genómicas y ecológicas: i) genoma pequeño y rico en AT ii) vida libre y gran tamaño poblacional.

Esta tesis de maestría consta de dos capítulos, en el primero se analiza la composición nucleotídica de los genomas y su evolución a partir del estudio de patrones sustitucionales en el clado. El segundo capítulo explora la variación en el uso de codones sinónimos en *Ca P. ubique* e investiga las posibles causas de la variación que tradicionalmente se debaten entre la mutación y la slección natural.

Objetivo general

Estudiar los patrones evolutivos que moldean la particular composición nucleotídica de los genomas de la familia Pelagibacteriaceae.

Objetivos específicos y preguntas asociadas

Capítulo I

1. Estudiar los patrones de sustitución en el clado y en los subclados que puedan definirse dentro de la familia Pelagibacteriaceae.
   - ¿Cómo varían los patrones de sustitución en el tiempo? ¿Cómo varían según las características funcionales de los sitios (sinónimos, no sinónimos, *4fold*) ¿De qué procesos evolutivos son producto?
2. Definir el sesgo mutacional que opera en el clado y su relación con la composición nucleotídica del genoma.
   - ¿Son la mutación y la selección negativa los únicos involucrados en la determinación del contenido G+C en Pelagibacteriaceae?
3. Eventualmente, buscar fuerzas selectivas que estén manteniendo la composición nucleotídica de los genomas.

Capítulo II

1. Estudiar el patrón de uso de codones en el clado Pelagibacteriaceae.
   ¿Se observan variaciones ínter o intra genómicas en el clado?
2. Detectar si existen fuerzas selectivas afectando el uso de codones sinónimos (UCS) en Pelagibacteriaceae
   ¿Existe selección operando a nivel de la traducción?

Capítulo I: Patrones Sustitucionales en *Pelagibacteriaceae*

Introducción

Evolución de la composición de bases de genomas procariotas

Lograr discernir la selección natural de procesos neutrales sigue siendo un problema central en el área de la evolución molecular. Una de las controversias más importantes en este sentido es sobre la causa de variabilidad en la composición de bases de los genomas procariotas.

Como se mencionó anteriormente, el contenido en G+C en procariotas muestra una amplia variación entre especies (Belozersky and Spirin, 1958; Sueoka, 1961). Se han descrito especies como el endosimbionte *Carsonella ruddii*, con un contenido en G+C de 16,5%, hasta *Anaeromyxobacter dehalogens*, una bacteria cuyo contenido en G+C es de 75% (Hershberg and Petrov, 2010a). En general se observa que los genomas actualmente secuenciados muestran una importante correlación entre el contenido en G+C y el tamaño del genoma (Andersson and Kurland, 1998), los genomas pequeños pertenecientes generalmente a bacterias endosimbiontes son ricos en A+T, mientras que los genomas grandes que corresponden a bacterias de vida libre y ambientes complejos son ricos en G+C (Foerstner et al., 2005). De todas formas existen excepciones como el caso de el endosimbionte *Candidatus Hodgkinia cicadicola* que posee uno de los genomas más pequeños conocidos con un contenido en G+C de 58.4% (McCutcheon et al., 2009).

Dado que la mayoría de las bacterias tienen poco ADN intergénico y que se ha observado correlación entre el porcentaje de G+C en las primeras y segundas posiciones de los codones, y el G+C genómico, se establece que la composición de bases de un genoma se encuentra directamente relacionada con la composición de aminoácidos, el tipo de proteínas que se producen, el tamaño de los genomas y muchos aspectos potencialmente relevantes a nivel fenotípico (Rocha and Feil, 2010; Lightfield, Fram, y Ely 2011).

Qué factores generan la variabilidad en el contenido G+C, tanto entre genomas como dentro de un mismo genoma (entre genes o entre regiones codificantes y no codificantes), y a través de qué mecanismos este sesgo se modifica en respuesta a la variabilidad del ambiente o constitución genética del organismo, son aún interrogantes (Lind and Andersson, 2008).

La primera explicación que se planteó para explicar la variabilidad de composición nucleotídica entre genomas de distintas especies, fue la existencia de sesgo mutacional diferencial. Lo que se propone es que la fuerza principal moldeando el contenido nucleotídico en Bacteria es la mutación, direccionada distintamente, en diferentes especies (Sueoka, 1961). El sesgo mutacional afectaría tanto a sitios sinónimos como no-sinónimos, por lo tanto la evolución de las proteínas estaría influenciada principalmente por fuerzas neutrales (Sueoka, 1961). A esto se suma la interpretación de la variación en composición entre sitios, dado que las terceras posiciones muestran un contenido extremo, y tienen menos restricciones selectivas (Muto and Osawa, 1987), un postulado bastante aceptado es que la composición de bases de los genomas se debe a efectos mutacionales neutrales y tiene poca relevancia adaptativa (Rocha and Feil, 2010). Bajo este supuesto, los clados que son ricos en G+C, son clados en los cuales la mutación ha estado constantemente sesgada hacia

G+C, mientras que en los clados ricos en A+T, la mutación ha estado constantemente sesgada hacia A+T. Esto implica que el sesgo mutacional en bacterias es extremadamente variable (Hershberg and Petrov, 2010a).

Pero existe otro punto de vista, que sostiene que la selección natural tiene un rol preponderante sobre la composición de bases de los genomas. Las hipótesis en este contexto sostienen que no es la variación en el sesgo mutacional que lleva a una variación en el contenido nucleotídico de los genomas, sino la variación en la probabilidad relativa de fijación de las mutaciones (Bentley and Parkhill, 2004). La variación puede entonces originarse debido a diferencias en la fuerza y dirección de la selección natural (Rocha and Feil, 2010). La selección natural afecta la probabilidad de fijación de un alelo basado en la ventaja o desventaja, o sea el fitness del mismo, y el tamaño poblacional efectivo del organismo en cuestión (Bentley y Parkhill 2004).

Buscando justamente el nexo entre el contenido G+C de los genomas y las fuerzas selectivas operantes, muchos factores eco-fisiológicos se han asociado a la variación en composición nucleotídica tanto entre como dentro de los genomas (Rocha y Danchin 2002; Naya et al. 2002; Musto et al. 2004; Singer and Ames, 1970;C E McEwan, 1998)(Tabla 1.1). No es sorprendente que dada la complejidad de la ecología microbiana, ninguna de estas teorías provea de una explicación universal de la naturaleza de las fuerzas selectivas que otorgan ventaja a un alto o bajo contenido en G+C (Lind and Andersson, 2008).

Tabla 1.1: Resumen de variables eco-fisiológicas asociadas a la variación en la composición de bases en genomas procariotas.

Variable	**Explicación**	**Problemas**
Ambiente (Foerstner et al., 2005)	Diferentes ambientes contienen bacterias diferencialmente enriquecidas en GC	Los mecanismos tras esta variable no son claros y podrían ser en realidad combinaciones de muchas variables
Largo de los genes (Xia et al., 2003)	El alto contenido en GC favorece genes más largos reduciendo la frecuencia de mutaciones non-sense. El largo de los genes se correlaciona con su contenido GC.	El contenido GC genómico está, como mucho, correlacionado con el largo promedio de genes, el cual no varía mucho entre genomas.
Largo del genoma (Heddi et al., 1998)	La reducción genómica frecuentemente se observa como consecuencia de reducción de tamaño efectivo poblacional. Genomas pequeños poseen bajo contenido en GC y los genomas grandes, son ricos en GC.	Debido a que la densidad de genes en procariotas es muy alta, el largo del genoma es proxy de muchas variables. Esto hace difíciles las interpretaciones biológicas.
Presiones mutacionales (Hershberg and Petrov, 2010a; Hildebrand et al., 2010; Lind and Andersson, 2008)	Las mutaciones son ricas en AT, la pérdida de genes de reparación lleva a enriquecimiento en AT	No explica el gap composicional entre patrones mutacionales y composición real de los genomas. No explica la existencia de genomas ricos en GC.
Fijación de Nitrógeno (C E McEwan, 1998)	Selección que favorece el ahorro en el uso de N en ADN y ARN (ambas macromoléculas ricas en N). El contenido en GC es más alto en fijadores de Nitrógeno.	El contenido en GC es más alto en dos géneros de aerobios fijadores pero más bajo en dos géneros anaerobios. La mayoría de los procariotas no son fijadores.
Oxígeno (Naya et al., 2002)	Los tres puentes de H que unen las bases G y C generan una molécula de ADN más fuertemente empaquetada y menos susceptible a la oxidación. Las G sinónimas podrían tener un importante rol en ambientes oxidantes. Aerobios son ricos en GC.	En general, G es la base mas susceptible a la oxidación.

Parasitismo (Rocha and Danchin, 2002)	Patógenos, plásmidos, elementos transponibles y bacteriófagos están enriquecidos en AT, cuyo costo es menor y son más abundantes.	No explica la existencia de genomas ricos en GC.
Composición y plegamiento de proteínas (Seligmann, 2003)	Codones ricos en GC codifican aminoácidos más baratos. La susceptibilidad a la oxidación y la estabilidad del plegamiento covarían con el contenido GC.	La selección sobre GC no debería estar dirigida por la composición de proteínas ya que la selección purificadora sobre el contenido GC es mayor en sitios intergénicos y degenerados.
Plegamiento del ARN (Daubin and Perrière, 2003)	Prácticamente todas las posiciones en los genomas bacterianos son transcritas, y las estructuras de ARN ricas en GC son más estables.	Los genes básicos (core genes) tienen un contenido de bases homogéneo, las excepciones no forman parte de operones altamente expresados. Los operones ADNr, altamente expresados en crecimiento exponencial, son ricos en GC en genomas pobres en GC y viceversa.
Temperatura (Musto et al., 2004)	La riqueza en GC aumenta la termoestabilidad del ADN, ARN y del apareamiento codón-anticodón.	La asociación entre temperatura óptima de crecimiento y contenido en GC genómico es débil. Cepas de Pasteurella, que evolucionaron a altas temperaturas son ricas en AT.
Radiación UV (Singer and Ames, 1970)	Dinucleótidos ricos en AT son más susceptibles a la formación de dímeros de pirimidinas bajo radiación UV.	No se encuentra selección en contra de los dinucleótidos susceptibles.
Plegamiento del ADN (Travers, 2004)	GC aumenta la estabilidad	No está claro que la estabilidad debida al alto contenido en GC esté siendo seleccionada
Conversión génica (Duret and Galtier, 2009)	La conversión resultante de recombinación entre secuencias mal apareadas aumenta el contenido en GC. Las zonas de alta recombinación en E.coli son ricas en GC.	Las tasas de recombinación/mutación observadas no correlacionan con GC.

Investigaciones recientes desde otro enfoque, se abocaron al estudio del sesgo mutacional en procariotas encontrando resultados que socavan de manera importante las hipótesis neutralistas. Estos estudios se proponían estudiar la dirección del sesgo mutacional en un amplio rango de procariotas, dado que si la evolución del contenido G+C es principalmente moldeada por las mutaciones, deberían encontrarse sesgos con muy diferentes sentido e intensidad. Al contrario lo que encontraron fue un sesgo mutacional universal dirigido hacia A+T. Los resultados mostraron que el patrón mutacional es en realidad muy similar entre bacterias y que las mutaciones más frecuentemente observadas son las transiciones de C/G a T/A generando un sesgo mutacional hacia A+T en todas las especies estudiadas, incluso en bacterias con contenidos altos en G+C (Hershberg and Petrov, 2010a; Hildebrand et al., 2010).

De acuerdo a estos resultados, los organismos pobres en G+C, lo son debido al sesgo mutacional mientras que los ricos, estarían subiendo su contenido en G+C a través de otros mecanismos evolutivos, que pueden ser tanto selección natural como g-BGC (*del inglés GC - Biased Gene Conversion*). El fenómeno conocido como g-BGC, favorece los nucleótidos G/C en sitios polimórficos durante la conversión de los intermediarios de recombinación (Duret and Galtier, 2009). Aunque g-BGC como proceso no está relacionado con la selección natural, afecta la probabilidad de fijación de un alelo de forma similar, puede de hecho dejar marcas en el

genoma que imitan las de la selección positiva, o interferir con la selección promoviendo la fijación de alelos deletéreos (Ratnakumar et al., 2010). Existe amplia evidencia de este proceso a nivel eucariota y recientemente se ha propuesto que en Bacteria y Archea, g-BGC tiene también un importante rol (Lassalle et al., 2015). Hildebrand et al, descartaron la existencia de g-BGC en su análisis de sesgo mutacional porque el exceso de mutaciones G/C→A/T no cambió al analizar sets de datos con evidencia de recombinación y tampoco encontraron correlación entre el contenido en G+C y la tasa de recombinación en el rango de especies estudiadas (Hildebrand et al., 2010). Las nuevas evidencias indican que la recombinación afecta el contenido G+C genómico de la mayoría de los phylum bacterianos (Lassalle et al., 2015).

La implicación tanto de la selección natural como de g-BGC en la evolución del contenido en G+C de los genomas, torna el estudio composicional de la familia Pelagibacteriaceae muy interesante. Dadas sus características, cualquiera de estos mecanismos podrá estar en funcionamiento en la evolución de los genomas de la familia. Pelagibacteriaceae tiene un tamaño poblacional efectivo (Ne) estimado muy grande, que permite un gran impacto de la selección aún cuando ésta es débil. También se ha estimado que posee altas tasas de recombinación por lo tanto podría estar influyendo en su contenido G+C genómico el proceso de g-BGC.

Los organismos ricos en A+T que fueron estudiados tanto en Hildebrand et al., 2010 como en Hersberg et al., 2010 eran parásitos intracelulares obligatorios de genomas muy reducidos, altas tasas de mutación y pequeños tamaños poblacionales. Los patógenos que han divergido recientemente, como las cepas monoclonales, que poseen un exceso de polimorfismos A+T, presentan selección relajada y baja tasa de recombinación, lo cual podría explicar la disminución en el %G+C de sus genomas (Andersson and Kurland, 1998; Rocha and Feil, 2010; Toft and Andersson, 2010). Ya se ha observado que la adopción de asociaciones con eucariotas, como la endosimbiosis y el parasitismo, llevan a una reducción del tamaño poblacional efectivo y enriquecimiento en A+T, posiblemente debido a una menor capacidad de eliminación de las mutaciones GC -> AT (Rocha and Feil, 2010). Excepto por la pérdida de mecanismos de reparación, todas las demás explicaciones de enriquecimiento en A+T se basan en la forma de vida parásita (Foerstner et al., 2005; Lind and Andersson, 2008; Rocha, 2008; Rocha and Feil, 2010), y no aplican a la familia Pelagibacteriaceae, las características genómicas de la misma generan incluso una contradicción con las explicaciones al enriquecimiento en A+T anteriormente propuestas.

Determinar la causa subyacente de la variación composicional en los genomas procariotas es un problema fundamental, si la variación es resultado de la selección entonces el contenido en G+C es en sí mismo un carácter que contribuye al *fitness* de los organismos, mientras que si la variabilidad es moldeada por mecanismos no adaptativos, entonces caracterizarlos es de gran importancia para la detección de selección (Lassalle et al., 2015).

Metodología

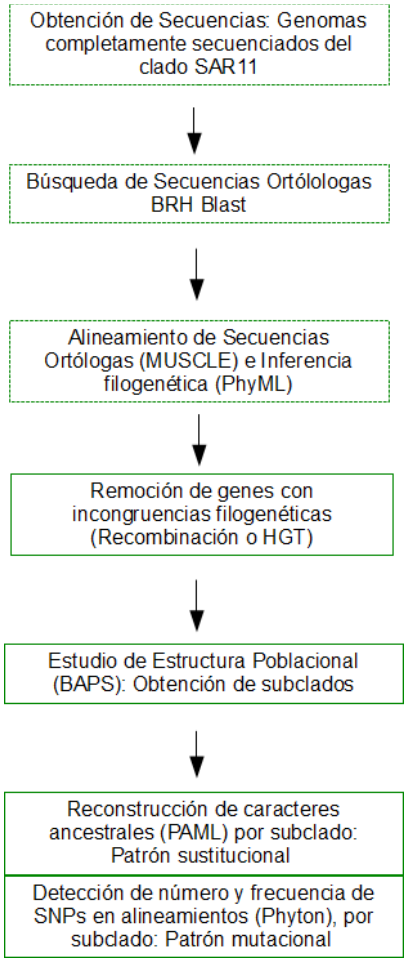
DATOS

Se trabajó con 16 genomas completamente secuenciados pertenecientes a Pelagibacteriaceae. Los mismos se listan en la Tabla 1.2. Los genomas se descargaron del portal *Joint Genome Institute.*

Tabla 1.2: especies de la familia Pelagibacteriaceae en estudio.

Genome Name / Sample Name	Geographic Location	Genome Size	Gene Count	GC
Candidatus Pelagibacter-like (SAR11) HIMB140	Kaneohe Bay, Oahu, Hawaii.	1437930	1535	.29
Candidatus Pelagibacter ubique SAR11 HTCC1002	Coastal Pacific Ocean, Newport, Oregon, cold	1327604	1428	.3
Candidatus Pelagibacter ubique HTCC7214	BATS Bermuda-Atlantic-Time-Series study site	1375060	1475	.29
Candidatus Pelagibacter ubique HTCC7217	BATS Bermuda-Atlantic-Time-Series study site	1433611	1537	.29
Candidatus Pelagibacter-like SAR11 HIMB4	Kaneohe Bay, Oahu, Hawaii.	1382291	1489	.29
Candidatus Pelagibacter ubique SAR11 HTCC8051	Coastal Pacific Ocean, Newport, Oregon, cold	1395115	1498	.29
Candidatus Pelagibacter ubique HTCC1040	Coast of Oregon	1274624	1381	.3
Candidatus Pelagibacter sp. IMCC9063	Svalbard, Norway	1284727	1482	.32
Candidatus Pelagibacter sp. HTCC7211	BATS Bermuda-Atlantic-Time-Series study site	1456888	1576	.29
Candidatus Pelagibacter ubique SAR11 HTCC1062	Coast of Oregon, cold	1308759	1394	.3
Candidatus Pelagibacter-like (SAR11) HIMB083		1395997	1505	.29
Candidatus Pelagibacter ubique SAR11 HTCC9022	Coastal Pacific Ocean, Newport, Oregon, cold	1362422	1464	.3
Candidatus Pelagibacter ubique HIMB058		1115049	1269	.3
Candidatus Pelagibacter ubique SAR11 HTCC1013	Coastal Pacific Ocean, Newport, Oregon, cold	1302704	1411	.3
Candidatus Pelagibacter ubique SAR11 HTCC9565	NE Pacific Open Ocean	1279674	1386	.29
Candidatus Pelagibacter ubique HTCC1016	Coastal Pacific Ocean, Newport, Oregon, cold	1297098	1404	.3

DISEÑO EXPERIMENTAL



Búsqueda de genes ortólogos

Los conceptos de genes ortólogos y parálogos son de importancia en el entendimiento de la evolución de los genomas y el surgimiento de novedades genéticas. Los ortólogos surgen de un proceso de especiación, y son producto de descendencia vertical desde un ancestro común a diferentes especies. Los parálogos son el resultado de eventos de duplicación génica y se los ha propuesto como una fuente de innovación (Fitch, 1970). Saber reconocer la diferencia entre ambos es muy importante especialmente cuando se quiere reconstruir una filogenia precisa o determinar las fuerzas de la selección natural.

La tarea de encontrar genes homólogos a una secuencia de interés (*query*) en una base de datos que contiene muchas otras secuencias (*subjects*), se puede conceptualizar como: conseguir el mejor alineamiento posible del *query* con los *subjects*, otorgar un *score* a estos alineamientos, y elegir aquellos cuyos *scores* sobrepasan un umbral dado (Ward and Moreno-Hagelsieb, 2014).

BLAST (del inglés: Basic Local Alignment Search Tool) RBH (del inglés: Reciprocal Best Hits) es un método computacional muy común para predecir presuntos genes ortólogos. Los RBH son un indicador común para ortología en genómica comparativa, se encuentran cuando las proteínas codificadas por mas de un gen, cada uno en diferentes genomas, son entre ellas el mejor *hit*, con el mejor *score* (Ward and Moreno-Hagelsieb, 2014) (Fig.1.2).

Las secuencias ortólogas entre las 16 especies analizadas se identificaron realizando un BLASTP *query* (Altschul et al., 1990) sobre todas las secuencias codificantes siguiendo el criterio de RBH y un *e-value* mínimo de 1e-5.

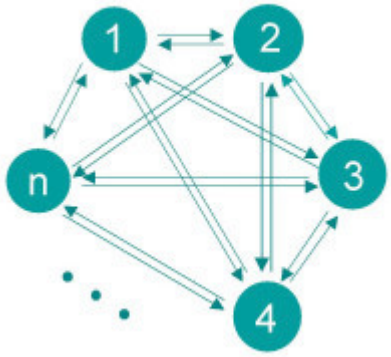


Figura 1.2: Esquema que ilustra la búsqueda de ortólogos por RBH (Poptsova and Gogarten, 2007)

Reconstrucción filogenética

Las secuencias ortólogas se alinearon con MUSCLE (Edgar, 2004) a nivel de proteínas y con tranalign (Rice et

al., 2000) a nivel nucleotídico. Tranalign genera un alineamiento de ADN a través de el correspondiente alineamiento proteico (Bininda-Emonds, 2005).

La inferencia filogenética se realizó sobre el alineamiento a nivel de nucleótidos de todos los genes ortólogos concatenados, a través del programa PhyML (Guindon et al. 2010). Este programa utiliza el método de máxima verosimilitud (ML) para reconstruir el árbol que mejor explique la historia evolutiva de las secuencias. A diferencia de métodos como máxima parsimonia, la aproximación probabilista de ML permite testar varias hipótesis, como por ejemplo si dos topologías de un árbol son significativamente distintas.

El proceso de inferencia filogenética por ML, comienza por generar una matriz de distancias evolutivas entre las secuencias, esto requiere la comparación entre todos los pares de secuencias. Luego se construye un árbol inicial, usando BIONJ. Se calculan las verosimilitudes condicionales para todos los sitios, subárboles (dentro del árbol) y para el árbol en su totalidad. Para aumentar la verosimilitud de la filogenia de inicio, se ajustan los parámetros del modelo evolutivo de sustituciones (tasas de transición y trasversión, y la distribución gamma que modela la variación de tasas entre sitios). Estos parámetros son re-estimados en cada proceso de refinamiento de la filogenia. El árbol es entonces refinado hasta alcanzar convergencia, en cada evento de refinamiento se chequea que la verosimilitud aumente. El refinamiento del árbol concluye cuando no hay más cambios posibles y el largo de las ramas es estable (Guindon and Gascuel, 2003).

El modelo de sustitución nucleotídica usado fue HKY85 (Hasegawa et al., 1985). El modelo incorpora múltiples parámetros para crear una simulación más realista del comportamiento de las secuencias, el modelo asume que la tasa de sustitución varía entre nucleótidos ($\pi A \neq \pi C \neq \pi T \neq \pi G$), también atribuye diferentes frecuencias a transiciones y trasversiones. Yang et al. 2003 demostró que los modelos que tienen en cuenta estas características, aumentan significativamente la verosimilitud del árbol.

Búsqueda de incongruencias filogenéticas en genes ortólogos

Un supuesto común de los métodos que realizan reconstrucción de secuencias ancestrales *(ASR: ancestral sequence reconstruction*), es que todas las posiciones en el alineamiento han evolucionado bajo la misma filogenia, y existe por lo tanto un único ancestro común más cercano (MRCA) para todas las secuencias de la muestra. El hecho es que pueden existir regiones recombinantes u obtenidas por THG en las secuencias, y por lo tanto lo anterior no se cumple, ya que éstas tendrán una historia evolutiva diferente (Arenas and Posada, 2010a).

Detección de genes recombinantes

Para disminuir los errores en la estimación de estados ancestrales, se debe tener en cuenta entonces la estructura de la población, particularmente el nivel de intercambio de material genético entre las especies. Para esto se evaluó el nivel de recombinación entre las especies, esto nos permite evaluar qué porción del genoma proviene de THG.

SplitsTree (Huson and Bryant, 2006) es un programa que permite reconstruir árboles y redes filogenéticas, las redes a diferencia de los árboles, son una representación de una presunta historia evolutiva que puede

involucrar eventos de reticulación. Las redes filogenéticas se pueden calcular a partir de un amplio rango de datos, alineamientos múltiples, matrices de distancia, grupos de árboles entre otros (Huson and Bryant, 2006).
A partir del concatenado de los alineamientos de cada gen, se construyó una red que representa la historia evolutiva de las secuencias en presencia de eventos como transferencia horizontal o recombinación.
En este caso, las redes proveen una representación explícita de la historia evolutiva de los genomas, se representan como un árbol filogenético con ramas adicionales, los nodos internos representan secuencias ancestrales, los que tienen más de dos ancestros corresponden a eventos como recombinación (Huson and Bryant, 2006). Esta aproximación es de carácter exploratorio, si el resultado muestra indicios de recombinación entre las especies, entonces se procede a detectar los genes afectados por eventos de recombinación y removerlos de la ASR.
La detección de genes recombinantes se realizó con el programa PhiPack que implementa el test Phi (Bruen et al., 2006). Lo que permite hacer este test es determinar si las relaciones filogenéticas o grados de similitud entre secuencias del set de datos, varían a lo largo de las mismas de forma congruente con la existencia de un evento de recombinación. En esencia, busca sitios filogenéticamente incompatibles, o sea sitios dentro de un alineamiento cuyo patrón de variación apoya diferentes filogenias. PhiPack detecta los genes recombinantes, comparando la frecuencia y distribución de los sitios filogenéticamente incompatibles observados, versus los esperados en ausencia de recombinación (Martin et al., 2011).

Congruencia en las reconstrucciones filogenéticas

Este procedimiento consiste en comparar la similitud entre la topologías del árbol construido con el concatenado de los genes ortólogos, que puede considerarse la filogenia del clado, y el árbol de cada gen. Los genes ortólogos que respetan la filogenia del clado son los obtenidos únicamente por herencia vertical, y los cambios que entre ellos se puedan detectar van a deberse solo al proceso de sustitución nucleotídica.
PhyML puede leer uno o varios árboles ingresados al programa como *input*. Estos árboles son generalmente usados como estimaciones de ML iniciales, para ser subsecuentemente ajustados por el algoritmo de reconstrucción filogenética. En este caso, se realiza una reconstrucción filogenética para cada gen ortólogo agregando al *input* el árbol del clado. Luego de finalizada la reconstrucción filogenética para el gen, se obtiene un valor de verosimilitud distinto al inicial, y a través de un test estadístico se determina si estos valores son significativamente distintos.
CONSEL (Shimodaira and Hasegawa, 2001) es un paquete de programas escritos en lenguaje C. Se utiliza para calcular el p-valor en problemas de selección de modelos para evaluar la confianza, aunque puede aplicarse a cualquier tipo de problemas de selección, está diseñado para el problema de selección de árboles filogenéticos. CONSEL no estima árboles sino que trabaja con el *output* de programas que sí lo hacen, como PhyML. Los p-valores son calculados usando varios tests: Kishino-Hasegawa, Shimodaira-Hasegawa, *weighted* Shimodaira-Hasegawa (Shimodaira and Hasegawa, 2001).
Una vez obtenidos con CONSEL los p-valores, se tiene un criterio con el cual evaluar la similitud de las topologías y retirar o no un gen del set de secuencias ortólogas.

Estimaciones dN y dS

Tradicionalmente, las tasas de sustitución sinónimas (dS) y no sinónimas (dN) se definen en el contexto de la comparación entre dos secuencias de ADN, siendo dS el número de sustituciones sinónimas por sitio sinónimo, y dN, el número de sustituciones no sinónimas por sitios de igual característica (Yang y Bielawski 2000). Los sitios sinónimos son aquellos en los que un cambio nucleotídico, no genera variación en el aminoácido codificado, o sea, no tiene impacto a nivel de proteína, los cambios en sitios no sinónimos si alteran la identidad del aminoácido codificado. La estimación de ambas tasas es importante para el entendimiento de la dinámica evolutiva a nivel molecular (Yang and Nielsen, 2000).

La tasa dN/dS, también conocida como w, cuantifica el tipo y la fuerza de la selección operando en las secuencias. Diferentes presiones selectivas operan sobre los sitios sinónimos y no sinónimos, al comparar los cambios sinónimos, asumidos neutrales, y los no sinónimos, expuestos a la selección debido a que sus modificaciones alteran la identidad de los aminoácidos y la composición de las proteínas (Mugal et al., 2014) se puede revelar la dirección y magnitud de la selección natural actuando sobre una proteína (Kimura, 1977).

Existen dos tipos de métodos para estimar las tasas de sustitución. Por un lado los que implican contar los sitios sinónimos (S) y no sinónimos (N) en las secuencias, contar las diferencias sinónimas y no sinónimas entre las secuencias y corregir por múltiples sustituciones en un mismo sitio. Estos métodos -Miyata and Yasunaga (Miyata and Yasunaga, 1980),y Nei and Gojobori (Nei y Gojobori 1986)- realizan asunciones simplistas sobre el proceso de sustitución nucleotídica, como que las tasas de transiciones (T↔C and A↔G) y trasversiones (T,C ↔A,G) son las mismas, y que las secuencias tienen un uso de codones uniforme. Dado que a nivel de la tercera posición de los codones las transiciones tienen mayor probabilidad de ser silenciosas que las trasversiones, es importante tener en cuenta la diferencia entre ambas tasas de cambio (Yang y Bielawski 2000; Z. Yang y Rannala 2012).

Por otro lado, existen los métodos que estiman dS y dN por máxima verosimilitud basados en modelos explícitos de sustitución de codones (Goldman and Yang, 1994). Los parámetros en estos modelos -tiempo de divergencia (t), tasa de transición/trasversión (k), dN/dS (v)- son estimados a partir de los datos por ML y usados para calcular dN y dS de acuerdo a sus definiciones (Yang y Bielawski 2000). En este caso, la teoría de probabilidad asociada al método consigue la difícil tarea de estimar los parámetros mutacionales como k, y ponderar las diferentes vías de cambio entre codones.

En general se recomienda el uso de métodos basados en estimaciones por ML, esto se puede llevar a cabo en Codeml, con los mismos datos y al tiempo que se realiza la ASR.

Se estimaron dS y dN para cada *cluster* de genes ortólogos y para el concatenado de los mismos. Con esto se pretende obtener una visión general de las fuerzas evolutivas operando en el set de genes ortólogos, y el nivel de divergencia que existe entre las especies.

Estructura Poblacional

Se trata de determinar si existe estructura genética en la población de secuencias, identificar si existieron o

existen barreras para el flujo génico, de manera tal de determinar grupos de especies dentro de los cuales estudiar el patrón de sustituciones. Se utilizó BAPS (Corander et al., 2003), una aproximación bayesiana para inferir estructura genética en una población usando secuencias nucleotídicas. BAPS usa los SNPs (single nucleotide polimorphisms) obtenidos a partir del alineamiento de los genes ortólogos para identificar grupos estructurados en los cuales llevar a cabo el estudio del sesgo, identifica, si existen, subpoblaciones en la muestra total de secuencias.

Patrón Sustitucional

El resultado de Codeml genera un archivo a partir del cual, se pueden identificar la cantidad y la polaridad de las sustituciones a nivel de sitios sinónimos y no sinónimos. Cuando se analizan secuencias codificantes, se posee la ventaja de distinguir entre estos dos tipos de sitios. Las sustituciones sinónimas o silenciosas son aquellas que no modifican el aminoácido codificado, las no sinónimas sí lo hacen.

En este caso, se clasificaron como sinónimos aquellos cambios ocurridos en las terceras posiciones de los codones degenerados, conocidos como *4-fold,* dado que sea cual sea la modificación observada, el aminoácido codificado no cambia. Los cambios no sinónimos, quedaron definidos como aquellos que ocurren en la primera y/o segunda posición de los codones, en estos sitios las variaciones generan cambios a nivel proteico. Como la selección natural actúa a nivel de proteínas principalmente, las mutaciones sinónimas y las no sinónimas están bajo presiones selectivas diferentes y son fijadas a tasas diferentes (Yang, 2007).

Existen 12 cambios posibles entre las bases del ADN:

AT | AC | AG | CT | CA | CG | GT | GC | GA | TA | TC | TG

Para cada uno de ellos se contabilizó el número de ocurrencias, que se transformó en la frecuencia de ocurrencia de los mismos al estandarizar por la composición de bases:

AT | AC | AG / n° sitios A
CT | CA | CG / n° sitios C
GT | GC | GA / n° sitios G
TA | TC | TG / n° sitios T

Una vez que obtenemos la frecuencia de sustituciones podemos calcular el contenido G+C en equilibrio, entendido como el contenido en G+C que presentarían las especies si la composición de bases del genoma estuviera determinada solamente por el sesgo mutacional (Hershberg y Petrov 2010) o sea estuviera en equilibrio mutacional.

Si la selección no afecta de manera significativa al contenido nucleotídico de los genomas, entonces éstos deberían de encontrarse en equilibrio mutacional, o sea, deberían presentar el mismo número de cambios de GC -> AT y de AT -> GC (Hershberg and Petrov, 2010b). Se calculan las tasas de cambio de GC -> AT y de AT -> GC, y con ellas se estima el GC en equilibrio:

rAT->GC = # AT->GC/# AT

rGC->AT = # GC->AT/# GC

GCeq = rAT->GC/ (rAT->GC + rGC->AT)

Aproximación al Estudio de Patrones Mutacionales

El objetivo es detectar diferencias nucleotídicas entre las secuencias que permitan aproximarnos al patrón mutacional del clado en cuestión e independizarnos de procesos selectivos. Los cambios neutrales se acumulan en los genomas de forma estocástica, proporcionalmente a la tasa de mutación espontánea. Por el contrario los cambios con efectos fuertes, positivos o deletéreos, son rápidamente fijados o eliminados de las poblaciones (Rocha, 2008). Por este motivo se analizaron las posiciones sinónimas, terceras posiciones de los codones degenerados.

Se trabajó con los alineamientos de genes no recombinantes de las especies más cercanamente emparentadas. Los SNPs se extrajeron del alineamiento y la identidad del estado ancestral o derivado se determinó de acuerdo al genoma usado como referencia. Al proceso de identificación del estado ancestral del sitio en cuestión le llamamos polarización.

Además se clasificaron los posibles cambios a observar en tres categorías que se muestran en la tabla 1.3.

Tabla 1.3: Polarización de las mutaciones de acuerdo a un genoma de referencia.Se ilustra el criterio a través del cual se determina cuál es el estado ancestral de un sitio. No se usaron sitios con más de dos bases distintas, ejemplificadas por x y 0. Se usa un genoma como referencia que se escoge a partir de las relaciones filogenéticas.

	1	2	3
Genoma de referencia:	x	x	0
Sp 1	0	0	0
Sp 2	x	0	x
Sp 3	x	x	x
Sp 4	x	x	x
polarización	x -> 0	x -> 0	x -> 0

Este procedimiento se llevó a cabo con un script de python que permite detectar los sitios del alineamiento que varían y de acuerdo al criterio expuesto en la Tabla 1.2, identificar la dirección en que sucedió la variación, o sea identificar el estado ancestral del sitio.

Como resultado se obtiene el número de veces que cada uno de los doce cambios pudo observarse y de la misma forma que se explicó para el estudio del patrón sustitucional, los datos se estandarizaron de acuerdo a la composición de bases para obtener la frecuencia de ocurrencia de la mutación. También se estimó el contenido

G+C en equilibrio dado por el patrón mutacional.

Reconstrucción de caracteres ancestrales

Una forma de estudiar el sesgo mutacional de un clado, es a través de los SNPs presentes entre los taxa que lo conforman. Para poder establecer la dirección del sesgo, es necesario contar con la cantidad y la polaridad de estos cambios. Obtener la cantidad de SNPs es un problema menor, pero polarizar los cambios, o sea, detectar el estado ancestral de un sitio es un problema más complejo.

Una filogenia describe el patrón jerárquico de descendencia de un grupo de especies a partir de un ancestro común. Si hay información disponible sobre los estados de los caracteres en los taxa contemporáneos, entonces surge la posibilidad de usar esa información combinada con la filogenia para reconstruir los acontecimientos históricos de la evolución (Blanchette et al., 2008).

La reconstrucción de los estados ancestrales de las secuencias, permite testar hipótesis sobre la evolución de genes y genomas (Arenas y Posada 2010).

Las secuencias ancestrales se pueden inferir a través de varios métodos, principalmente máxima parsimonia (MP), máxima verosimilitud y métodos bayesianos. Los últimos dos son herramientas estadísticas estándar, y los modelos probabilistas más usados actualmente tanto en reconstrucción filogenética como en ASR. Los estados ancestrales en este trabajo se obtuvieron utilizando el método de ML que ha demostrado ser un método preciso cuya principal desventaja es la demanda computacional del proceso (Holder and Lewis, 2003; Whelan et al., 2001).

En ML, una hipótesis es juzgada de acuerdo a cuán bien predice los datos observados, o sea, las secuencias actuales. El proceso de ML comienza con un alineamiento de secuencias, una filogenia que las relaciona y un modelo (estadístico) evolutivo:

- Alineamiento de secuencias actuales: para la ASR se necesita contar con un alineamiento de la mayor cantidad de secuencias ortólogas posibles, cuanto más información se posea más precisa será la reconstrucción (Blanchette et al., 2008).
- Árbol filogenético: un componente importante de la ASR por ML es el conocimiento de las relaciones filogenéticas entre las secuencias. Poseer la topología del árbol y estimar sus largos de rama es crucial para una reconstrucción precisa (Blanchette et al., 2008).
- Modelo (estadístico) evolutivo: estos modelos están compuestos por parámetros que rigen los aspectos del proceso evolutivo como las probabilidades relativas de transversiones versus transiciones y el grado en el cual la tasa de evolución varía entre sitios. Los valores correctos de estos parámetros se estiman durante el proceso y se escoge el valor de los mismos que maximiza la verosimilitud de los datos (Holder and Lewis, 2003).

Un análisis por verosimilitud requiere de la formulación de un modelo probabilista de sustitución, en este caso se utilizó un modelo de sustitución de codones que provee Codeml:

$$Q_{ij} = \begin{cases} 0, \text{ si los dos codones difieren en mas de una posición} \\ \pi_{ij}, \text{ transversiones sinónimas} \\ k\pi j, \text{ transiciones sinónimas} \\ w\pi j, \text{ transversiones no sinónimas} \\ wk\pi j, \text{ transiciones no sinónimas} \end{cases}$$

Siendo π_i la frecuencia en equilibrio del codón i, Pij(t) es la probabilidad de que el carácter i sea reemplazado por el carácter j a lo largo de una rama de largo t. En todos los modelos evolutivos probabilistas, las probabilidades se expresan en forma de sumas o multiplicaciones de factores Pij(t). La verosimilitud de los datos describe la probabilidad de observar los caracteres actuales dados la topología del árbol, los largos de rama y los factores Pij(t).

Se puede calcular la probabilidad de observar los datos x, como la suma sobre todas las posibilidades de y:

$$f(x;\theta)=\sum f(y)\, f(x \mid y;\, \theta)$$

Los nodos internos del árbol representan las secuencias ancestrales probables. Para cada nodo interno de la filogenia y para cada sitio en la secuencia, se calcula la verosimilitud del posible estado ancestral (que se define como la probabilidad de observar todos los estados del presente dados el carácter ancestral, la filogenia y el modelo).

$$\frac{f(y \mid x;\theta)=\sum f(y)\, f(x \mid y;\, \theta)}{f(x\,;\,\theta)}$$

El carácter que el procedimiento de ML indica como ancestral es el que posea la verosimilitud más alta (Hanson-Smith et al., 2010).

Siguiendo la teoría de ML, se escoge la configuración de ancestros, largos de rama y topología del árbol que maximice la verosimilitud de las secuencias observadas. Es importante destacar que cada carácter de las secuencias (sitio en el alineamiento) es tratado de forma independiente, o sea, la probabilidad del alineamiento entero es el producto de las probabilidades para cada posición (Blanchette et al., 2008).

La aproximación por máxima verosimilitud asume que el alineamiento, el árbol, el modelo y los parámetros del mismo, son correctos *a priori*. En la práctica esta asunción no es válida, para solucionar estos problemas, se han planteado métodos Bayesianos para la ASR, que pueden incorporar la incertidumbre, pero el hecho es que no se pudo demostrar que el uso de los mismos mejore la precisión en la ASR (Hanson-Smith et al., 2010).

El procedimiento de ASR se realizó con PAML (*Phylogenetic analysis by maximum likelihood*) (Yang, 2007) un paquete de programas, que permite el análisis filogenético de secuencias nucleotídicas y proteicas por máxima verosimilitud. El paquete incluye el programa Codeml, que implementa modelos de sustitución de nucleótidos, codones y aminoácidos para secuencias codificantes (Yang 2007).

En Codeml se introdujo como *input* el alineamiento de las secuencias y las relaciones filogenéticas entre ellas. El modelo evolutivo especificado fue HKY85 y la reconstrucción se realizó usando como unidad evolutiva el codón (modelo de sustitución de codones).

Resultados

Filogenia y genes recombinantes

La búsqueda realizada con el BLAST BRH, dio como resultado 399 grupos de genes ortólogos entre los 16 genomas con promedio de porcentaje de similitud de 80.4. Con este set de datos se realizó la reconstrucción filogenética del clado que se muestra en la Figura 1.3.

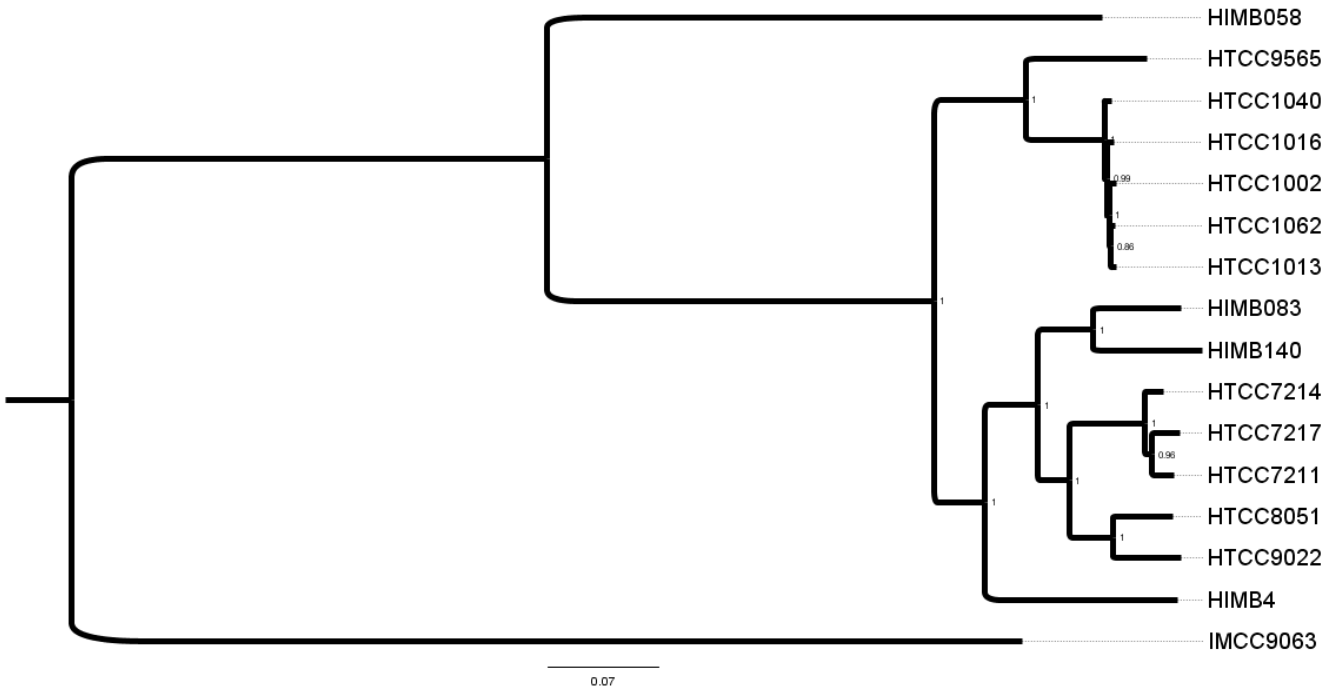


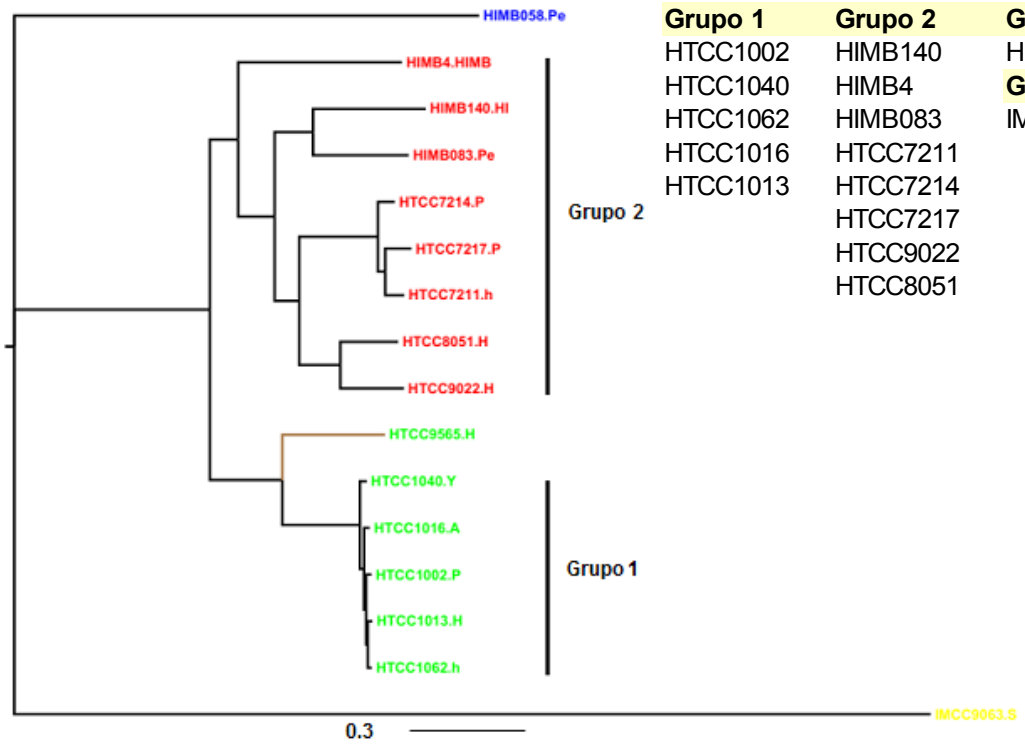
Figura 1.3: Reconstrucción filogenética del clado ralizada a partir del concatenado de genes ortólogos

El árbol filogenético (Fig. 1.3) se enraizó en la rama más larga, perteneciente al taxa IMCC9063, se pueden diferenciar dos grandes subclados dentro de los cuales se observan diferencias en cuanto a los niveles internos de divergencia. Uno de los subclados, está compuesto por las especies HTCC1013, HTCC1062, HTCC1016, HTCC1040, HTCC1002 y HTCC9565, y muestra niveles de divergencia muy bajos (por lo menos entre HTCC1013, HTCC1062, HTCC1016, HTCC1040 y HTCC1002), indicando que posiblemente estos linajes se han separado recientemente. Por otro lado se puede identificar un subclado compuesto por HIMB083, HIMB140,HTCC7214, HTCC7217, HTCC7211, HTCC8051, HTCC9022 y HIMB4, en este caso los niveles de divergencia son mayores y aunque dentro del clado hay varios grupos de especies hermanas la divergencia entre ellas es superior a la observada en el otro subclado.

Como se explicó en la metodología, el nivel de divergencia entre taxas es clave para realizar estudios mutacionales, a primera vista el clado compuesto por HTCC1013, HTCC1062, HTCC1016, HTCC1040, HTCC1002 y HTCC9565 es un buen candidato, pero esta elección no depende solamente lo evidenciado por la reconstrucción filogenética.

En bacterias, el concepto de especie es algo problemático ya que no se reproducen de la manera sexual

convencional, de todas formas, la genética de poblaciones existe en la medida en que existen cepas que experimentan juntas deriva génica y selección (Hildebrand et al., 2010). El estudio de estructura poblacional realizado con BAPS determinó que en la muestra existen cuatro grupos (Fig. 1.4).



Grupo 1	Grupo 2	Grupo 3
HTCC1002	HIMB140	HIMB058
HTCC1040	HIMB4	**Grupo 4**
HTCC1062	HIMB083	IMCC9063
HTCC1016	HTCC7211	
HTCC1013	HTCC7214	
	HTCC7217	
	HTCC9022	
	HTCC8051	

Figura 1.4: Se muestra como quedaron conformados los grupos encontrados con BAPS en la filogenia.

La especie HTCC9565 no se pudo adjudicar a ninguno de los 4 grupos, el estudio indicó que tiene similitudes con los cuatro. El resultado de BAPS reafirma los clados que mostró el análisis filogenético, e indica además que el intercambio genético entre los grupos, aunque exista, no es suficiente para homogeneizar las poblaciones, o sea, existen barreras para el flujo de material genético entre los clados. Llama la atención que aún dentro del grupo 2 que presenta una divergencia relativamente importante el análisis de BAPS sugiere un flujo de material genético importante.

Los grupos 3 y 4 quedaron excluidos de futuros análisis por estar conformados por un solo taxa, los análisis de sustitución nucleotídica se llevaron a cabo con los grupos 1 y 2 indicados en la figura 1.4.

La red filogenética construida con Splitstree (Figura 1.5) muestra un cierto grado de reticulación, indicando eventos de recombinación o transferencia horizontal entre las especies, que han afectado al set de genes ortólogos.

Tabla 1.4: Número de genes recombinantes encontrados por el test Phi

Genes	Grupo 1	Grupo 2
Recombinantes	68	95
No Recombinantes	331	304
Ortólogos	399	399

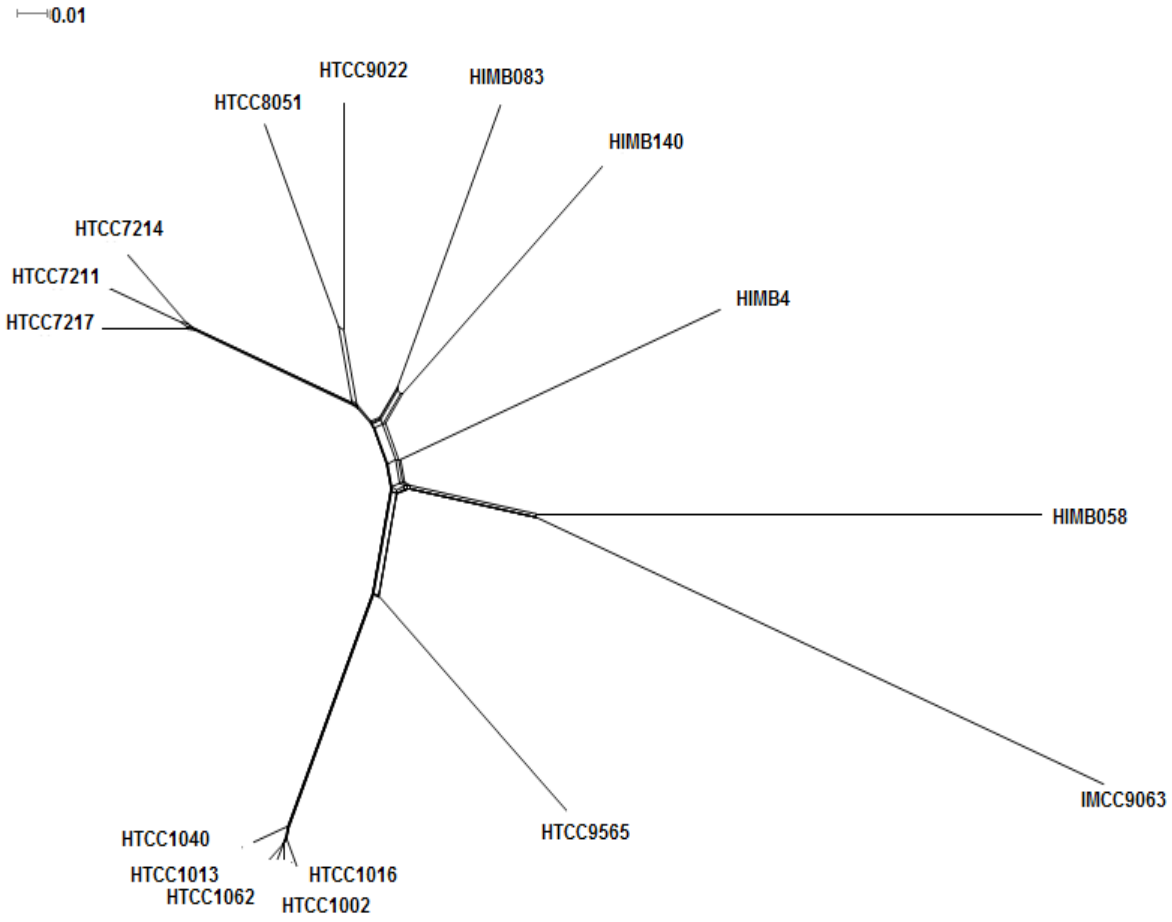


Figura 1.5: Red filogenética reconstruída a partir del set de 399 genes ortólogos con el programa Splitstree

La reticulación evidencia que no todas las secuencias ortólogas comparten una misma historia evolutiva (Huson and Bryant, 2006).

En concordancia con el resultado obtenido con Splitstree, el test Phi realizado *a posteriori,* permitió detectar 68 genes con indicios de recombinación en el grupo 1 y 95 en el grupo 2 (Tabla 1.4). El test Phi se realizó en cada grupo p or separado ya que la ASR también se realizó por grupo, y es con el fin de tornarla más precisa que se diferencian los genes que han recombinado de los que no. Los genes recombinantes se removieron en los estudios de ASR.

En este sentido, intentamos comparar la historia evolutiva de cada gen presuntamente ortólogo con la del árbol de especies obtenido utilizando el test de comparación de filogenias de Shimodiara-Hasegawa (Shimodaira and Hasegawa, 2001).

De los 399 genes, hay 79 en los que la filogenia de la especie no difiere significativamente de la del gen, sugiriendo que estos habrían sido transferidos en forma vertical. Este subconjunto de genes es el que cumple con los criterios más exigentes para los posteriores análisis. Los resultados generales no difieren del grupo de 399 genes así que los resultados que se presentan en la tesis son los del grupo mayor.

Estimaciones de dS, dN y w

Se estimó la tasa de variación sinónima (dS) y la de variación no sinónima (dN). Estas estimaciones nos permiten obtener una visión global de las presiones selectivas operantes sobre el set de genes ortólogos. Los valores estimados para cada gen se pueden observar en la Figura 1.6. Es importante destacar que las estimaciones del clado dos muestran valores de dS muy altos (>1) lo que representa sitios saturados de sustituciones en la mayoría de genes ortólogos (Luo et al., 2014).

El clado 1 no muestra saturación a nivel de sitios sinónimos (dS<1), pero si valores de dN/dS bajos (<0.1), indicando que existe una fuerte presión evolutiva para minimizar las sustituciones aminoacídicas (selección negativa o purificadora). Dadas las características de la familia bacteriana con la que se está trabajando (gran tamaño poblacional, vida libre, *genome streamlining*), los valores obtenidos en las estimaciones de dN y dS son esperables. Cuando una especie posee gran tamaño poblacional efectivo, la selección aunque débil, puede tener un gran efecto en la evolución de su genoma (Charlesworth, 2009).

Los patrones de sesgo mutacional obtenidos en estudios anteriores (Hershberg and Petrov, 2010a; Hildebrand et al., 2010), provienen de cepas clonales que no mostraban señales de selección, presentaban valores de w de aproximadamente 0.6, asegurando un estado relajado de las presiones selectivas, importante para el estudio a nivel mutacional (Rocha, 2008). Se esperaba encontrar esta situación en los genes no recombinantes del clado 1, pero los resultados muestran un dN/dS de 0.02 (estimado sobre el concatenado de los genes ortólogos no recombinantes), es un valor muy bajo pero, como se dijo anteriormente, hay que tener en cuenta que el gran tamaño poblacional efectivo (que se estima tienen estas especies) influye disminuyendo el dN/dS (Luo et al., 2014).

De todas formas, el estudio de las sustituciones en las terceras posiciones de codones degenerados (*4fold*) a nivel de genes no recombinantes del clado 1 podrían acercarnos al patrón mutacional: al no verse afectados por recombinación, la diversidad nucleotídica en estos genes sería resultado solamente de la mutación (Arenas and Posada, 2010b), y además la presión selectiva sobre las posiciones *4fold*, se encontraría relajada dado que los cambios en estos sitios no generan grandes alteraciones a nivel de la traducción. Otros indicios a favor del estudio del patrón sustitucional a nivel del grupo 1 es el valor de dS encontrado, de 0.22, que indica que las posiciones no se encuentran saturadas de cambios (Rocha et al., 2006; Luo et al., 2014).

Tabla 1.5: dS estimado entre pares de especies a través del método Yang&Nielsen

Sub 1		Sub 2					
Pares	dS	Pares	dS	Pares	dS	Pares	dS
HTCC1016_1002	0.0638	HTCC7211_HIMB083	1.0215	HIMB4_HTCC7211	1.0911	HTCC9022_7217	0.9191
HTCC1040_1002	0.0859	HTCC7214_HIMB083	1.0299	HTCC8051_7211	0.8955	HIMB4_HIMB140	1.1019
HTCC1062_1002	0.0549	HTCC7217_HIMB083	1.0339	HTCC9022_7211	0.9239	HTCC8051_HIMB140	1.0662
HTCC1013_1002	0.0586	HIMB140_HIMB083	1.0045	HTCC7217_7214	0.2562	HTCC9022_HIMB140	1.093
HTCC1040_1016	0.0856	HIMB4_HIMB083	1.1141	HIMB140_HTCC7214	1.0629	HTCC8051_HIMB4	1.0998
HTCC1062_1016	0.0624	HTCC8051_HIMB083	1.0084	HIMB4_HTCC7214	1.0952	HTCC9022_HIMB4	1.1086
HTCC1013_1016	0.065	HTCC9022_HIMB083	1.0371	HTCC8051_7214	0.9023	HTCC9022_8051	0.5976
HTCC1062_1040	0.0846	HTCC7214_7211	0.226	HTCC9022_7214	0.9214	HTCC8051_7217	0.9001

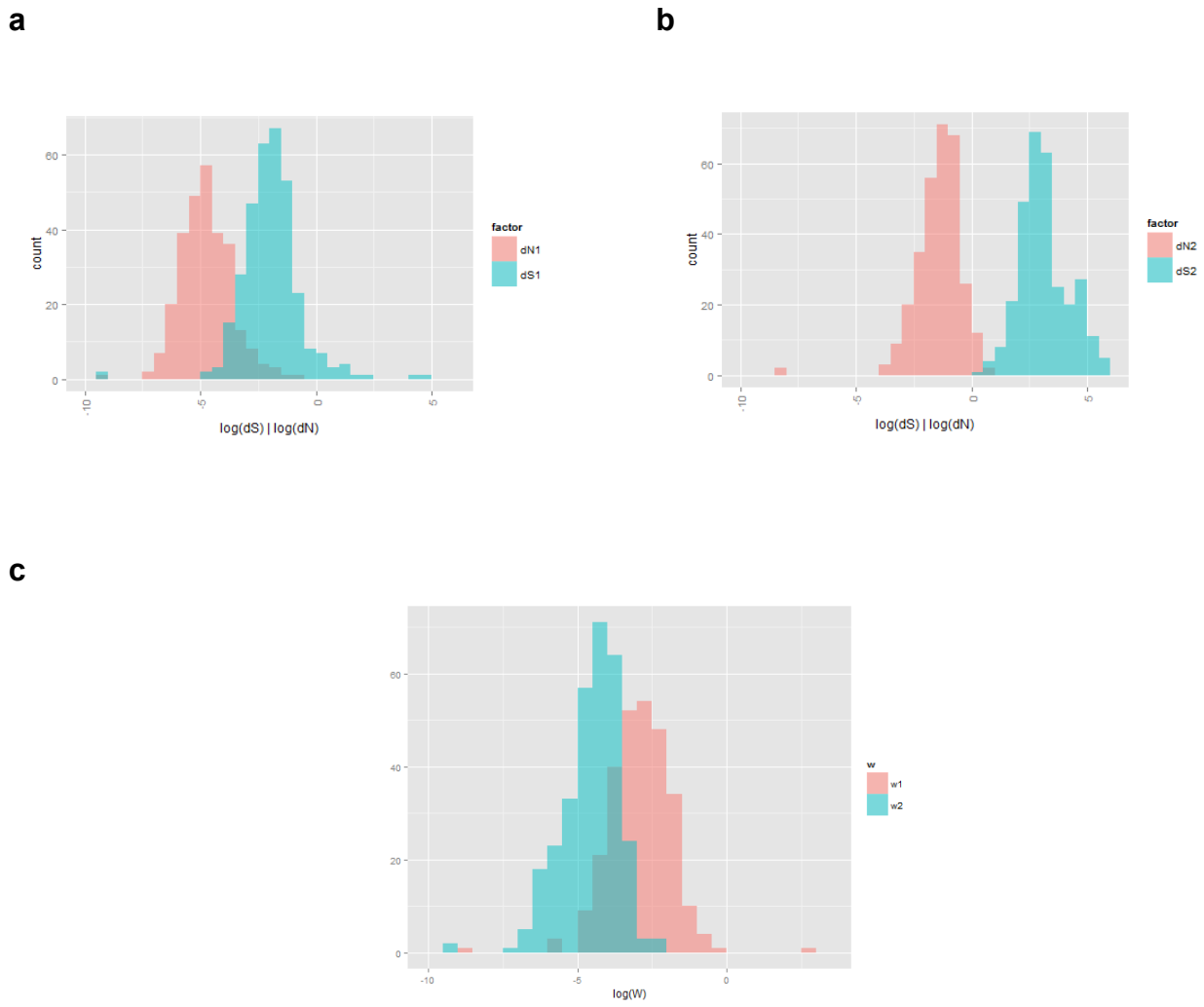


Figura 1.6: Distribución de los valores de dS y dN (escala logarítmica) en el set de genes ortólogos para el grupo 1 (a) y para el grupo 2 (b). En c se muestra la distribución de w (escala logarítmica) para el mismo set de genes de ambos grupos.

Patrones sustitucionales

Los patrones de sustitución que se presentan son el resultado de la polarización y posterior conteo de los SNPs estimados a través de la reconstrucción de caracteres ancestrales con el programa Codeml. Los cambios se clasificaron en seis tipos de sustituciones: G/C a A/T, G/C a T/A, G/C a C/G, A/T a G/C, A/T a C/G, y A/T a T/A La frecuencia relativa de cada categoría se obtuvo normalizando por el contenido en G+C de las posiciones sinónimas y no sinónimas. Las tasas de sustituciones sinónimas y no sinónimas para cada categoría de cambio se muestran en la figura 1.7.

Las frecuencias estimadas indican que los cambios que implican la sustitución de G por A y C por T, son los más observados sin importar de qué sitios se trate, sinónimos, o no sinónimos, esto es consistente con los trabajos abocados al estudio de patrones mutacionales mencionados en los antecedentes (Hershberg and Petrov, 2010a; Hildebrand et al., 2010).

Se pueden observar diferencias entre el patrón sustitucional del clado 1 y el del clado 2, en el uno, la incidencia de los cambios GC->AT es destacadamente mayor que el resto, mientras que en el clado 2, otros tipos de

cambios cobran protagonismo, este es el caso de las sustituciones AT->TA y AG->TC. Las frecuencias de sustituciones varían también debido a la naturaleza de los sitos, los sitios sinónimos que son más permisivos, presentan mayor frecuencia de sustituciones que los no sinónimos. El nivel de variación se incrementa con la divergencia entre las especies tanto a nivel de sitios sinónimos (S) como no sinónimos (NS), por lo tanto, son mayores las frecuencias de cambios nucleotídicos en el grupo dos, debido a la acumulación de los mismos.

El contenido nucleotídico es un carácter que evoluciona lentamente. Por lo tanto, ya sea impulsado por la selección natural como por g-BGC, la composición nucleotídica entre especies cercanas podría no haber tenido tiempo suficiente para alcanzar un nuevo equilibrio mutacional. Sin embargo, si el contenido en G+C de un genoma se debe predominantemente a la mutación, y ni la selección ni g-BGC tienen un rol importante en la evolución del mismo, entonces, el contenido nucleotídico de los genomas ya habrá alcanzado un equilibrio mutacional. Para testar esto se compara el número de cambios que son GC->AT y AT->GC, en equilibrio estos deberían ser iguales (Hershberg and Petrov, 2010a). A su vez, se estimó el GC en equilibrio (GC_{eq}) resultante del patrón sustitucional sinónimo y no sinónimo en ambos grupos (Tabla 1.6). Este índice representa el contenido en G+C que alcanzarían los genomas si estuvieran evolucionando solo por mutación.

Tabla 1.6: G+C en equilibrio

Grupo 1	
sinónimos	**No-sinónimos**
0.17	**0.26**
Grupo 2	
sinónimos	**No-sinónimos**
0.34	**0.36**

La frecuencia de cambios que son GC->AT y AT->GC son diferentes, indicando que las especies no están en equilibrio mutacional. El GC_{eq} estimado a partir de las sustituciones sinónimas en el grupo 1, coincide con el contenido GC3 de los genomas (Tabla 1.7) en la mayoría de los casos, pero el estimado a partir de los cambios no sinónimos (0.26) es menor al contenido G+C tanto de la primera como de la segunda posición. Estos resultados están indicando que en los sitios no sinónimos puede existir una elevada tasa de fijación de cambios hacia G+C.

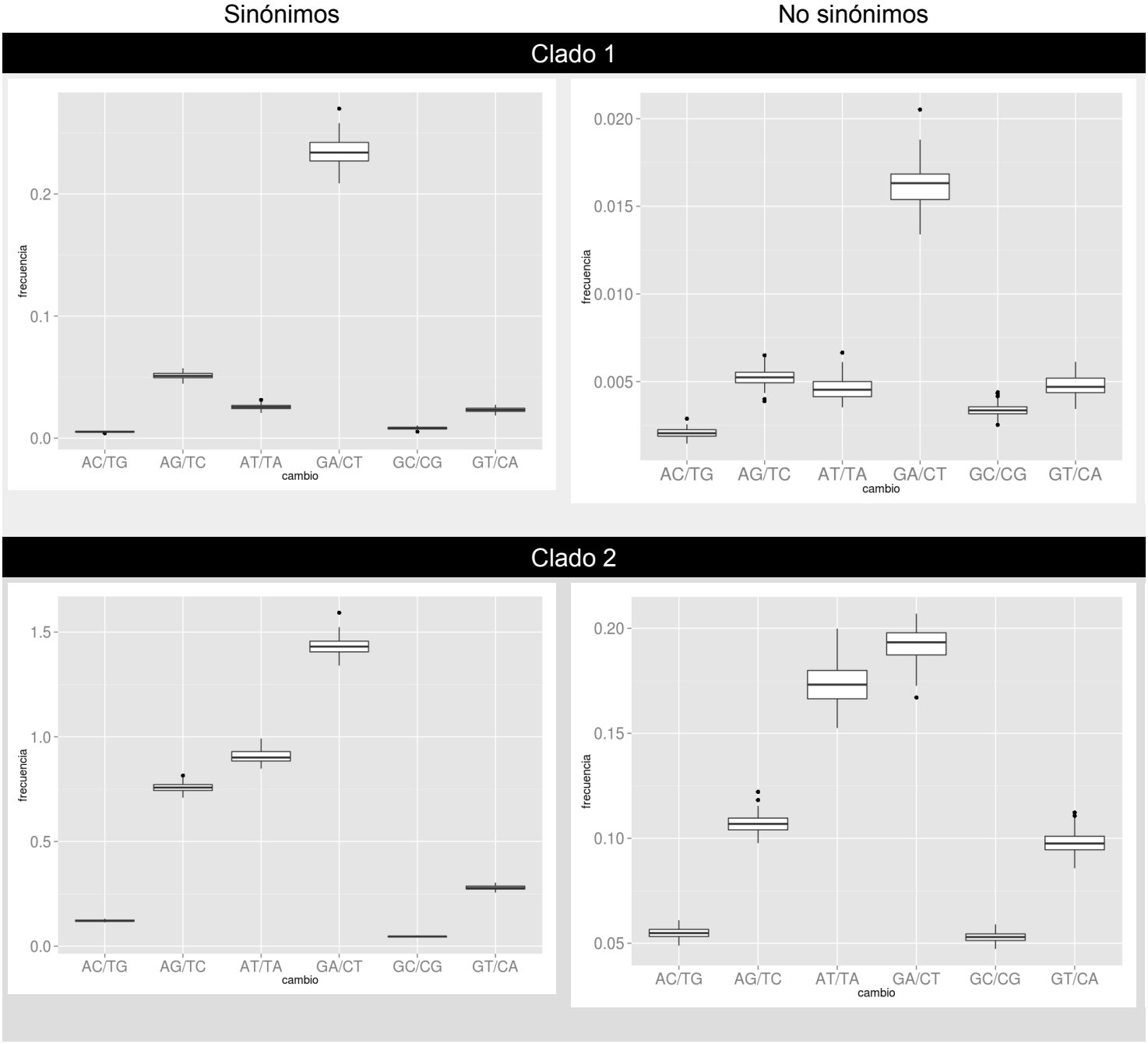


Figura 1.7 : Se muestran las frecuencias de los seis tipos de cambio y el GC en equilibrio estimado a partir de ellos, para el clado uno y el dos de acuerdo a los sitios estudiados, sinónimos y no sinónimos

Las diferencias entre genomas lejanamente emparentados deberían reflejar los efectos de la selección y/o g-BGC, el GC_{eq} calculado a partir de estas diferencias debería ser más similar al contenido G+C observado para las especies que el estimado a partir de SNPs de cepas cercanamente emparentadas (Hershberg and Petrov, 2010a). Para el clado dos ambas estimaciones arrojaron valores cercanos a 0.3, para el caso de los sitios sinónimos este es un número excesivo, pero para los no sinónimos es concordante con el G+C observado.

Tabla 1.7:: Contenido GC para cada Posición de los codones

	GC1	GC2	GC3
HIMB058	0.38	0.31	0.19
HTCC9565	0.38	0.31	0.17
HTCC1040	0.39	0.31	0.18
HTCC1016	0.39	0.31	0.18
HTCC1002	0.39	0.31	0.18
HTCC1062	0.39	0.31	0.18
HTCC1013	0.39	0.31	0.18
HIMB083	0.39	0.31	0.17
HIMB140	0.39	0.31	0.17
HTCC7214	0.39	0.31	0.17
HTCC7217	0.39	0.31	0.17
HTCC7211	0.39	0.31	0.17
HTCC8051	0.39	0.31	0.17
HTCC9022	0.39	0.31	0.17
HIMB4	0.39	0.30	0.17
IMCC9063	0.40	0.32	0.23

Las causas de las variaciones en este sentido son difíciles de determinar. Las diferencias entre genomas más distantemente relacionados, reflejadas en este estudio por el grupo 2, deberían reflejar los efectos de la selección natural y/o BGC. El contenido GC en equilibrio estimado a partir de estas diferencias debería ser similar al contenido G+C observado (Hershberg and Petrov, 2010a).

Para observar el cambio en el patrón mutacional a medida que aumentan los niveles de divergencia (Fig 1.8), se tomaron las sustituciones, sinónimas y no sinónimas detectadas en diferentes largos de rama o “caminos”. Nuevamente, tomando en cuentas distancias evolutivas pequeñas, la frecuencia de cambios GA->CT es la mas destacada. Esta tendencia aparentemente pierde intensidad con el tiempo, los GC_{eq} estimados a diferentes niveles de divergencia, van incrementando su valor a medida que la distancia evolutiva aumenta, este incremento parece deberse principalmente a la disminución en la frecuencia de las sustituciones GA-CT (Fig1.8-b). A medida que pasa el tiempo los cambios de AT-TA incrementan su frecuencia relativa, pero esto no debería afectar la estimación de GC_{eq}, ya que no se usa este valor para su cálculo.

El cambio en el patrón sustitucional en distancias profundas puede estar indicando la dirección de la selección, en este caso una disminución en la frecuencia relativa de las sustituciones GA->CT, debido probablemente al purgado de las mismas.

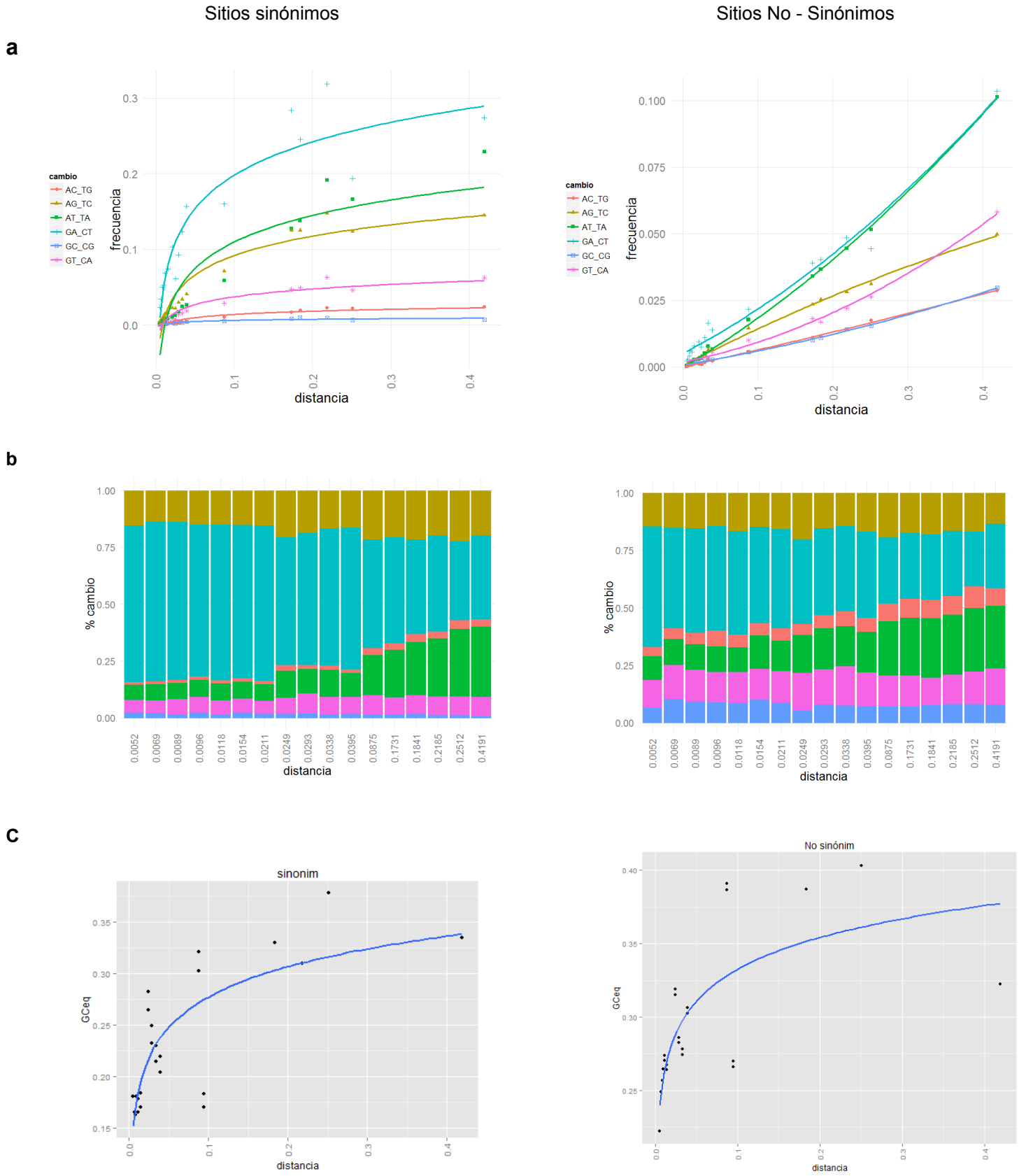


Figura 1.8: a- Frecuencia estimada para los distintos tipos de cambios en diferentes niveles de divergencia. b- Porcentaje de contribución de cada tipo de cambio en cada nivel de distancia. c- Contenido GC en equilibrio estimado a partir de los patrones sustitucionales característico de cada nivel de divergencia

A medida que aumenta el nivel de divergencia los tipos de cambios que prevalecen varían, la figura 1.8-b, muestra cómo cambia con el tiempo la contribución de cada tipo de sustitución sobre total de diferencias estimadas, se ve claramente que el aporte de las sustituciones GA->CT disminuye y es reemplazado por los cambios AT->TA.

Aproximación al Patrón Mutacional (cambios 4fold)

Los cambios *4fold* se estimaron polarizando y contabilizando las diferencias en las terceras posiciones de los codones degenerados. Con la intención de aproximar al parón mutacional, se tomaron en cuenta solo las sustituciones en sitios 4fold entre las especies con la mínima distancia evolutiva del clado, el set de alineamientos usado en este caso se reduce a los genes no recombinantes del clado 1. Se obtuvo la frecuencia para los doce tipos de cambios (Fig 1.9). Las transiciones de G a A y de C a T son las variaciones mas frecuentemente observadas en todos los casos. Como se había visto en cierta medida en los análisis anteriores, el sesgo mutacional en Pelagibacteriaceae es, al igual que en todos los clados bacterianos estudiados hasta ahora, de G+C hacia A+T.

Se compararon los patrones sustitucionales encontrados en los diferentes sets de datos analizados, entre el grupo uno y el dos, separando entre sitios sinónimos y no sinónimos, a nivel de genes recombinantes y no recombinantes, y se encontraron importantes correlaciones entre ellas (Figura 1.10).

El contenido GC_{eq} estimado a partir de los sitios *4fold* (Tabla 1.8) muestra coincidencia con el observado en las terceras posiciones de los genomas (~0.17), indicando que la composición nucleotídica en estos sitios está dominada por la mutación.

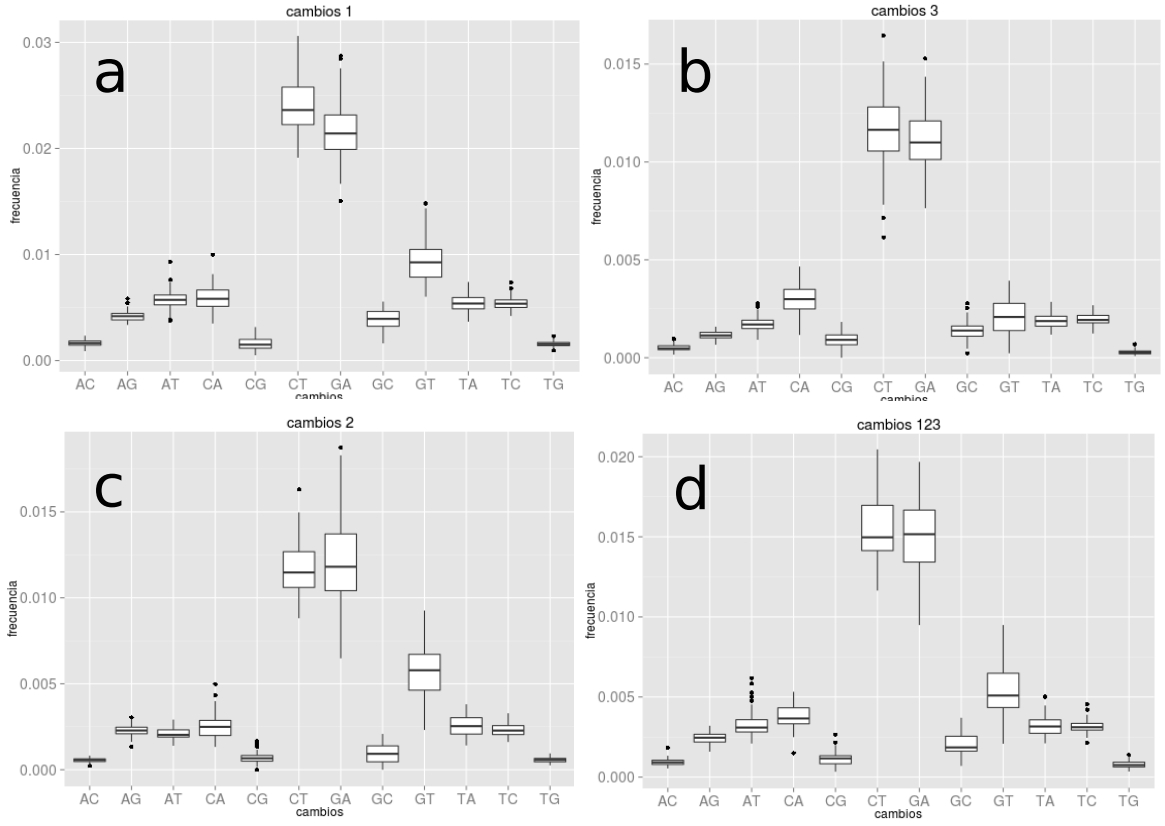


Figura 1.9: Frecuencia de cambios 4fold para los tres tipos de mutaciones estudiadas (b-1,c-2,c-3) y la suma de ellas (a).

Tabla 1.8: Contenido GC en equilibrio estimado a partir de cambios 4fold

4fold	Gc_{eq}
Categoría 1	**0.17**
Categoría 2	**0.15**
Categoría 3	**0.12**

Discusión

Las transiciones de G a A y sus complementarias de C a T, son las variaciones mas frecuentemente observadas en todos los casos cuando normalizamos por el contenido nucleotídico.

Nuestros resultados muestran que el sesgo mutacional en Pelagibacteriaceae es, al igual que en todos los clados bacterianos estudiados hasta ahora, de G+C hacia A+T. Se puede observar en general un exceso de cambios de C a T y de G a A para todas las posiciones estudiadas (sinónimas, no sinónimas y *4fold*). Dependiendo del nivel de divergencia al cual se esté llevando a cabo el análisis, se ve modificada la abundancia relativa de otros tipos de cambios, como AT->TA y GT->CA, que cobran mayor importancia a medida que se van acumulando diferencias y por lo tanto distancia evolutiva.

Lo que evidencian los resultados es que el sesgo mutacional tiende a enriquecer los genomas de Pelagibacteriaceae en los nucleótidos A y T, generando un exceso de sustituciones en ese sentido (GC->AT > AT->GC).

El análisis realizado por Hildebrand et al, sobre el sesgo mutacional en especies ricas en A+T no fue concluyente (Hildebrand et al., 2010) pero Hershberg et al, trabajó con dos especies de parásitos intracelulares ricos en A+T, *S.typhi* y *Y.pestis*, y encontró que el sesgo mutacional se dirige hacia A+T y que el GC_{eq} se corresponde con el contenido G+C genómico observado (Hershberg and Petrov, 2010a).

En Pelagibacteriaceae, el sesgo mutacional parece dominar el contenido de las terceras posiciones de los codones (GC_{eq} en posiciones sinónimas ~ 0.17, GC3 ~ 0.17), pero no de todo el genoma, como en los endosimbiontes (Andersson and Kurland, 1998; Andersson and Sharp, 1996; Hershberg and Petrov, 2010a). El GC_{eq} estimado a partir de las sustituciones no sinónimas en el clado 1 revela que, si estuvieran guiadas solamente por la mutación, las posiciones no sinónimas (1° y 2°) tendrían un G+C aproximado de 0.26. Sin embargo, los valores observados en estos sitios son de ~0.39 y ~0.31, esto indica que el sesgo mutacional opera sobre todo el genoma pero su impacto se ve limitado a nivel no sinónimo por otros mecanismos evolutivos (Hershberg and Petrov, 2010a). O sea, eventualmente el exceso en A+T generado por el sesgo se pierde. A través de qué mecanismos, no se puede responder exactamente pero como se explicó en la introducción, el purgado del exceso en A+T tiene principalmente, dos explicaciones.

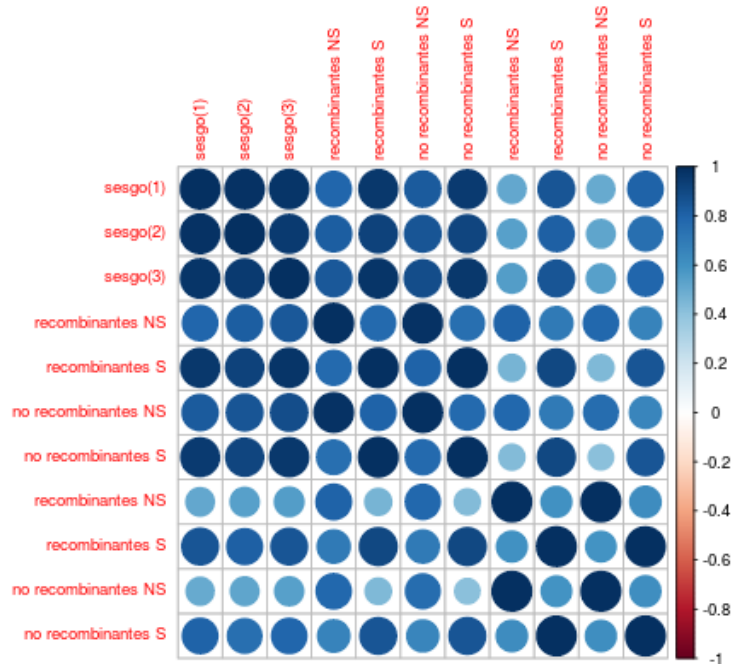


Figura 1.10: Se ven representadas las correlaciones entre los diferentes patrones de sustitución encontrados. Los coeficientes de correlación más altos corresponden a los círculos más grandes. Sesgo (1), (2) y (3) corresponden a los cambios 4fold. Luego están las estimaciones para el clado 1 y para el 2 en ese orden. NS=sitios no sinónimos. S=sitios sinónimos

Por un lado, dado que encontramos un número importante de genes recombinantes y que existen estudios que muestran la importancia de la recombinación en la evolución del clado (Vergin et al., 2007), el mecanismo de g-BGC, podría estar favoreciendo la fijación de los nucleótidos G+C en sitios polimórficos. Un indicio de que g-BGC está afectando la composición nucleotídica de los genomas, es la existencia de diferencias significativas entre el contenido G+C de los genes recombinantes y los no recombinantes. Si está surtiendo efecto el proceso de g-BGC, los genes recombinantes deberían de poseer mayor cantidad de pares de bases G+C. Este no es el caso en Pelagibacteriaceae, la composición de bases entre ambos tipos de genes se comparó (no se muestran los resultados) y no se hallaron diferencias significativas entre ellos. Esto no descarta por completo la participación de g-BCG en la evolución de la composición nucleotídica del clado pero es un dato desalentador.

Por otro lado, la probabilidad de fijación de las mutaciones hacia A+T puede estar siendo modificada por presiones selectivas actuando sobre los sitios no sinónimos. Los resultados mostraron indicios de selección purificadora actuando en los linajes, o sea hay señales de que las mutaciones que ocurren a nivel de sitios no sinónimos, que se vio son más frecuentes en sentido GC->AT, son eliminadas. En bacterias ricas en G+C se observó también sesgo mutacional hacia A+T y se postuló que la selección estaría actuando sobre todas las posiciones de los codones de manera de mantener el GC3 extremo de estos organismos (Hershberg and Petrov, 2010a). En Pelagibacteriaceae este no parece ser el caso, a menos que el enriquecimiento en A+T se vea favorecido cosa que no podríamos distinguir de el proceso mutacional.

Cuando se comparan genomas bacterianos cercanamente emparentados, donde el concepto de especie no es

tan claro como en otros organismos, no es válido asumir que los efectos de la selección sobre los cambios no sinónimos sean instantáneos. Se debe considerar la posibilidad de que la tasa de dN/dS cambia a lo largo del tiempo debido a un defasaje en la remoción de mutaciones levemente deletéreas (Ohta, 2000; Rocha et al., 2006). Esto puede explicar el desbalance encontrado a nivel de sitios no sinónimos en el clado 1 (GC_{eq}=0.26 vs GC1~0.39 y GC2~0.31) y que no se observa en el clado 2 (GC_{eq}=0.33 vs GC1~0.39 y GC2~0.31). Es posible que la vasta mayoría de las mutaciones en los sitios no sinónimos tengan un costo selectivo lo suficientemente pequeño como para mantenerse en la población por un período de tiempo, antes de ser "purificadas" por selección natural (Rocha et al., 2006). A su vez, la observación de un exceso de cambios hacia A+T en los sitios sinónimos del clado uno en comparación con los del clado dos tiene sentido, ya que las sustituciones levemente deletéreas se limitan a niveles de divergencia bajos debido a su corto "tiempo de vida" (McDonald, 2006).

Se ha propuesto que la capacidad de las poblaciones de bacterioplancton de eliminar mutaciones deletéreas y esparcir las ventajosas durante la adaptación a un ambiente fluctuante está directamente relacionada con su tamaño poblacional efectivo (Luo et al., 2014), los resultados aquí presentados apoyan esta idea porque están indicando que la composición de bases de los genomas de Pelagibacteriaceae no es producto solamente de la mutación como se ha visto en genomas emparentados, sino que hay indicios de selección sobre este factor y las características que distinguen al clado son su tamaño poblacional y estilo de vida libre.

Capítulo II: Uso de codones en Pelagibacteriaceae

Introducción

El descubrimiento del código genético, develó que los aminoácidos que componen las proteínas, excepto Met y Trp, pueden ser codificados por más de un codón, entre dos y seis mas precisamente. Estos codones difieren típicamente en la tercera posición y se denominan codones sinónimos (Fig. 2.1). A partir de la disponibilidad de genomas completamente secuenciados de muchos organismos procariotas, se hizo evidente que el uso de los codones sinónimos no es azaroso, su frecuencia de ocurrencia difiere entre organismos y entre los genes de un organismo (Grantham et al., 1980; Gouy and Gautier, 1982), fenómeno conocido como sesgo en el *uso de codones sinónimos* (UCS).

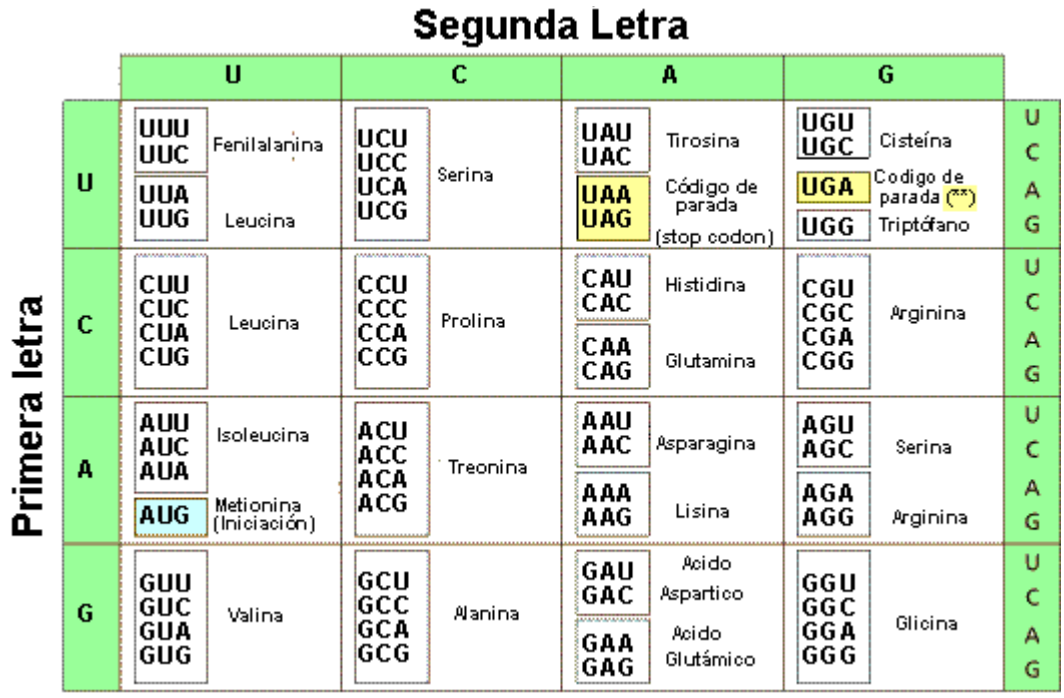


Figura 2.1: Código Genético

El sesgo en el UCS es similar entre especies que comparten características como contenido G+C, población de ARNt y que están emparentadas filogenéticamente (Chen et al., 2004). A su vez, estudios recientes sobre diferentes comunidades bacterianas indican que bajo las mismas condiciones ambientales y presiones evolutivas, y con la posibilidad de transferir material genético, las diferentes especies que componen una comunidad comparten el patrón de UCS (Roller et al., 2013).

A nivel intragenómico, se ha visto que el sesgo en el uso de codones sinónimos depende de varios fenómenos: estructura del ARNm, largo del gen, composición aminoacídica y nivel de expresión (Chen et al., 2004). En efecto, el sesgo en el UCS se usa de manera rutinaria para predecir genes altamente expresados y

funcionalmente relevantes (Plotkin and Kudla, 2011).

Las diferencias entre genes de un genoma, en el patrón de UCS, también son indicios de THG. Un gen adquirido recientemente, posee el UC característico del genoma de origen, que eventualmente se asemejará más al nuevo contexto genómico, a través del proceso conocido como amelioración. A veces es posible estimar la tasa de dicho proceso y predecir la edad del evento de transferencia (Ermolaeva, 2001).

Los motivos subyacentes a la variación en el UCS, han sido explicados tanto en contexto neutralista como seleccionista.

La explicación neutral se basa principalmente en la existencia de un sesgo mutacional característico del genoma e íntimamente asociado al patrón de sustitución de cada especie. El patrón de mutaciones determinaría la composición de bases y ésta, la variación en el UCS (Muto and Osawa, 1987). De hecho, a través de análisis multivariados se ha visto que el contenido G+C es la variable que permite explicar la mayor parte de la variación en el UCS (Chen et al., 2004; Sharp et al., 2010). El contenido en G+C es un factor importante principalmente en genomas con composición de bases extrema, las especies con alto contenido en G+C usan con mayor frecuencia codones terminados en G y C, las de extremo contenido en A+T, utilizan codones ricos en A y T (U). En muchos casos, el sesgo en el UCS no se explica solamente por el patrón de sustitución de una especie, por ejemplo las diferencias en el UCS encontradas entre los HEG y el resto de los genes del genoma, se ha interpretado como reflejo de la selección natural (Sharp et al., 2010).

Es generalmente aceptado que existe presión selectiva afectando el UCS en bacteria, insectos y mamíferos (Hershberg and Petrov, 2012). Para organismos unicelulares como las bacterias, existe una relación directa entre las tasas de los procesos celulares como la traducción, y las tasas de crecimiento y división celular. Por lo tanto, los cambios que impliquen mejoras en la traducción, aumentarán el *fitness* del organismo (Ran and Higgs, 2012). Dado que la traducción es el proceso energéticamente más costoso en las células que se encuentran en fase de crecimiento exponencial, su eficiencia se encuentra bajo fuertes presiones selectivas. Bajo estas condiciones fisiológicas, un pequeño set de genes captura la mayoría de la transcripción y traducción ocurriendo en la célula. Generalmente este set de genes incluye genes relacionados a la transcripción, traducción y metabolismo energético (Andersson and Kurland, 1998; Rocha, 2004).

Las diferencias selectivas entre codones sinónimos reflejan dos aspectos del *pool* de ARNt isoaceptores presente en las células (Ikemura, 1985). Los ARNt isoaceptores son ARN de transferencia cargados con el mismo aminoácido pero que responden usualmente a diferentes codones que codifican para ese aminoácido. La visión seleccionista tradicional sostiene que el uso de codones cuyo ARNt isoaceptor es más abundante, aumenta la eficiencia de la traducción. Los ARNt isoaceptores abundantes permitirían una traducción más veloz, dejando a los ribosomas, elemento limitante en este caso, disponibles rápidamente, lo que permite una mayor tasa de crecimiento para la célula. Existe evidencia experimental de que los ARNm compuestos por codones preferidos, se traducen de manera más rápida que ARNm modificados artificialmente, compuestos por codones raros (Ermolaeva, 2001).

La abundancia relativa de las especies de ARNt varía entre genomas y probablemente entre estadíos de crecimiento. La preferencia de codones se correlaciona con la composición del *pool* de ARNt (Ikemura, 1985), pero no es claro si esto es causa, o consecuencia del sesgo en la abundancia de ARNt. El número típico de de genes de ARNt en los genomas bacterianos es de alrededor de 50, pero varía desde 30 a 120. Se ha propuesto

que el número de copias de genes de ARNt evoluciona en respuesta a la selección traduccional. En organismos que se multiplican rápidamente, el tiempo invertido en la traducción es un factor limitante en la división celular; por lo tanto es beneficioso duplicar el número de genes ARNt, aumentando su concentración en la célula y la velocidad de la traducción. En organismos con bajas tasas de división celular, el tiempo empleado en la traducción no es un factor limitante, entonces el costo total de la traducción se minimiza reduciendo el número de copias de los genes ARNt. Por consiguiente el uso de codones y la población de ARNt coevolucionan a un estado mutuamente adaptado (Higgs and Ran, 2008; Rocha, 2004). Además de la velocidad a la que se traduce una proteína, la precisión con la que se hace es un factor muy relevante, hay datos que muestran que aproximadamente, una de cada 10^3-10^4 lecturas ribosomales resultan con errores (Ogle and Ramakrishnan, 2005). Hay una perspectiva seleccionista relacionada a esto que plantea que el uso de codones cuyo isoaceptor tiene complementariedad codón-anticodón perfecta, aumenta la precisión en la traducción. Un solo ARNt puede usualmente responder a muchos codones como fue propuesto por Crick en 1966 a través de la "*wobble hypothesis*", pero los codones con los que forman interacciones débiles serían evitados (Rocha, 2004). La razón de un elevado sesgo en el UCS en HEG deriva de su alto nivel de conservación que sugiere la existencia de fuertes presiones selectivas sobre la precisión en la traducción (Rocha, 2004). Los sitios donde la identidad del aminoácido es fundamental para la proteína es donde se necesita mayor precisión y se espera que sean los sitios más conservados entre especies (Sharp et al., 2005).

Ambas hipótesis selectivas han sido muy estudiadas, se ha detectado selección natural actuando sobre sitios sinónimos, para hacer la traducción de proteínas más eficiente (más rápida) y/o precisa, lo que sugiere que el proceso de adaptación puede ocurrir a través de cambios sinónimos que afectan la naturaleza de la interacción mRNA/tRNA (Botzman and Margalit, 2011). Estas explicaciones son las que en la actualidad ofrecen mayor evidencia, y no son mutuamente excluyentes (Hershberg and Petrov, 2012).

Es esperable que ambos tipos de selección afecten mas fuertemente a los genes de alta expresión, porque si estos no son traducidos de forma eficiente, se daría una disminución en la tasa de traducción que afectaría eventualmente la traducción de todos los genes del genoma. A su vez, si la traducción de los HEG no es precisa, una gran proporción de las proteínas podrían presentar errores llevando al mal funcionamiento y/o plegamiento de las mismas. Como resultado se podría observar la pérdida de funciones en parte del proteoma e incluso muerte celular (Ran and Higgs, 2012).

Dentro de un genoma, los codones que son favorecidos por la fuerza selectiva operante, como los usados para optimizar la traducción se conocen como codones óptimos. Mutaciones hacia un codón óptimo desde un codón no óptimo serán beneficiosas mientras que cambios en la dirección opuesta serán deletéreos. Por lo tanto, la probabilidad de que una población aumente la frecuencia de esta mutación dependerá del tamaño poblacional efectivo (Ne), del coeficiente de selección (s) y del sesgo mutacional (Iriarte et al., 2013).

En bacterias, en genes altamente expresados, los cuartetos y los duetos poseen preferencias opuestas para C y U en la tercera posición. Una explicación para este fenómeno es la estabilización del nivel de energía para el apareamiento codón-anticodón a través de la elección en la tercera base del codón, conocido como modelo de estabilidad (Crick, 1966). Las primeras dos posiciones de los codones que forman cuartetos son ricas en G y C , y se observa que existe la tendencia a utilizar U en la tercera posición, lo que contribuiría a estabilizar el nivel de energía del apareamiento codón-anticodón. En el caso de los duetos, el código genético muestra que las dos

primeras posiciones de los codones son pobres en G y C, observándose una preferencia al uso de C como tercera base (Sharp et al., 2005). Los codones preferidos para cada aminoácido pueden variar entre organismos de acuerdo al repertorio de ARNt y otros factores (Botzman and Margalit, 2011).

El modelo explicativo de la variación en el UCS a nivel intra-genómico que parece tener mayor aceptación es conocido como el modelo de balance mutación-selección-deriva (Bulmer, 1991). La selección sobre los codones sinónimos produce un sesgo en el UC entre los genes de un genoma, aumentando la frecuencia de algunos codones sobre sus sinónimos. Aunque la selección operando a nivel de uso de codones no es la única fuerza selectiva que afecta la identidad nucleotídica de un genoma, es la principal en muchas bacterias, permitiendo la existencia de codones no óptimos como resultado de la mutación y la deriva genética (Bulmer, 1991). El sesgo aumenta con el nivel de expresión del gen, indicando mayor fuerza selectiva, por ejemplo en aquellos genes que codifican para funciones fisiológicas relevantes, que son generalmente pero no siempre, las proteínas ribosomales (Retchless and Lawrence, 2012). Las bacterias de crecimiento rápido, presentan poblaciones de ARNt abundantes, esto lleva a un mayor sesgo en el uso de codones en los HEG, en relación al resto del genoma. Como resultado, las bacterias de corto tiempo generacional presentan un sesgo en el uso de codones más fuerte que el que puede observarse en organismos procariotas de crecimiento lento (Rocha, 2004).

A diferencia de las comunidades de ambientes ricos cuyas funciones dominantes son la síntesis ribosomal y las actividades dedicadas a incrementar la tasa de crecimiento, las funciones en comunidades microbianas de ambientes con recursos limitados, como el ambiente oligotrófico del océano, se basan en la adquisición de recursos y asimilación de proteínas (Grzymski & Dussaq, 2012). Se ha postulado que en estas condiciones la supervivencia de los organismos depende de una competencia eficiente por los recursos que se encuentran en baja densidad en el ambiente, esto requeriría la expresión de transportadores de alta afinidad para diferentes sustratos (Sowell et al., 2009). Estudios sobre las proteínas del clado SAR11 en el mar de los Sargazos indicaron que las más frecuentemente detectadas incluyen proteínas de transporte para sustratos como fosfato y aminoácidos (Sowell et al., 2009). Este es el caso del clado Pelagibacteriaceae, por lo tanto la aproximación tradicional, que utiliza los genes de ARNr como referencia de genes de alta expresión, no es válida en este caso, y es necesario obtener datos de expresión para analizar el uso de codones sinónimos en relación a HEG. Como ya se ha mencionado, Pelagibacteriaceae es un clado adaptado a un ambiente oligotrófico, con un pequeño genoma rico en A+T, cuyo tamaño dadas las evidencias sería resultado de fuertes presiones selectivas, y posee una tasa de crecimiento lento. En este trabajo se estudia el patrón de variación en el UCS de esta familia a nivel intragenómico. De acuerdo a los antecedentes la variación intragenómica en especies de crecimiento lento puede ser mínima (Rocha, 2004), pero presiones selectivas detectables sobre subsets de codones sinónimos son características de los organismos de vida libre (*E.coli*, *Salmonella*) como Pelagibacteriaceae, en ese caso, se debe testar la hipótesis de que existe selección operando a nivel de la traducción. Es además de interés general extender el rango de especies en las cuales el patrón de uso de codones se ha caracterizado, de forma de obtener una base sólida sobre la cual responder preguntas sobre los procesos evolutivos que dirigen la divergencia entre especies en el uso de codones .

Metodología

Datos

Se utilizaron los genomas completamente secuenciados de las 16 especies de Pelagibacteriaceae que se detallan en el capítulo I. Los mismos análisis se llevaron a cabo con el conjunto de genes ortólogos cuyo procedimiento de obtención se detalla en el capítulo I.

Identificación del Patrón de Sesgo en el Uso de Codones

Uso de codones

El uso de codones se estimó con el paquete *seqinr* de R (Charif, D. and Lobry, J.R. 2007), este paquete ofrece programas para el análisis exploratorio y visualización de datos de secuencias biológicas (ADN y proteínas).
La función que calcula el uso de codones, *uco*, devuelve el conteo de codones, la frecuencia relativa y el rscu (del inglés: *relative sinonymous codon usage*) a partir de las secuencias nucleotídicas. Se calculó el uso de codones para, todos los genes del genoma y para el set de genes ortólogos, en cada especie (la lista de especies, la obtención de genomas y de genes ortólogos se detalla en el capítulo I).

Análisis de Correspondencia

Para estudiar la variabilidad en el uso de codones se realizó un análisis de correspondencia (CA). Este análisis se usa para detectar y cuantificar de forma confiable los patrones de uso de codones sinónimos, es un método de estadística multivariada que puede resumir una matriz de datos de múltiples dimensiones, como conteo de codones por gen, reduciéndolos a un número limitado de variables llamadas ejes. Los ejes retienen mucha de la información sobre la variación de uso de codones entre genes, haciendo esas diferencias más fáciles de comprender (Suzuki et al., 2008). El CA es una técnica gráfica que representa información contenida en una tabla de contingencia de dos vías la cual representa la totalización de las observaciones de una muestra dada, para una tabla cruzada de dos variables categóricas. En este caso la matriz de entrada es el conteo de uso de cada codón (columnas) por gen (filas). Con los perfiles de filas y columnas descritos se elabora la matriz de coordenadas (distancias) utilizando la distancia chi-cuadrado, que permitirá calibrar la magnitud de las diferencias entre la tabla de datos analizada y una tabla de datos sin relación entre las variables, el objetivo es construir un sistema de coordenadas (generalmente,bidimensional) asociado a las filas y columnas de la tabla de contingencia, que refleje las relaciones existentes entre dichas filas y columnas. Las distancias no se miden entre dos filas o dos columnas sino con relación al perfil medio de fila o columna, es decir, con relación al promedio de las coordenadas de esa fila (o columna) ponderada por su masa (peso proporcional a su importancia en el conjunto).
Con el CA se construye una gráfica que señala la interacción de dos variables categóricas a través de la relación de las filas y de las columnas entre sí. La finalidad principal de este estudio es observar la relación dependencia entre los genes (filas de la matriz) de acuerdo al uso de codones de los mismos (columnas). La Inercia Total cuantifica el grado de dependencia entre las variables, las proporciones de inercia acumulada

explicada por las i - ésimas primeras dimensiones permiten decidir el número mínimo de dimensiones necesario para explicar dichas dependencias. En la representación gráfica, dos puntos-filas que están muy juntos representan filas con perfiles similares a través de las columnas (Legendre and Legendre, 1998).

El análisis de correspondencia del uso de codones es una herramienta ampliamente utilizada en el análisis de secuencias, pero la variabilidad en composición aminoacídica entre las comúnmentees un factor de confusión cuando se intenta analizar el uso de codones sinónimos entre secuencias nucleotídicas (Charif et al., 2005). Diferentes aproximaciones se han usado para remover los efectos de la composición aminoacídica, comunmente, el CA se realiza sobre datos de uso de codones que han sido ajustados a la frecuencia del aa que codifican. Las frecuencias relativas y el rscu se utilizan rutinariamente en lugar de la frecuencia absoluta de codones, pero los estudios han mostrado que fallan en remover el sesgo asociado a la baja frecuencia de cisteína en las proteínas (Charif et al., 2005).

Una forma simple de enfrentar este problema es usando el *within-group correspondence analysis* (WCA).

El WCA se realizó en R con la librería ade4 (Thioulouse et al., 1997). El WCA se realiza sobre el conteo de codones de forma tal de no introducir pesos injustificados sobre los datos (http://pbil.univ-lyon1.fr/datasets/charif04/). El WCA es más efectivo que otros métodos a la hora de generar ejes que reflejen la variación en el UCS, además los estudios a traves de WCA han revelado fuentes de variación previamente no identificadas (Suzuki et al., 2008).

El resultado del WCA permitirá analizar el uso de codones por gen y el grado de similitud entre ellos reflejando el sesgo dentro del genoma.

Identificación de Factores Asociados al Sesgo en el UCS

Fuentes de variación

Para identificar las principales fuentes de variación entre genes en los ejes generados por el WCA sobre el uso de codones se llevaron a cabo análisis considerando características comunmente usadas de los genes codificantes: GC3, CAI (codon adaptation index), tasa de sustitución sinónima (dS), MELP (MILC-based expression level predictor).

Como se describió en la introducción, existen variables genómicas que se han visto típicamente asociadas a la variación en el UCS. Entre ellas se encuentra GC3, el contenido en G+C en la tercera posición de los codones para cada gen, esta característica refleja la intensidad del sesgo mutacional. GC1, GC2 y GC3 se estimaron con el paquete seqinr de R.

Selección para la Traducción: Tasa de sustitución sinónima y estimaciones de CAI (Codon Adaptation Index) y MELP (MILC-based Expression Level Predictor)

La distancia sinónima (dS) entre los genes (cantidad de cambios sinónimos/sitios sinónimos), refleja la existencia de la selección mostrando valores de dS más bajos cuando se trata de genes de alta expresión, con uso de codones conservado. La tasa de sustitución sinónima se estimó con el paquete PAML por el método de Yang & Nielsen (Yang, 2007).

CAI (Sharp & Li 1987) es un índice especie-específico de adaptación relativa en el uso de codones de un gen, en relación al uso de codones de un set de genes de referencia, usualmente genes de alta expresión. Cuando existen codones óptimos para algunos aminoácidos, los genes de alta expresión parecen preferir su uso. La adaptabilidad relativa de cada codón (w) es la relación entre el uso de cada codón y el del más abundante para ese aminoácido. CAI se define como la media geométrica de estos valores de adaptación relativa, tiene un rango de 0 a 1, siendo 1 el valor para un gen que usa sólo codones óptimos.

Para calcular el CAI de cada gen se necesita el uso de codones de genes los HEG, para Pelagibacteriaceae se tomaron de bibliografía. Se encontraron 23 genes, identificados como genes de alta expresión en trabajos previos realizados con la especie HTCC1062 (Ottesen et al., 2013). Este fue el grupo de genes usado como referencia para la estimación de CAI (Tabla 2.1).

Actualmente hay varios métodos (o medidas) para cuantifcar el uso de codones en genes y predicciones de niveles de expresión (CAI). Estas medidas se ven generalmente influenciadas por propiedades de las secuencias como su largo o composición nucleotídica. Esto puede introducir fuertes sesgos metodológicos, por lo tanto se utilizó además del CAI, un índice resistente a cambios en largo y composición nucleotídica de los genes, MELP (*MILC-based Expression Level Predictor*) (Supek and Vlahoviček, 2005). MILC (*Measure Independent of Length and Composition*), es la base de MELP y un métdo de estimación de sesgo en el uso de codones entre genes:

$$MELP=MILC_{(genoma)}/MILC_{(HEG)}$$

Para cada taxa, los ejes del WCA reflejando la variación en el UC, se correlacionaron con los índices descriptivos de cada gen, G+C3, CAI, MELP y dS, a través de las funciones *cor* y rcorr de los paquetes *stats y Hmisc* en R.

Tabla 2.1: Genes identificados como de alta expresión en HTCC1062

Gen	Proteina
yhdW	Transportador ABC
SAR11_0359_Candidatu	Porina
potD	Spermidine/putrescine-binding protein
smoM	Mannitol-binding protein
atpD	F1-ATP synthase beta chain
SAR11_0769_Candidatu	Probable ABC sugar transporter
tufB	Translation elongation factor EF-Tu
tauA	Taurine transport system
dctP	TRAP dicarboxylate transporter
rpoA	DNA-directed RNA polymerase
rpoB	DNA-directed RNA polymerase
groEL	60-kDa chaperonin
hupA/hupB	DNA-binding protein Hbsu
rpoC	DNA-directed RNA polymerase
dnaK	Chaperone protein
rpsD	Ribosomal protein S4
atpA	H+-transporting tw o-sector ATPase
nusA	Transcription termination factor
hflK	Probable integral membrane proteinase
pnp	Polyribonucleotide nucleotidyltransferase
rplL	Ribosomal protein L7/L12 (L8)

Análisis comparativo de la variabilidad en uso de codones

Para obtener una idea global de la similitud en la variación de uso de codones entre los organismos se realizaron estudios de correlación entre las posiciones de los genes ortólogos en los diferentes ejes generados por WCA.

Para estudiar si la fuerza de asociación variaba con la distancia filogenética se compararon ambas distancias, definiendo la distancia en similaridad de uso de codones como 1-r, siendo r el coeficiente de correlación entre los ejes 1 del WCA ortólogo de dos especies. Para esto se realizó un test de Mantel, diseñado para datos que no son independientes entre sí. Los árboles filogenéticos se utilizan de manera rutinaria para analizar las familias de genes, así como las relaciones entre especies. Una distancia patrística es la suma de las longitudes de las ramas que unen dos nodos en un árbol, donde los nodos son típicamente nodos terminales que representan secuencias de genes existentes o especies. Una matriz de distancias patrísticas calculada a partir de un árbol para todos los pares de genes o especies resume el cambio genético, o el cambio filogenético, representado en el árbol y en los datos. Las distancias pueden ser utilizados para analizar la velocidad de cambio y pueden ser comparadas con otras medidas de diferencia genética (Fourment and Gibbs, 2006).

Codones Óptimos

Para detectar codones óptimos se utilizó un método comparativo. Esta aproximación nos permite identificar los codones favorecidos comparando el uso de codones de alta expresión y el uso de codones en el resto del genoma. Esta metodología supone varias cosas, una de ellas, que los genes de alta expresión se encuentran bajo mayor presión selectiva para el uso de codones óptimo. Otra asunción del método es que las diferencias en el uso de codones entre ambos sets de genes debidas a la selección sobre la eficiencia y/o precisión de la traducción, son significativamente mayores a otras presiones selectivas que podrían actuar sobre los genes HEG (Hershberg and Petrov, 2008).

Para identificar los codones óptimos de un genoma en particular, se calculó para cada codón la frecuencia dentro de su familia (grupo de codones sinónimos) para todos los genes y se comparó con las frecuencias estimadas para el set de genes de alta expresión. Un test de contingencia Chi cuadrado se utilizó para encontrar diferencias significativas entre os tripletes preferidos por los genes de alta expresión y el resto de los genes del genoma.

Resultados

El uso de codones se analizó en todas las regiones codificantes anotadas en los genomas, y en el set de 399 genes ortólogos cuyo proceso de obtención se detalla en el capítulo I.

Variación Inter - Genómica

Se estimó la frecuencia relativa de uso de cada codón, este valor indica la frecuencia de uso de un codón con respecto a sus sinónimos, 1 es la frecuencia del codón sinónimo más frecuentemente observado. El UCS no mostró una amplia variación entre los taxas en estudio. Los codones que muestran mayores frecuencias relativas son los mismos para todos los taxa, y son siempre codones ricos en A+T (U) (Tabla 2.2).

Los perfiles de contenido en G+C (Fig. 2.2) estimados muestran similar distribución en todas las posiciones de los codones. La tercera posición presenta un contenido más extremo como es usual en los genomas procariotas, en general el contenido de los rangos de valores, GC1, GC2, GC3 obtenido, es el esperable para un genoma rico en A+T (Muto and Osawa 1987; Lloyd and Sharp 1993).

A este nivel de estudio, las características de clado son compartidas con *Rickettsia*, un género de alfaproteobacterias cercanamente emparentadas con SAR11 (ver Introducción). *Rickettsia*, con un genoma igualmente rico en A+T y de pequeño tamaño, presenta un perfil de G+C para las 1°, 2° y 3° posición de los codones de 40.4, 31.2 y 18.6% respectivamente. Además, los estudios de uso de codones en *Rickettsia* indican que no existe sesgo entre genes, incluso tomándose en cuenta genes de altos niveles de expresión, por lo tanto no se ha evidenciado tampoco la existencia de selección para la traducción (Andersson and Sharp, 1996; Andersson et al., 1998).

Aunque, como se dijo anteriormente, el patrón general de UC en Pelagibacteriaceae no varía mucho y los codones mas frecuentes son los terminados en A o U, la frecuencia de cada uno de ellos presenta leves variaciones entre familias de aminoácidos, y entre genomas. También varía en ambos niveles, la frecuencia relativa de uso de codones terminados en G y C. IMCC9063, es el taxa que muestra la mayor frecuencia de uso de codones terminados en G y C respecto del resto del clado. Se calculó también el patrón de frecuencias relativas de codones sinónimos obtenido a partir del set de genes ortólogos, y se correlacionó con el calculado para los genomas, se vio fuertemente correlacionado en todas las especies.

- Variación Intra-genómica

Para estudiar la variabilidad de UCS entre los genes se realizó un WCA sobre el conteo de UC a nivel de genomas (WCAg) y a nivel del set de genes ortólogos (WCAo). La variación explicada por los ejes generados en cada análisis es baja, esto es un indicio de que existe poca variación en el UCS entre genes (Tabla 2.3). El gráfico de los ejes del WCA señala la relación de dependencia entre los genes en cuanto al UCS, algunos de estos *biplots* se muestran en la Figura 2.3. Se trata de los taxa HTCC1062 del subclado I, HTCC8051, HTCC7211 y HTCC7214 del subclado II y IMCC9063, la especie más distante filogenéticamente al resto. Se

Tabla 2.2: Tabla de frecuencia relativa de uso de codones para cada genomas

AA	Ala				Arg						Asn		Glu		Ile			Pro			
codones	GCU	GCC	GCA	GCG	CGU	CGC	CGA	CGG	AGA	AGG	AAU	AAC	GAA	GAG	AUU	AUC	AUA	CCU	CCC	CCA	CCG
HIMB058	1.000	0.167	0.950	0.151	0.062	0.012	0.062	0.012	1.000	0.123	1.000	0.272	1.000	0.298	1.000	0.190	0.611	0.868	0.111	1.000	0.132
HTCC9565	0.906	0.131	1.000	0.116	0.059	0.006	0.069	0.006	1.000	0.093	1.000	0.204	1.000	0.262	1.000	0.154	0.556	0.722	0.101	1.000	0.080
HTCC1040	0.950	0.135	1.000	0.108	0.057	0.006	0.066	0.007	1.000	0.089	1.000	0.221	1.000	0.261	1.000	0.155	0.529	0.744	0.108	1.000	0.077
HTCC1016	0.899	0.130	1.000	0.109	0.061	0.007	0.071	0.007	1.000	0.092	1.000	0.207	1.000	0.260	1.000	0.151	0.553	0.730	0.101	1.000	0.076
HTCC1002	0.934	0.164	1.000	0.112	0.077	0.006	0.065	0.008	1.000	0.095	1.000	0.252	1.000	0.252	1.000	0.165	0.534	0.774	0.143	1.000	0.075
HTCC1062	0.933	0.141	1.000	0.117	0.062	0.006	0.070	0.006	1.000	0.089	1.000	0.236	1.000	0.270	1.000	0.161	0.520	0.743	0.127	1.000	0.075
HTCC1013	0.928	0.157	1.000	0.115	0.074	0.005	0.064	0.007	1.000	0.092	1.000	0.253	1.000	0.254	1.000	0.162	0.528	0.770	0.145	1.000	0.069
HIMB083	0.909	0.128	1.000	0.115	0.055	0.006	0.066	0.007	1.000	0.093	1.000	0.236	1.000	0.278	1.000	0.161	0.558	0.742	0.094	1.000	0.078
HIMB140	0.931	0.163	1.000	0.112	0.075	0.006	0.065	0.007	1.000	0.095	1.000	0.253	1.000	0.258	1.000	0.163	0.531	0.773	0.145	1.000	0.069
HTCC7214	0.879	0.137	1.000	0.111	0.059	0.008	0.079	0.007	1.000	0.096	1.000	0.210	1.000	0.272	1.000	0.173	0.516	0.718	0.099	1.000	0.075
HTCC7217	0.890	0.114	1.000	0.111	0.049	0.004	0.062	0.004	1.000	0.085	1.000	0.251	1.000	0.276	1.000	0.171	0.543	0.734	0.088	1.000	0.072
HTCC7211	0.937	0.169	1.000	0.107	0.066	0.009	0.056	0.008	1.000	0.117	1.000	0.233	1.000	0.247	1.000	0.152	0.595	0.715	0.142	1.000	0.067
HTCC8051	0.940	0.164	1.000	0.112	0.076	0.007	0.061	0.007	1.000	0.094	1.000	0.254	1.000	0.258	1.000	0.163	0.535	0.765	0.146	1.000	0.075
HTCC9022	0.908	0.133	1.000	0.115	0.060	0.007	0.069	0.007	1.000	0.088	1.000	0.207	1.000	0.264	1.000	0.155	0.547	0.713	0.096	1.000	0.082
HIMB4	0.929	0.162	1.000	0.114	0.077	0.006	0.063	0.007	1.000	0.090	1.000	0.255	1.000	0.253	1.000	0.163	0.526	0.770	0.146	1.000	0.072
IMCC9063	0.982	0.310	1.000	0.182	0.114	0.032	0.136	0.020	1.000	0.170	1.000	0.299	1.000	0.318	1.000	0.189	0.442	1.000	0.300	1.000	0.159

AA	Asp		Cys		Gln		Gly				Val				Phe		His		Lys	
codones	GAU	GAC	UGU	UGC	CAA	CAG	GGU	GGC	GGA	GGG	GUU	GUC	GUA	GUG	UUU	UUC	CAU	CAC	AAA	AAG
HIMB058	1.000	0.248	1.000	0.340	1.000	0.207	0.933	0.284	1.000	0.160	1.000	0.167	0.950	0.151	1.000	0.173	1.000	0.397	1.000	0.154
HTCC9565	1.000	0.192	1.000	0.234	1.000	0.168	1.000	0.211	0.992	0.139	0.906	0.131	1.000	0.116	1.000	0.118	1.000	0.273	1.000	0.136
HTCC1040	1.000	0.194	1.000	0.237	1.000	0.176	1.000	0.191	0.946	0.146	0.950	0.135	1.000	0.108	1.000	0.119	1.000	0.298	1.000	0.143
HTCC1016	1.000	0.187	1.000	0.228	1.000	0.167	1.000	0.206	0.980	0.130	0.899	0.130	1.000	0.109	1.000	0.119	1.000	0.277	1.000	0.137
HTCC1002	1.000	0.217	1.000	0.255	1.000	0.194	1.000	0.225	0.907	0.160	0.934	0.164	1.000	0.112	1.000	0.126	1.000	0.364	1.000	0.158
HTCC1062	1.000	0.190	1.000	0.256	1.000	0.175	1.000	0.202	0.975	0.168	0.933	0.141	1.000	0.117	1.000	0.120	1.000	0.315	1.000	0.145
HTCC1013	1.000	0.212	1.000	0.264	1.000	0.180	1.000	0.224	0.900	0.163	0.928	0.157	1.000	0.115	1.000	0.125	1.000	0.356	1.000	0.158
HIMB083	1.000	0.201	1.000	0.250	1.000	0.180	0.978	0.185	1.000	0.139	0.909	0.128	1.000	0.115	1.000	0.142	1.000	0.291	1.000	0.135
HIMB140	1.000	0.213	1.000	0.261	1.000	0.181	1.000	0.222	0.904	0.161	0.931	0.163	1.000	0.112	1.000	0.127	1.000	0.357	1.000	0.162
HTCC7214	1.000	0.183	1.000	0.237	1.000	0.180	1.000	0.220	0.906	0.142	0.879	0.137	1.000	0.111	1.000	0.130	1.000	0.284	1.000	0.142
HTCC7217	1.000	0.198	1.000	0.255	1.000	0.187	0.986	0.174	1.000	0.129	0.890	0.114	1.000	0.111	1.000	0.149	1.000	0.330	1.000	0.140
HTCC7211	1.000	0.211	1.000	0.235	1.000	0.191	1.000	0.229	0.923	0.159	0.937	0.169	1.000	0.107	1.000	0.117	1.000	0.339	1.000	0.157
HTCC8051	1.000	0.215	1.000	0.265	1.000	0.184	1.000	0.234	0.910	0.163	0.940	0.164	1.000	0.112	1.000	0.125	1.000	0.361	1.000	0.158
HTCC9022	1.000	0.191	1.000	0.232	1.000	0.164	1.000	0.215	0.989	0.135	0.908	0.133	1.000	0.115	1.000	0.121	1.000	0.280	1.000	0.141
HIMB4	1.000	0.211	1.000	0.257	1.000	0.177	1.000	0.222	0.911	0.161	0.929	0.162	1.000	0.114	1.000	0.126	1.000	0.357	1.000	0.158
IMCC9063	1.000	0.305	1.000	0.556	1.000	0.310	0.720	0.263	1.000	0.185	0.982	0.310	1.000	0.182	1.000	0.134	1.000	0.459	1.000	0.212

AA	Tre				Tyr		Leu						Ser					
codones	ACU	ACC	ACA	ACG	UAU	UAC	UUA	UUG	CUU	CUC	CUA	CUG	UCC	UCA	UCG	AGU	AGC	UCU
HIMB058	1.000	0.373	1.000	0.124	0.702	0.172	1.000	0.171	0.325	0.044	0.191	0.037	0.118	1.000	0.126	0.572	0.280	0.912
HTCC9565	1.000	0.236	1.000	0.105	0.648	0.140	1.000	0.182	0.315	0.030	0.179	0.033	0.090	1.000	0.091	0.586	0.179	0.816
HTCC1040	1.000	0.251	1.000	0.098	0.619	0.142	1.000	0.188	0.330	0.032	0.190	0.034	0.091	1.000	0.090	0.597	0.176	0.838
HTCC1016	1.000	0.245	1.000	0.108	0.640	0.137	1.000	0.183	0.327	0.031	0.182	0.032	0.086	1.000	0.091	0.582	0.179	0.798
HTCC1002	1.000	0.274	1.000	0.097	0.545	0.150	1.000	0.167	0.341	0.030	0.207	0.042	0.106	1.000	0.091	0.648	0.225	0.902
HTCC1062	1.000	0.261	1.000	0.105	0.583	0.152	1.000	0.194	0.328	0.031	0.196	0.036	0.096	1.000	0.091	0.600	0.190	0.825
HTCC1013	1.000	0.277	1.000	0.096	0.555	0.146	1.000	0.165	0.337	0.030	0.204	0.037	0.104	1.000	0.087	0.644	0.219	0.888
HIMB083	1.000	0.270	1.000	0.102	0.632	0.140	1.000	0.192	0.304	0.031	0.172	0.036	0.093	1.000	0.096	0.608	0.191	0.859
HIMB140	1.000	0.277	1.000	0.098	0.555	0.147	1.000	0.163	0.337	0.029	0.204	0.038	0.103	1.000	0.085	0.649	0.217	0.887
HTCC7214	1.000	0.238	1.000	0.111	0.578	0.159	1.000	0.196	0.319	0.033	0.182	0.037	0.101	1.000	0.087	0.613	0.180	0.844
HTCC7217	1.000	0.299	1.000	0.098	0.648	0.132	1.000	0.213	0.305	0.030	0.177	0.036	0.093	1.000	0.091	0.619	0.188	0.868
HTCC7211	1.000	0.256	1.000	0.111	0.587	0.145	1.000	0.189	0.325	0.035	0.203	0.039	0.110	1.000	0.090	0.641	0.193	0.895
HTCC8051	1.000	0.281	1.000	0.099	0.543	0.145	1.000	0.169	0.338	0.032	0.204	0.039	0.109	1.000	0.093	0.641	0.217	0.922
HTCC9022	1.000	0.245	1.000	0.105	0.631	0.135	1.000	0.181	0.320	0.029	0.180	0.032	0.084	1.000	0.090	0.585	0.182	0.824
HIMB4	1.000	0.282	1.000	0.097	0.548	0.145	1.000	0.168	0.337	0.031	0.205	0.038	0.111	1.000	0.090	0.650	0.216	0.900
IMCC9063	1.000	0.403	1.000	0.185	0.661	0.283	1.000	0.368	0.405	0.064	0.283	0.094	0.241	0.622	0.129	0.503	0.304	1.000

pueden ver distintas distribuciones, por ejemplo en los gráficos de HTCC8051 y HTCC1062 se ve que ambos ejes logran captar la variación entre genes mientras que en los gráficos de HTCC7211 y HTCC7214 esto no se consigue y alguno de los ejes concentra los puntos cerca del 0, y el otro abarca toda la variación. Este problema se evidencia a su vez cuando se realiza un test de correlación entre los ejes 1 y 2 del WCAg, versus ejes 1 y 2 de WCAo (Tabla 2.4).

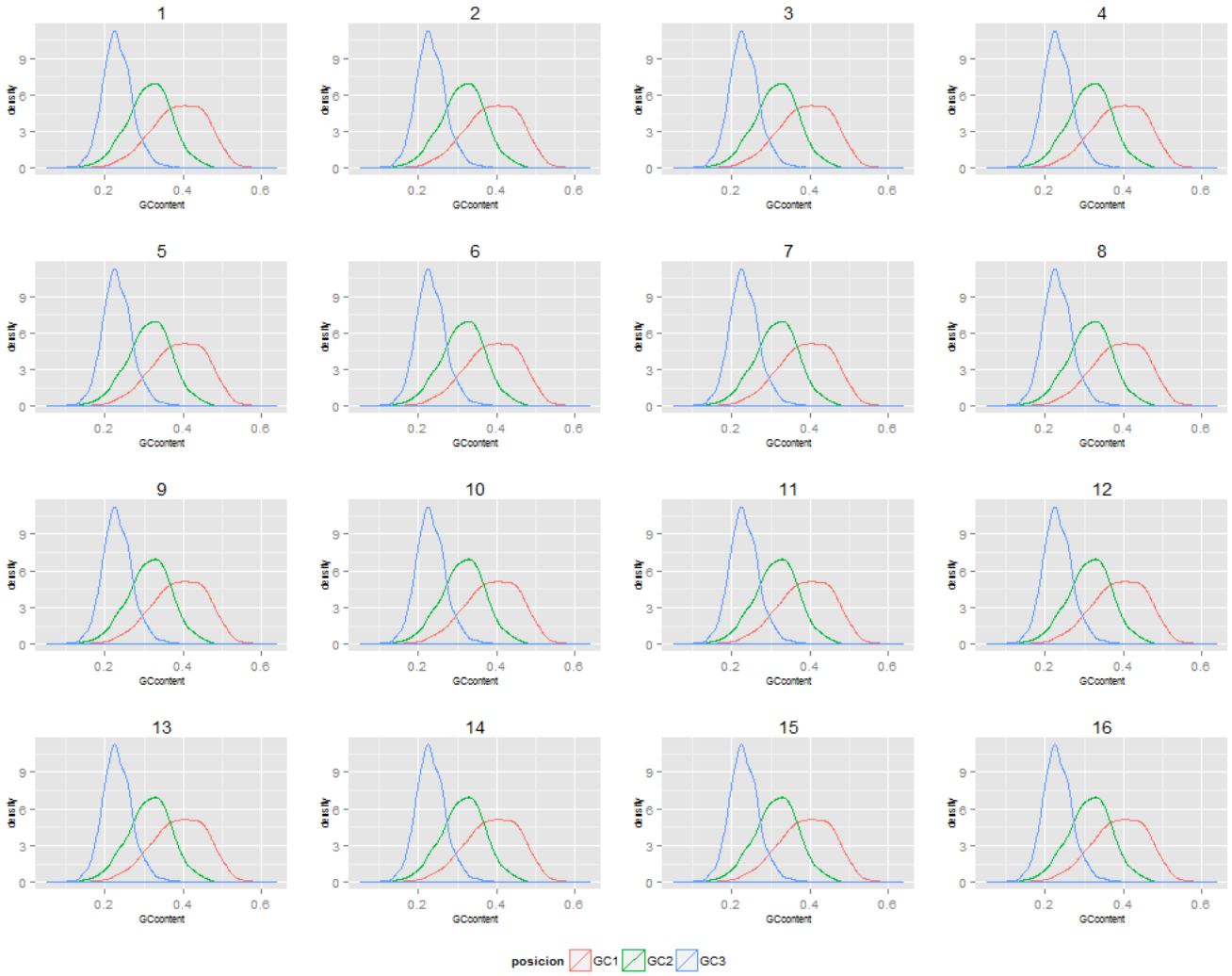


Figura 2.2: Gráficos de densidad de la distribución de G+C en las tres posiciones de los codones para todos los genomas del clado

En la mayoría de los casos, las correlaciones son altas y significativas, esto es muy importante porque que permite extrapolar análisis llevados a cabo con los genes ortólogos a todo el genoma, pero en otros casos, como los nombrados anteriormente, HTCC7214 y HTCC7211, y en un taxa del subclado I, HTCC1002, los ejes del WCAo y WCAg no se correlacionan. Lo contradictorio es, que cuando se compara la variación entre especies a nivel ortólogo, o sea se estudia la correlación entre los ejes del WCAo (Fig 2.3), estos mismos taxas, muestran alta correspondencia con sus especies hermanas. El problema está entonces en el WCAg.

Varios motivos pueden estar generando las discordancia, puede deberse a un factor biológico afectando la variabilidad entre genes en algunos genomas, por ejemplo genes obtenidos recientemente por HGT, cuyo patrón de UC sea significativamente distinto al resto del genoma. Pero las diferencias entre WCAg y WCAo también pueden deberse a artefactos generados por el análisis. Por lo tanto, si trabajamos a nivel de genes ortólogos nos enfocamos en la herencia vertical, reduciendo el efecto de la HGT y la recombinación y comparando el sesgo en el UC entre genes similares de distintos genomas (Iriarte et al., 2013).

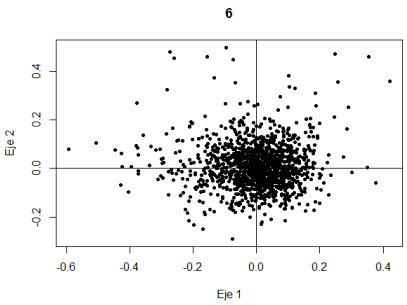
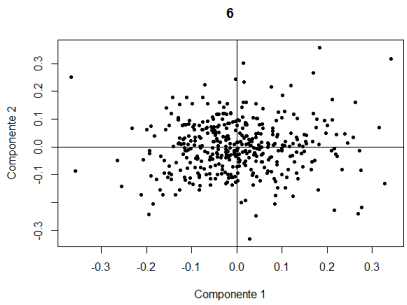
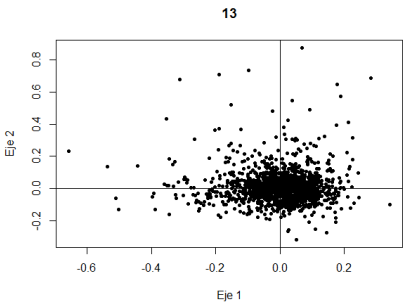
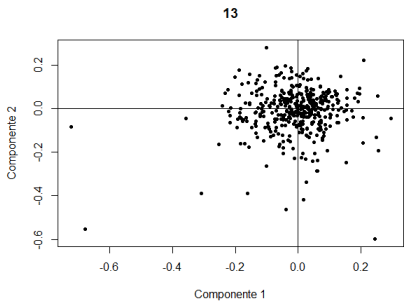
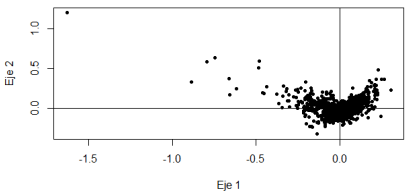
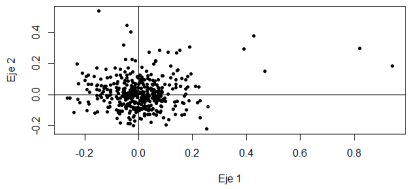
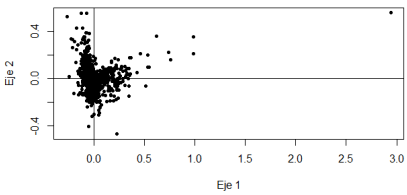
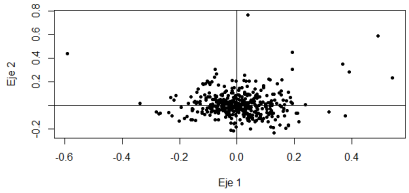
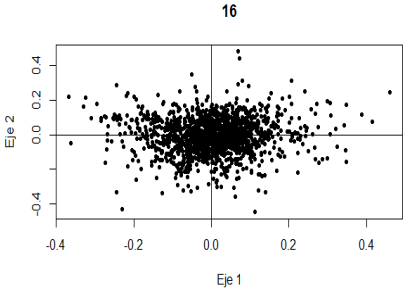
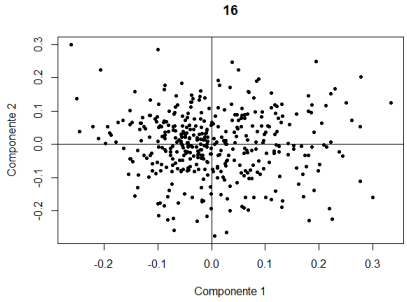
Especie	Genomas	Ortólogos
HTCC1062 Genoma-Eje1 y Eje2 6.183254 3.972905 Ortólogos-Eje1 y Eje2 6.229055 4.177582		
HTCC8051 Genoma-Eje1 y Eje2 5.624802 3.672399 Ortólogos-Eje1 y Eje2 5.155706 4.365966		
HTCC7217 Genoma-Eje1 y Eje2 5.180535 4.791739 Ortólogos-Eje1 y Eje2 4.817044 4.189997		
HTCC7214 Genoma-Eje1 y Eje2 7.346239 5.076495 Ortólogos-Eje1 y Eje2 5.072791 4.327924		
IMCC9063 Genoma-Eje1 y Eje2 4.623357 4.352021 Ortólogos-Eje1 y Eje2 5.400946 4.573729		

Figura 2.3 : Biplots de genomas y de ortólogos que reflejan la variedad de resultados obtenidos

Tabla 2.3: Porcentaje de variabilidad explicada por ejes del WCA
Para genomas y ortólogos

	WCA genomas		WCA ortólogos	
	Eje 1	Eje 2	Eje 1	Eje 2
HIMB058	6.806172	5.609114	4.737984	4.066717
HTCC9565	5.114782	3.852696	4.540417	4.19232
HTCC1040	6.951696	5.448644	6.015952	4.088594
HTCC1016	9.344319	6.480152	6.112506	4.105715
HTCC1002	6.098672	4.629232	6.057947	4.230091
HTCC1062	6.183254	3.972905	6.229055	4.177582
HTCC1013	5.990329	3.898379	6.109183	3.969196
HIMB083	5.863633	3.54709	5.237197	4.064987
HIMB140	6.450416	3.590783	5.442815	4.211147
HTCC7214	7.346239	5.076495	5.072791	4.327924
HTCC7217	5.180535	4.791739	4.817044	4.189997
HTCC7211	5.313885	4.167505	5.206886	4.43853
HTCC8051	5.624802	3.672399	5.155706	4.365966
HTCC9022	5.685419	3.655532	5.524382	4.087863
HIMB4	8.79208	4.930483	4.867473	4.159364
IMCC9063	4.623357	4.352021	5.400946	4.573729

Tabla 2.4: Correlaciones entre ejes 1 y 2 de WCA
Para genomas y para ortólogos

	eje1	eje2
HIMB058	-0.6355030654	0.4185730108
HTCC9565	0.7955383641	-0.6154709676
HTCC1040	-0.6847882323	0.15014356
HTCC1016	-0.4856919576	0.0108257714
HTCC1002	-0.968711205	-0.0501468648
HTCC1062	-0.9693331825	-0.5603258737
HTCC1013	0.9633334574	0.0584333229
HIMB083	0.9220301622	-0.8479592482
HIMB140	0.9536845746	-0.7820844582
HTCC7214	-0.5577358931	-0.0461780238
HTCC7217	-0.6550735351	0.0865215315
HTCC7211	-0.7644264302	0.5919922365
HTCC8051	0.9340489176	-0.8616067909
HTCC9022	-0.9583507729	0.8564043534
HIMB4	0.1768383632	-0.1497057769
IMCC9063	-0.9326874967	0.7620612676

- Contenido en G+C y sesgo en UC

Para el set de genes ortólogos se estudió la correlación entre los ejes 1 y 2 del WCA de las distintas especies (Fig 2.5). Lo que se constató es que existe una correspondencia muy importante, entre los ejes tanto 1 como 2, de las especies pertenecientes al subclado I (definido en la filogenia). Se estudió la relación entre el contenido en G+C de los genes, específicamente el contenido en G+C de la tercera posición de los codones (GC3), y la variación en el uso de codones entre genes, representada por los ejes del WCA.

Se obtuvieron correlaciones importantes del eje 2 con GC3 en la mayoría de las especies, en algunos casos se ve el eje 1 asociado al GC3 (Tabla 2.5 y Fig 2.6). Por ejemplo en la especie IMCC9063, GC3 se correlaciona con el eje 1 y en HIMB083, con ambos ejes. Esto se debe a que no toda la variación existente a nivel de la composición de bases en la tercera posición se puede capturar a través de un solo eje del WCAo. De todas formas la tendencia general entre las especies indica que la composición de bases es uno de los factores explicativos de la variabilidad en el UC.

-Selección para la traducción

Observar que existe un efecto de la selección operando a nivel de la velocidad de traducción implica que el nivel de expresión de un gen es un factor principal a la hora de explicar la variabilidad en el UCS, y que los genes de alta expresión usados como referencia, presentan un sesgo mayor en el UCS. A su vez, los genes de alta expresión presentan generalmente una divergencia sinónima menor (dS) y preferencia por los cuatro codones universales (UUC, AUC, UAC y AAC) definidos por Sharp (Sharp et al., 2005).

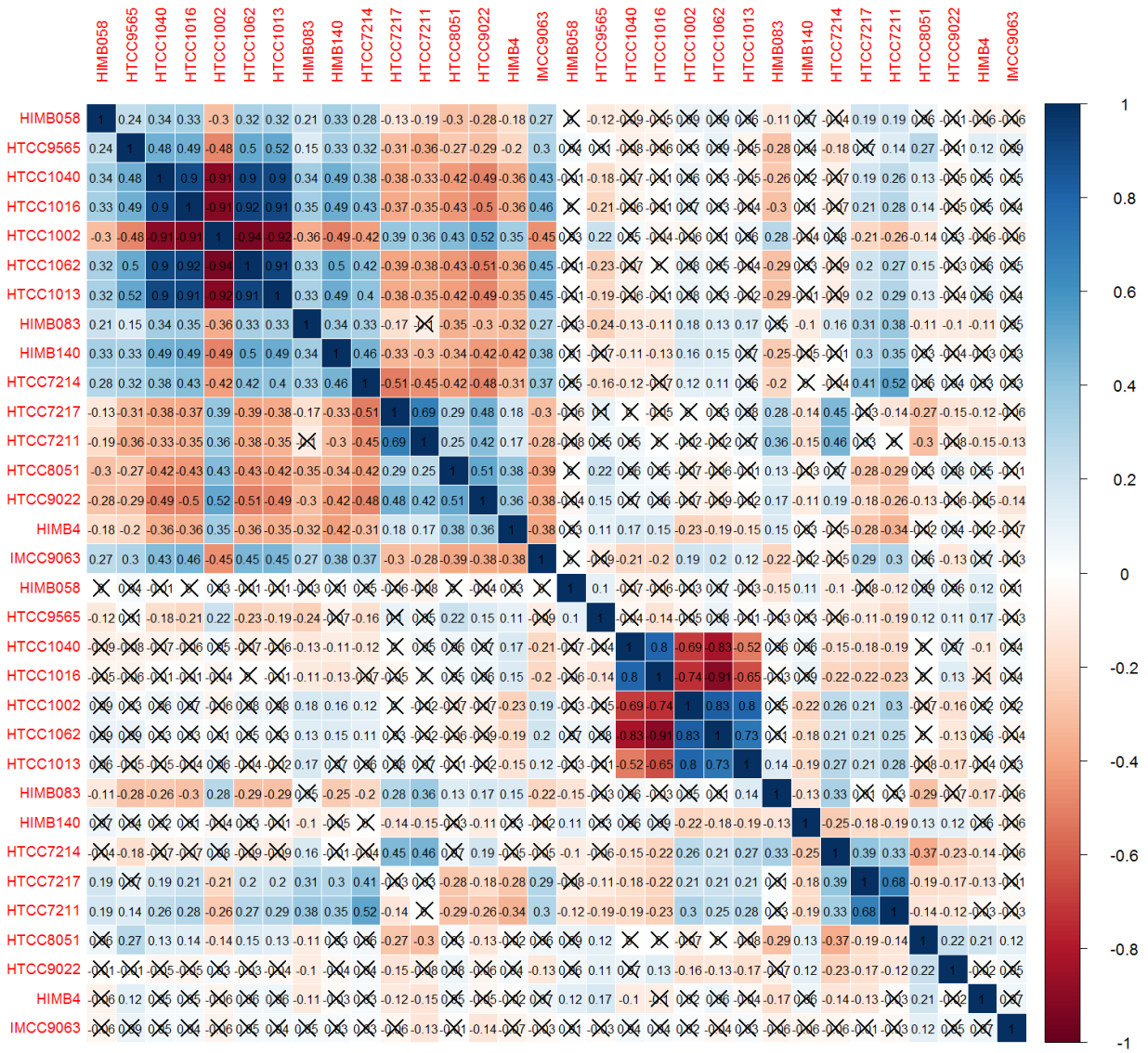


Figura 2.5: Correlación de ejes 1 y 2 del WCA de genes ortólogos entre especies.

Los sesgos en regiones conservadas, o las preferencias de codones óptimos en las regiones conservadas de los genes con respecto a las no conservadas, son indicativos de selección operando a nivel de la traducción (Sharp et al., 2010).

Para encontrar evidencia de selección sobre la traducción, se estimaron los índices CAI y MELP para cada gen en el conjunto de ortólogos usando un conjunto de genes de alta expresión obtenido de estudios de expresión en el clado SAR11, más específicamente, estudios realizados en el taxa HTCC1062 (Ottesen et al., 2013). Los resultados de la estimación de CAI mostraron que la mayoría de los genes tienen valores altos, este índice abarca valores del 0 al 1, siendo 1 el valor otorgado a un gen que usa sólo codones óptimos. La media del CAI estimado en los genomas de Pelagibacteriaceae es siempre mayor a 0.5, en todos los taxa, en la mayoría está

Tabla 2.5: Correlaciones de los ejes 1 y 2 del WCA en ortólogos vs GC3 , CAI, y MELP

	Eje1-GC3		Eje2-GC3		Eje1-CAI		Eje2-CAI		MELP-Eje1		MELP-Eje2	
	r	P-val	r	P-val	r	P-val	r	P-val	r	P-val	r	P-val
HIMB058	0.15	0.00E+000	-0.62	4.36E-139	0.24	6.02E-202	0.59	0.00E+000	0.73	0.00E+000	0.05	2.77E-001
HTCC9565	-0.44	1.45E-002	-0.54	0.00E+000	0.71	0.00E+000	0.41	3.14E-299	0.66	0.00E+000	-0.40	4.98E-017
HTCC1040	-0.21	8.88E-016	0.06	1.55E-005	0.52	4.01E-037	-0.17	2.22E-016	0.88	0.00E+000	-0.21	1.96E-005
HTCC1016	0.22	0.00E+000	0.30	3.94E-004	-0.51	1.55E-164	0.16	1.84E-013	0.87	0.00E+000	-0.20	3.99E-005
HTCC1002	-0.25	6.04E-026	0.37	1.67E-261	0.54	0.00E+000	-0.23	0.00E+000	-0.86	3.25E-118	0.26	2.17E-007
HTCC1062	-0.22	8.52E-002	0.02	0.00E+000	0.53	1.26E-034	-0.10	2.08E-174	0.88	0.00E+000	0.23	2.50E-006
HTCC1013	-0.20	3.68E-001	0.09	2.61E-111	0.53	0.00E+000	0.02	0.00E+000	0.87	0.00E+000	0.22	8.37E-006
HIMB083	0.52	0.00E+000	0.52	0.00E+000	-0.18	6.16E-002	-0.67	0.00E+000	0.69	0.00E+000	-0.28	8.40E-009
HIMB140	0.18	6.95E-012	-0.59	0.00E+000	0.29	0.00E+000	0.44	9.31E-202	0.85	0.00E+000	0.11	3.36E-002
HTCC7214	0.11	8.31E-013	0.77	1.06E-010	0.23	3.22E-021	-0.75	6.66E-016	0.77	0.00E+000	-0.14	6.38E-003
HTCC7217	0.36	3.87E-006	0.51	8.49E-010	-0.57	0.00E+000	-0.29	2.77E-007	-0.66	1.58E-051	0.34	2.64E-012
HTCC7211	0.48	3.30E-004	0.60	0.00E+000	-0.68	0.00E+000	-0.30	2.52E-263	-0.54	3.23E-031	0.52	0.00E+000
HTCC8051	-0.27	8.38E-019	-0.72	0.00E+000	-0.09	7.88E-007	0.75	1.70E-207	-0.80	1.27E-091	0.24	7.83E-007
HTCC9022	0.04	8.81E-001	-0.47	7.28E-096	-0.41	0.00E+000	0.40	0.00E+000	-0.83	6.14E-103	0.06	2.09E-001
HIMB4	-0.14	0.00E+000	-0.61	0.00E+000	-0.26	2.50E-135	0.57	7.86E-001	-0.81	1.97E-095	0.03	5.63E-001
IMCC9063	-0.31	1.62E-001	-0.24	5.62E-119	0.57	5.00E-021	0.24	0.00E+000	0.71	0.00E+000	0.07	1.65E-001

por encima de 0.6 y no presenta mucha variación entre genes (sd=0.03-0.04). Los organismos con composición de bases extrema, como es el caso del clado Pelagibacteriaceae, no pueden presentar bajos valores de CAI porque justamente su contenido en G+C define un repertorio limitado de codones que son compartido por los HEG y el resto de los genes (Botzman and Margalit, 2011).

Para corregir este sesgo metodológico se estimó también MELP, este índice muestra en general una media de 0.8 con desvío de 0.09.

Para determinar si el nivel de expresión del gen está influyendo en el UCS, se correlacionaron los ejes del WCAo con el CAI estimado (Tabla 2.5). En la mayoría de los taxa la correlación significativa encontrada es con el eje 1 del WCA, en algunos casos se ve correlación con el eje 2 y al igual que con GC3 hay una especie, HTCC9022, en la que CAI muestra correspondencia con ambos ejes.

Como se mencionó en la introducción, es de esperar que si el CAI es un factor influyente en el UCS de los genomas, también lo sea la tasa de sustitución sinónima (dS), dado que debe existir mayor presión selectiva sobre los genes de alta expresión, resultando en una menor cantidad de variación a este nivel.

Se pudo observar correlación entre MELP y el eje 1 del WCA ortólogo, todas con altos niveles de significancia (Tabla 2.5 y Figuras 2.7).

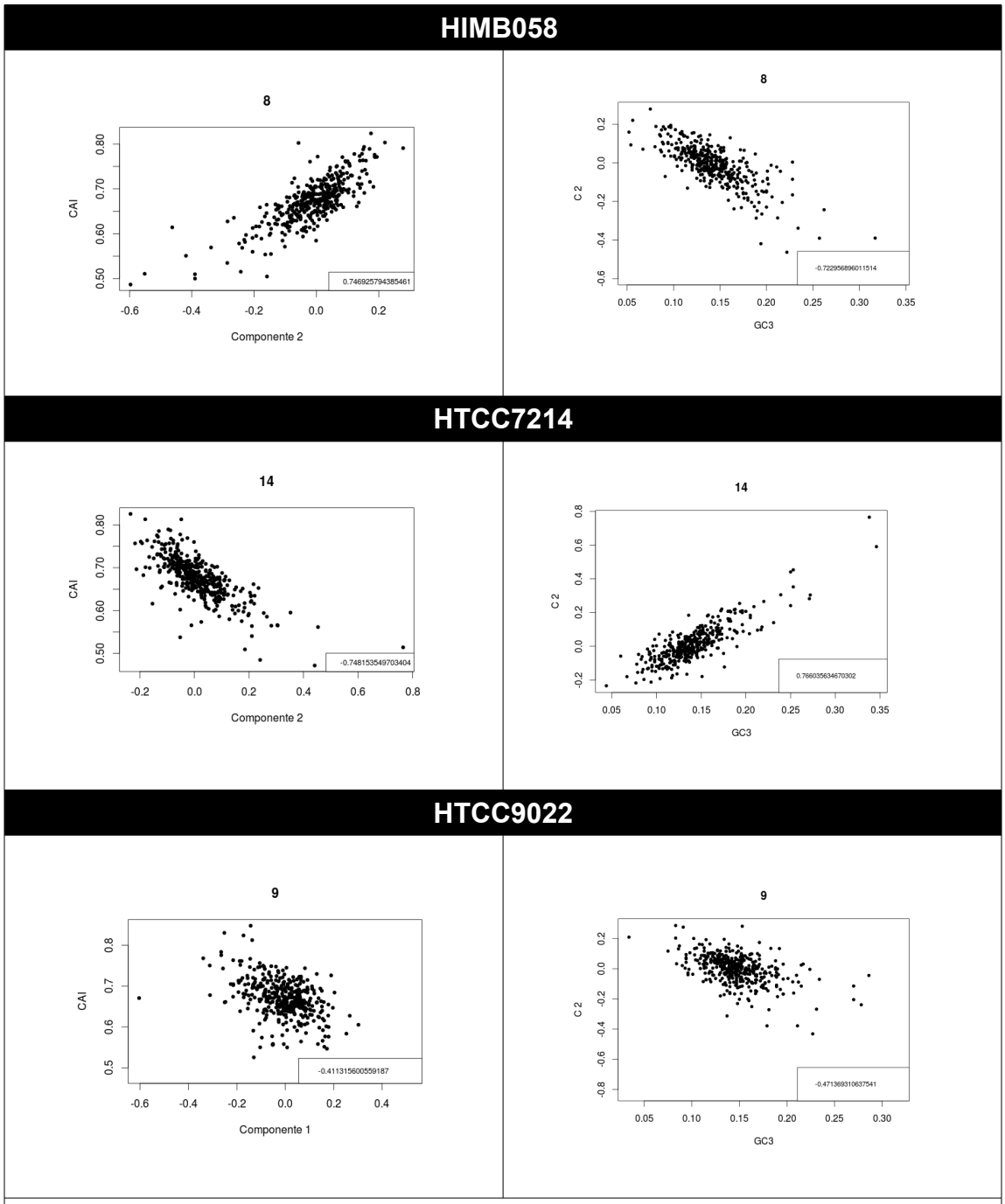


Figura 2.6: Correlaciones entre ejes y CAI o GC3 para tres especies.

Cuando el nivel de expresión es bajo, no hay restricciones y la tasa de sustitución sinónima puede variar mucho. Es usual que se detecten fuertes correlacones entre dS y el nivel de expresión de los genes cuando se estudia la selección para la traducción entre especies más lejanamente emparentadas, por ejemplo *Salmonella typhimurium* y *Escherichia coli.*

- Codones Óptimos

La existencia de diferencias en la preferencia de UCS entre HEG y el resto del genoma se evaluó a través de un test de Chi cuadrado (Tabla 2.6). Dado que la identidad de los codones óptimos puede variar entre especies, es importante enfocar la atención a nivel de cuatro aminoácidos cuyo codones preferidos deben ser siempre los mismos si está actuando la selección en favor de la precisión en la traducción (Sharp et al., 2010). El único ARNt para Phe conocido en bacterias es el anticodón GAA, por lo tanto el codón UUC siempre es preferido sobre su sinónimo UUU. De similar manera, para Tyr, Asn e Ile, G en la posición crítica del anticodón llevaría, siempre que actúe la selección, a preferir codones terminados en C, frente a codones terminados en U (Sharp et al., 2005).

Se encontraron diferencias significativas en el uso de codones entre los HEG y el resto de los genes del genoma, los HEG prefieren duetos terminados en G o C, mientras que usan preferentemente A y U en las terceras posiciones de los codones óptimos en cuartetos (Tabla 2.6).

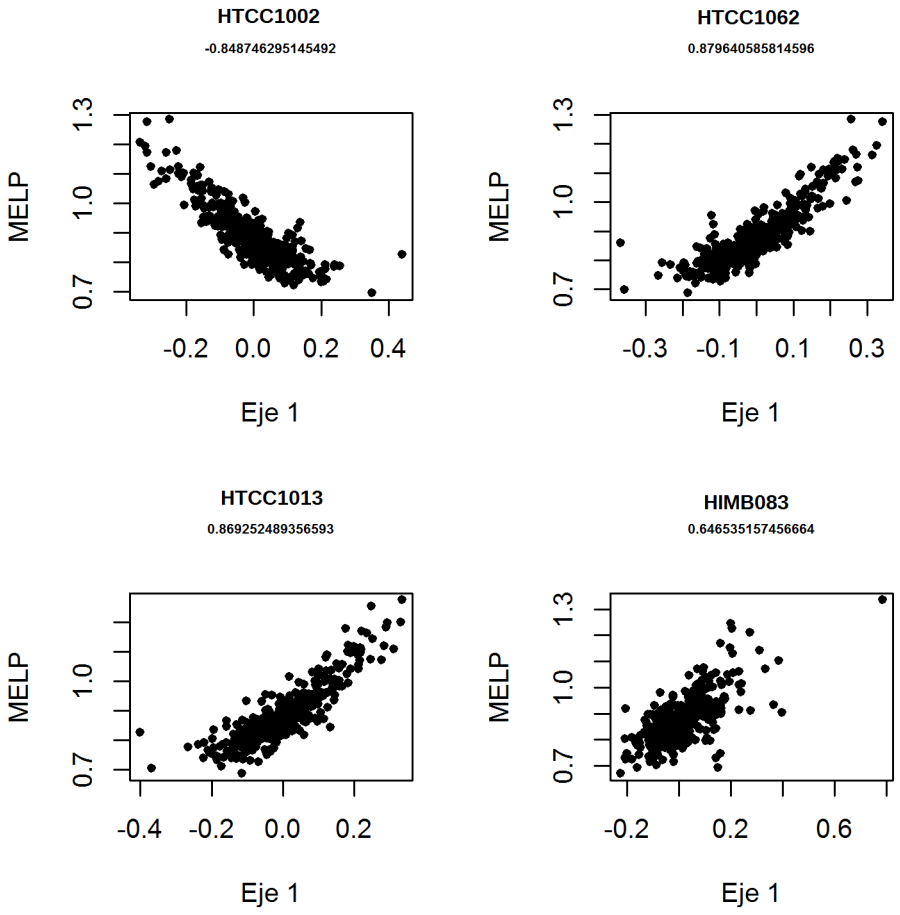


Figura 2.7: Correlaciones entre MELP y el eje 1 del WCAo para cuatro de las especies estudiadas

Se pudo constatar la preferencia de los HEG por los cuatro codones universales UUC (Phe), AUC (Ile), UAC (Tyr) y AAC (Asn), a los que se suman los codones GCU (Ala), GAC (Asp), GGU (Gly), CAC (His), AUU (Ile), CUU (Leu), CCU (Pro), AGA (Arg), UCA (Ser) y GUU (Val).

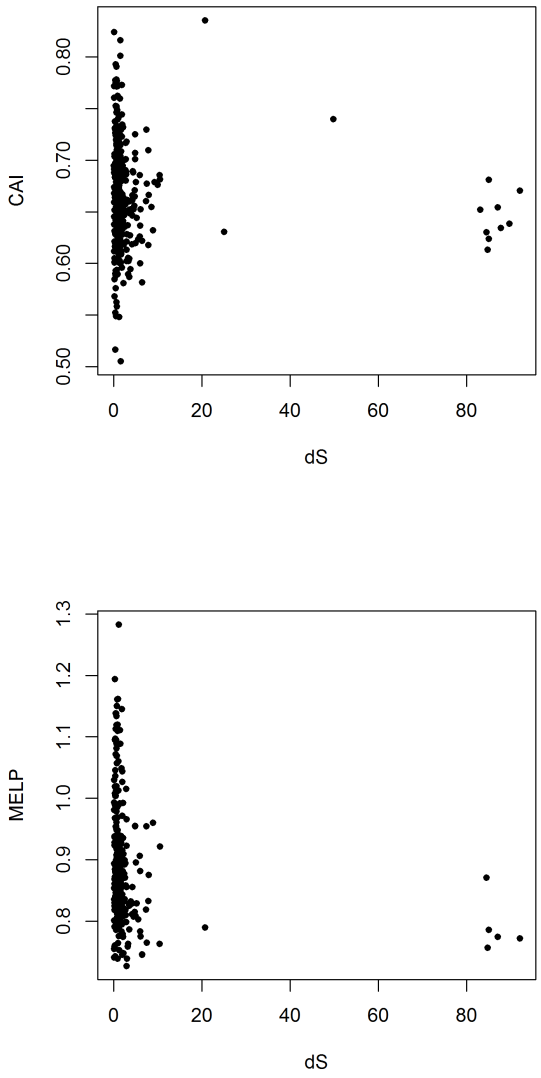


Figura 2.8: Correlaciones entre dS y MELP y entre dS y CAI para la especie HTCC1062

Discusión

El uso de codones en cualquier genoma se espera refleje el balance entre las fuerzas de la selección, la deriva y la mutación, por lo tanto es concebible que diferencias selectivas en el patrón de uso de codones entre genes de diferentes niveles de expresión, sean oscurecidas por un sesgo mutacional extremo.

Tabla 2.6: resultados del test de contongencia Chi cuadrado. Los codones sombreados en gris son los preferidos por los genes de alta expresión

AA		A				C		D		E		F		G				H		I			K		L						N	
codones		GCA	GCC	GCG	GCU	UGC	UGU	GAC	GAU	GAA	GAG	UUC	UUU	GGA	GGC	GGG	GGU	CAC	CAU	AUA	AUC	AUU	AAA	AAG	CUA	CUC	CUG	CUU	UUA	UUG	AAC	AAU
HIMB058	G	0.419	0.0737	0.0667	0.4408	0.254	0.746	0.199	0.801	0.77	0.2297	0.1472	0.853	0.4207	0.1193	0.067	0.39264	0.28418	0.7158	0.3393	0.105	0.555	0.867	0.133	0.108	0.02482	0.0207	0.1838	0.5658	0.097	0.214	0.786
	HEG	0.463	0.0393	0.0281	0.4698	0.18	0.82	0.209	0.791	0.824	0.1759	0.1621	0.838	0.3683	0.0784	0.046	0.507	0.4359	0.5641	0.2126	0.13	0.657	0.875	0.125	0.098	0.00415	0.0152	0.2185	0.5768	0.087	0.282	0.718
HTCC9565	G	0.464	0.061	0.054	0.4207	0.19	0.81	0.1607	0.839	0.792	0.2079	0.1055	0.895	0.4237	0.09	0.059	0.42703	0.21452	0.7855	0.3251	0.09	0.585	0.88	0.12	0.103	0.01719	0.0189	0.1814	0.575	0.104	0.17	0.83
	HEG	0.463	0.0393	0.0281	0.4698	0.18	0.82	0.209	0.791	0.824	0.1759	0.1621	0.838	0.3683	0.0784	0.046	0.507	0.4359	0.5641	0.2126	0.13	0.657	0.875	0.125	0.098	0.00415	0.0152	0.2185	0.5768	0.087	0.282	0.718
HTCC1040	G	0.456	0.0617	0.049	0.4333	0.191	0.809	0.1625	0.837	0.793	0.207	0.1063	0.894	0.4144	0.0836	0.064	0.43798	0.22938	0.7706	0.3141	0.092	0.594	0.875	0.125	0.107	0.01812	0.0194	0.1861	0.5635	0.106	0.181	0.819
	HEG	0.463	0.0393	0.0281	0.4698	0.18	0.82	0.209	0.791	0.824	0.1759	0.1621	0.838	0.3683	0.0784	0.046	0.507	0.4359	0.5641	0.2126	0.13	0.657	0.875	0.125	0.098	0.00415	0.0152	0.2185	0.5768	0.087	0.282	0.718
HTCC1016	G	0.468	0.0606	0.0512	0.4206	0.186	0.814	0.1572	0.843	0.793	0.2065	0.1063	0.894	0.4232	0.0889	0.056	0.43186	0.21699	0.783	0.3245	0.088	0.587	0.88	0.12	0.104	0.01751	0.0185	0.1861	0.5695	0.104	0.171	0.829
	HEG	0.463	0.0393	0.0281	0.4698	0.18	0.82	0.209	0.791	0.824	0.1759	0.1621	0.838	0.3683	0.0784	0.046	0.507	0.4359	0.5641	0.2126	0.13	0.657	0.875	0.125	0.098	0.00415	0.0152	0.2185	0.5768	0.087	0.282	0.718
HTCC1002	G	0.452	0.0744	0.0507	0.4224	0.203	0.797	0.178	0.822	0.798	0.2016	0.1118	0.888	0.3957	0.098	0.07	0.43646	0.2671	0.7329	0.3143	0.097	0.589	0.863	0.137	0.116	0.01701	0.0233	0.1908	0.5594	0.094	0.201	0.799
	HEG	0.463	0.0393	0.0281	0.4698	0.18	0.82	0.209	0.791	0.824	0.1759	0.1621	0.838	0.3683	0.0784	0.046	0.507	0.4359	0.5641	0.2126	0.13	0.657	0.875	0.125	0.098	0.00415	0.0152	0.2185	0.5768	0.087	0.282	0.718
HTCC1062	G	0.456	0.0645	0.0535	0.4257	0.204	0.796	0.16	0.84	0.787	0.2127	0.1069	0.893	0.4157	0.0863	0.071	0.4265	0.23946	0.7605	0.3091	0.096	0.595	0.873	0.127	0.11	0.01729	0.0203	0.1836	0.5605	0.109	0.191	0.809
	HEG	0.463	0.0393	0.0281	0.4698	0.18	0.82	0.209	0.791	0.824	0.1759	0.1621	0.838	0.3683	0.0784	0.046	0.507	0.4359	0.5641	0.2126	0.13	0.657	0.875	0.125	0.098	0.00415	0.0152	0.2185	0.5768	0.087	0.282	0.718
HTCC1013	G	0.455	0.0716	0.0521	0.4218	0.209	0.791	0.1747	0.825	0.798	0.2023	0.1112	0.889	0.3935	0.098	0.071	0.43722	0.26262	0.7374	0.3123	0.096	0.592	0.864	0.136	0.115	0.01677	0.0209	0.1901	0.564	0.093	0.202	0.798
	HEG	0.463	0.0393	0.0281	0.4698	0.18	0.82	0.209	0.791	0.824	0.1759	0.1621	0.838	0.3683	0.0784	0.046	0.507	0.4359	0.5641	0.2126	0.13	0.657	0.875	0.125	0.098	0.00415	0.0152	0.2185	0.5768	0.087	0.282	0.718
HIMB083	G	0.465	0.0596	0.0534	0.4223	0.2	0.8	0.1671	0.833	0.783	0.2173	0.1245	0.876	0.4344	0.0805	0.06	0.42482	0.22569	0.7743	0.3245	0.094	0.582	0.881	0.119	0.099	0.01789	0.0205	0.1752	0.5764	0.111	0.191	0.809
	HEG	0.463	0.0393	0.0281	0.4698	0.18	0.82	0.209	0.791	0.824	0.1759	0.1621	0.838	0.3683	0.0784	0.046	0.507	0.4359	0.5641	0.2126	0.13	0.657	0.875	0.125	0.098	0.00415	0.0152	0.2185	0.5768	0.087	0.282	0.718
HIMB140	G	0.453	0.0738	0.0508	0.4221	0.207	0.793	0.1759	0.824	0.795	0.205	0.113	0.887	0.3953	0.0969	0.07	0.43729	0.26304	0.737	0.3137	0.096	0.59	0.861	0.139	0.115	0.01648	0.0214	0.1901	0.5647	0.092	0.202	0.798
	HEG	0.463	0.0393	0.0281	0.4698	0.18	0.82	0.209	0.791	0.824	0.1759	0.1621	0.838	0.3683	0.0784	0.046	0.507	0.4359	0.5641	0.2126	0.13	0.657	0.875	0.125	0.098	0.00415	0.0152	0.2185	0.5768	0.087	0.282	0.718
HTCC7214	G	0.47	0.0645	0.0523	0.4132	0.192	0.808	0.1548	0.845	0.786	0.2138	0.1153	0.885	0.3993	0.097	0.063	0.44091	0.22124	0.7788	0.3055	0.102	0.592	0.876	0.124	0.103	0.01888	0.0207	0.1804	0.5661	0.111	0.174	0.826
	HEG	0.463	0.0393	0.0281	0.4698	0.18	0.82	0.209	0.791	0.824	0.1759	0.1621	0.838	0.3683	0.0784	0.046	0.507	0.4359	0.5641	0.2126	0.13	0.657	0.875	0.125	0.098	0.00415	0.0152	0.2185	0.5768	0.087	0.282	0.718
HTCC7217	G	0.473	0.054	0.0526	0.4206	0.203	0.797	0.1652	0.835	0.784	0.2161	0.1297	0.87	0.4369	0.0761	0.056	0.43083	0.24831	0.7517	0.3167	0.1	0.583	0.877	0.123	0.1	0.01708	0.0206	0.1733	0.5679	0.121	0.201	0.799
	HEG	0.463	0.0393	0.0281	0.4698	0.18	0.82	0.209	0.791	0.824	0.1759	0.1621	0.838	0.3683	0.0784	0.046	0.507	0.4359	0.5641	0.2126	0.13	0.657	0.875	0.125	0.098	0.00415	0.0152	0.2185	0.5768	0.087	0.282	0.718
HTCC7211	G	0.452	0.0764	0.0482	0.4234	0.19	0.81	0.1739	0.826	0.802	0.1982	0.1048	0.895	0.3994	0.099	0.069	0.43266	0.25291	0.7471	0.3405	0.087	0.572	0.864	0.136	0.113	0.0196	0.0219	0.1816	0.5583	0.105	0.189	0.811
	HEG	0.463	0.0393	0.0281	0.4698	0.18	0.82	0.209	0.791	0.824	0.1759	0.1621	0.838	0.3683	0.0784	0.046	0.507	0.4359	0.5641	0.2126	0.13	0.657	0.875	0.125	0.098	0.00415	0.0152	0.2185	0.5768	0.087	0.282	0.718
HTCC8051	G	0.451	0.0741	0.0504	0.4241	0.209	0.791	0.1768	0.823	0.795	0.2053	0.1112	0.889	0.3946	0.1014	0.071	0.43337	0.26543	0.7346	0.3151	0.096	0.589	0.864	0.136	0.114	0.01792	0.0218	0.1899	0.5612	0.095	0.203	0.797
	HEG	0.463	0.0393	0.0281	0.4698	0.18	0.82	0.209	0.791	0.824	0.1759	0.1621	0.838	0.3683	0.0784	0.046	0.507	0.4359	0.5641	0.2126	0.13	0.657	0.875	0.125	0.098	0.00415	0.0152	0.2185	0.5768	0.087	0.282	0.718
HTC9022	G	0.464	0.0617	0.0533	0.4213	0.188	0.812	0.1603	0.84	0.791	0.2087	0.108	0.892	0.423	0.0919	0.058	0.42752	0.21856	0.7814	0.3215	0.091	0.588	0.877	0.123	0.103	0.01665	0.0184	0.1837	0.574	0.104	0.172	0.828
	HEG	0.463	0.0393	0.0281	0.4698	0.18	0.82	0.209	0.791	0.824	0.1759	0.1621	0.838	0.3683	0.0784	0.046	0.507	0.4359	0.5641	0.2126	0.13	0.657	0.875	0.125	0.098	0.00415	0.0152	0.2185	0.5768	0.087	0.282	0.718
HIMB4	G	0.454	0.0734	0.0516	0.4214	0.205	0.795	0.1742	0.826	0.798	0.2019	0.112	0.888	0.3969	0.0969	0.07	0.43592	0.26285	0.7371	0.3115	0.097	0.592	0.864	0.136	0.115	0.0175	0.0215	0.1895	0.5621	0.094	0.203	0.797
	HEG	0.463	0.0393	0.0281	0.4698	0.18	0.82	0.209	0.791	0.824	0.1759	0.1621	0.838	0.3683	0.0784	0.046	0.507	0.4359	0.5641	0.2126	0.13	0.657	0.875	0.125	0.098	0.00415	0.0152	0.2185	0.5768	0.087	0.282	0.718
IMCC9063	G	0.404	0.1254	0.0735	0.397	0.357	0.643	0.2336	0.766	0.759	0.2413	0.1179	0.882	0.4612	0.1213	0.086	0.33194	0.31471	0.6853	0.2712	0.116	0.613	0.825	0.175	0.128	0.02869	0.0423	0.183	0.4518	0.166	0.23	0.77
	HEG	0.463	0.0393	0.0281	0.4698	0.18	0.82	0.209	0.791	0.824	0.1759	0.1621	0.838	0.3683	0.0784	0.046	0.507	0.4359	0.5641	0.2126	0.13	0.657	0.875	0.125	0.098	0.00415	0.0152	0.2185	0.5768	0.087	0.282	0.718

AA		P				Q		R						S						T			V				Y	
codones		CCA	CCC	CCG	CCU	CAA	CAG	AGA	AGG	CGA	CGC	CGG	CGU	AGC	AGU	UCA	UCC	UCG	UCU	ACA	ACC	ACG	ACU	GUA	GUC	GUG	GUU	UAC	UAU
HIMB058	G	0.474	0.0525	0.0625	0.4112	0.828	0.172	0.7864	0.097	0.049	0.0094	0.0095	0.049	0.0932	0.19	0.332	0.03931	0.04198	0.3032	0.4372	0.08	0.061	0.422	0.351	0.062	0.08628	0.5005	0.2716	0.7284
	HEG	0.489	0.0249	0.0387	0.4475	0.873	0.127	0.9002	0.031	0.01	0	0	0.06	0.053	0.1949	0.409	0.01368	0.02735	0.3026	0.4335	0.038	0.068	0.46	0.305	0.034	0.06009	0.6009	0.4043	0.5957
HTCC9565	G	0.525	0.0533	0.0419	0.3795	0.856	0.144	0.811	0.075	0.056	0.0052	0.0046	0.048	0.0649	0.2121	0.362	0.03243	0.03281	0.2955	0.4419	0.078	0.045	0.436	0.342	0.056	0.07406	0.528	0.1912	0.8088
	HEG	0.489	0.0249	0.0387	0.4475	0.873	0.127	0.9002	0.031	0.01	0	0	0.06	0.053	0.1949	0.409	0.01368	0.02735	0.3026	0.4335	0.038	0.068	0.46	0.305	0.034	0.06009	0.6009	0.4043	0.5957
HTCC1040	G	0.519	0.056	0.0398	0.3856	0.85	0.15	0.817	0.072	0.054	0.0053	0.0054	0.046	0.0631	0.2138	0.358	0.03244	0.03221	0.3002	0.4398	0.076	0.043	0.441	0.333	0.053	0.07648	0.5376	0.2009	0.7991
	HEG	0.489	0.0249	0.0387	0.4475	0.873	0.127	0.9002	0.031	0.01	0	0	0.06	0.053	0.1949	0.409	0.01368	0.02735	0.3026	0.4335	0.038	0.068	0.46	0.305	0.034	0.06009	0.6009	0.4043	0.5957
HTCC1016	G	0.524	0.0531	0.04	0.3827	0.857	0.143	0.8075	0.075	0.057	0.0053	0.0058	0.05	0.0655	0.2127	0.365	0.0316	0.03333	0.2916	0.4433	0.079	0.046	0.432	0.34	0.057	0.07275	0.5305	0.1971	0.8029
	HEG	0.489	0.0249	0.0387	0.4475	0.873	0.127	0.9002	0.031	0.01	0	0	0.06	0.053	0.1949	0.409	0.01368	0.02735	0.3026	0.4335	0.038	0.068	0.46	0.305	0.034	0.06009	0.6009	0.4043	0.5957
HTCC1002	G	0.502	0.0716	0.0376	0.3886	0.837	0.163	0.7986	0.076	0.052	0.0051	0.006	0.062	0.0756	0.218	0.336	0.03574	0.0306	0.3036	0.4359	0.087	0.057	0.42	0.304	0.054	0.08394	0.5578	0.2153	0.7847
	HEG	0.489	0.0249	0.0387	0.4475	0.873	0.127	0.9002	0.031	0.01	0	0	0.06	0.053	0.1949	0.409	0.01368	0.02735	0.3026	0.4335	0.038	0.068	0.46	0.305	0.034	0.06009	0.6009	0.4043	0.5957
HTCC1062	G	0.514	0.0651	0.0386	0.3821	0.851	0.149	0.811	0.072	0.057	0.0047	0.005	0.05	0.0679	0.2141	0.357	0.03411	0.03244	0.2946	0.4361	0.091	0.047	0.426	0.317	0.057	0.08248	0.5435	0.207	0.793
	HEG	0.489	0.0249	0.0387	0.4475	0.873	0.127	0.9002	0.031	0.01	0	0	0.06	0.053	0.1949	0.409	0.01368	0.02735	0.3026	0.4335	0.038	0.068	0.46	0.305	0.034	0.06009	0.6009	0.4043	0.5957
HTCC1013	G	0.504	0.0729	0.0348	0.3881	0.847	0.153	0.8051	0.074	0.051	0.0044	0.0053	0.059	0.0744	0.2188	0.34	0.03546	0.02956	0.3019	0.4371	0.089	0.055	0.419	0.309	0.053	0.08116	0.5566	0.2167	0.7833
	HEG	0.489	0.0249	0.0387	0.4475	0.873	0.127	0.9002	0.031	0.01	0	0	0.06	0.053	0.1949	0.409	0.01368	0.02735	0.3026	0.4335	0.038	0.068	0.46	0.305	0.034	0.06009	0.6009	0.4043	0.5957
HIMB083	G	0.523	0.0489	0.0406	0.3877	0.847	0.153	0.8156	0.076	0.054	0.0046	0.0054	0.045	0.0672	0.2137	0.351	0.03249	0.03374	0.3016	0.4482	0.079	0.045	0.427	0.338	0.054	0.07471	0.5336	0.2125	0.7875
	HEG	0.489	0.0249	0.0387	0.4475	0.873	0.127	0.9002	0.031	0.01	0	0	0.06	0.053	0.1949	0.409	0.01368	0.02735	0.3026	0.4335	0.038	0.068	0.46	0.305	0.034	0.06009	0.6009	0.4043	0.5957
HIMB140	G	0.503	0.0728	0.0347	0.3892	0.847	0.153	0.8014	0.076	0.052	0.0051	0.0057	0.06	0.0737	0.2206	0.34	0.03506	0.02883	0.3017	0.4337	0.088	0.054	0.424	0.308	0.055	0.0819	0.5554	0.2168	0.7832
	HEG	0.489	0.0249	0.0387	0.4475	0.873	0.127	0.9002	0.031	0.01	0	0	0.06	0.053	0.1949	0.409	0.01368	0.02735	0.3026	0.4335	0.038	0.068	0.46	0.305	0.034	0.06009	0.6009	0.4043	0.5957
HTCC7214	G	0.528	0.0525	0.0398	0.3794	0.847	0.153	0.8008	0.077	0.063	0.0064	0.0059	0.047	0.0639	0.217	0.354	0.03579	0.03062	0.2987	0.4394	0.086	0.045	0.43	0.313	0.06	0.08597	0.5411	0.1923	0.8077
	HEG	0.489	0.0249	0.0387	0.4475	0.873	0.127	0.9002	0.031	0.01	0	0	0.06	0.053	0.1949	0.409	0.01368	0.02735	0.3026	0.4335	0.038	0.068	0.46	0.305	0.034	0.06009	0.6009	0.4043	0.5957
HTCC7217	G	0.528	0.0466	0.038	0.3876	0.842	0.158	0.8311	0.071	0.051	0.003	0.0034	0.04	0.0657	0.2165	0.35	0.03239	0.03173	0.3038	0.4447	0.074	0.041	0.44	0.345	0.052	0.07005	0.5325	0.2304	0.7696
	HEG	0.489	0.0249	0.0387	0.4475	0.873	0.127	0.9002	0.031	0.01	0	0	0.06	0.053	0.1949	0.409	0.01368	0.02735	0.3026	0.4335	0.038	0.068	0.46	0.305	0.034	0.06009	0.6009	0.4043	0.5957
HTCC7211	G	0.52	0.0739	0.035	0.3716	0.84	0.16	0.7957	0.093	0.045	0.0074	0.0063	0.053	0.066	0.2187	0.341	0.03761	0.03074	0.3055	0.4572	0.079	0.045	0.419	0.319	0.06	0.07858	0.5427	0.2036	0.7964
	HEG	0.489	0.0249	0.0387	0.4475	0.873	0.127	0.9002	0.031	0.01	0	0	0.06	0.053	0.1949	0.409	0.01368	0.02735	0.3026	0.4335	0.038	0.068	0.46	0.305	0.034	0.06009	0.6009	0.4043	0.5957
HTCC8051	G	0.504	0.0737	0.0376	0.385	0.844	0.156	0.8034	0.076	0.049	0.0055	0.0058	0.061	0.0726	0.215	0.335	0.03668	0.03132	0.309	0.4354	0.087	0.053	0.425	0.304	0.056	0.08133	0.5592	0.2191	0.7809
	HEG	0.489	0.0249	0.0387	0.4475	0.873	0.127	0.9002	0.031	0.01	0	0	0.06	0.053	0.1949	0.409	0.01368	0.02735	0.3026	0.4335	0.038	0.068	0.46	0.305	0.034	0.06009	0.6009	0.4043	0.5957
HTC9022	G	0.529	0.0508	0.0436	0.3769	0.859	0.141	0.8124	0.071	0.056	0.0059	0.0057	0.049	0.0659	0.2115	0.362	0.03038	0.03246	0.2981	0.4425	0.076	0.047	0.435	0.337	0.056	0.07225	0.5345	0.197	0.803
	HEG	0.489	0.0249	0.0387	0.4475	0.873	0.127	0.9002	0.031	0.01	0	0	0.06	0.053	0.1949	0.409	0.01368	0.02735	0.3026	0.4335	0.038	0.068	0.46	0.305	0.034	0.06009	0.6009	0.4043	0.5957
HIMB4	G	0.503	0.0736	0.0363	0.3873	0.85	0.15	0.8044	0.073	0.051	0.0052	0.0054	0.062	0.0729	0.219	0.337	0.03743	0.03021	0.3033	0.4357	0.089	0.054	0.421	0.306	0.054	0.08096	0.5588	0.22	0.78
	HEG	0.489	0.0249	0.0387	0.4475	0.873	0.127	0.9002	0.031	0.01	0	0	0.06	0.053	0.1949	0.409	0.01368	0.02735	0.3026	0.4335	0.038	0.068	0.46	0.305	0.034	0.06009	0.6009	0.4043	0.5957
IMCC9063	G	0.407	0.122	0.0646	0.4067	0.763	0.237	0.6791	0.115	0.092	0.022	0.0136	0.077	0.1086	0.1797	0.222	0.0862	0.04608	0.3574	0.3481	0.153	0.082	0.417	0.31	0.087	0.13305	0.4698	0.2873	0.7127

< 0.001

< 0.01

< 0.05

De hecho en especies con contenido en G+C extremo no se detectan diferencias de UC entre genes (*Micrococcus luteus, Streptomyces ~ 74%, M. Capricolum~25% , Ricketsiales ~30%*) (Andersson and Sharp, 1996; Ohkubo et al., 1987), mientras que en el análisis de especies cuyo contenido en G+C se encuentra entre 50-65% (*Mycobacterium tuberculosis, E.coli)*, hay claras diferencias en el UC entre genes de distintos niveles de expresión (Akashi and Gojobori, 2002; Iriarte et al., 2013).

Los resultados de este estudio mostraron la existencia de sesgo en el UCS entre genes de Pelagibacteriaceae a pesar del contenido G+C extremo de su genoma, esto es interesante porque no es usual que genomas con esta característica, muestren variación intra-genómica en el UCS.

Por ejemplo en *Rickettsia*, clado de alfa-proteobacterias propuesto como grupo hermano de Pelagibacteriaceae (ver Introducción) con un genoma en extremo rico en A+T, no se han encontrado diferencias intra-genómicas en el UCS (Andersson & ShSharp, 1996).

Los patrones de variación en el UCS se vieron asociados tanto a la composición de bases (GC3) como al nivel

de expresión génico, medido a través de CAI y MELP. Esto indica que ambos factores influyen en las diferencias de UCS entre genes. Si la única correlación encontrada hubiera sido con GC3, entonces deberíamos aceptar que, como en la mayoría de las especies de crecimiento lento, composición de bases extrema y genoma pequeño, el UCS no responde a otra cosa que al sesgo sustitucional que sufren estos genomas. Pero las diferencias en el UCS se vieron también asociadas a las medidas del nivel de expresión de los genes, que es indicativo de la presencia de presiones selectivas operando a nivel de la traducción.

En especies como *E. coli* donde existe un tendencia dominante en el uso de codones asociada al nivel de expresión de los genes, el eje 1 obtenido de análisis de correspondencia revela esa tendencia, encontrándose fuertemente correlacionado con el CAI y la distancia sinónima. En este estudio, no se encontró correlación entre CAI y la distancia sinónima, tampoco entre los ejes del WCAo y la distancia sinónima. Cabe destacar que existe amplia variación entre genes en cuanto al dS (m=0.125,sd=0.3), mientras que la variabilidad en el uso de codones, reflejada en el cálculo del porcentaje de variación explicado por los ejes del análisis multivariado es baja. A su vez el CAI estimado para cada gen, mostró un rango acotado de valores (m=0.67,sd=0.04). Como se mencionó en la sección anterior, los organismos con composición de bases extrema, no pueden presentar bajos valores de CAI porque justamente su contenido en G+C define un repertorio limitado de codones que son compartidos por los HEG y el resto de los genes (Botzman & MaMargalit, 2011).

La estimación de MELP permite justamente obtener una medida del nivel de expresión independiente de la composición de bases y el largo de los genes. Los valores de MELP obtenidos para cada especie muestran mayor variación que los valores de CAI y las correlaciones de MELP con el eje 1 del WCAo son más fuertes que las encontradas para CAI.

Al igual que lo observado para CAI, las correlaciones de MELP con dS son bajas (Fig. 2.8) pero puede observarse que los genes con altos valores de MELP (alto nivel de expresión) tienen bajos valores de dS, en cuanto a los genes de bajo nivel de expresión, el dS varía ampliamente.

Cuando se observan frecuencias de codones traduccionalmente óptimos más altas en genes altamente expresados que en el resto de los genes del genoma, es un indicio de que estos codones están siendo favorecidos por la selección (Sharp et al., 2010). Este estudio indica que los codones traduccionalmente óptimos están favorecidos en los HEG, sobre el resto del genoma, pero faltan evidencias para afirmar que estas diferencias se deban a presiones selectivas.

Por un lado tenemos el antecedente de la falta de sesgo en el UCS en genomas de crecimiento lento, la correlación negativa entre la abundancia de ARNt (Pelagibacteriaceae presenta alrededor de 30 genes para ARNt) y el sesgo en el UCS con la tasa de crecimiento celular (Rocha, 2004) sugiriendo que no debería existir selección en la eficiencia de la traducción. Por otro lado, vemos que el sesgo en Pelagibacteriaceae existe a pesar del tamaño y composición de su genoma, y que los HEG usan codones universalmente óptimos a pesar de que estos son codones ricos en G+C en la tercera posición. Podría plantearse entonces que existe selección en favor de la precisión a nivel de la traducción, pero la evidencia presentada no es suficiente. La tasa de sustitución se torna un factor muy importante para contrastar la hipótesis de precisión dado que si fuera cierta, los genes HEG que mostraron preferencia por los codones óptimos deberían mostrar también una tasa de sustitución sinónima menor que refleje la fuerza de la selección purificadora sobre estos sitios. Sin embargo no se encontró el dS correlacionado al nivel de expresión de los genes. Las estimaciones de dS pueden tornarse

complejas cuando la composición de bases es extrema y los modelos tienen que estimar demasiados parámetros. La cantidad de cambio permitido en las secuencias de HEG con respecto a las demás podría estimarse de otra forma, por ejemplo comparando el número de veces que un codón óptimo cambió en una secuencia de alta expresión vs el número de veces que sufrió una sustitución en el resto del genoma. De esta forma obtendríamos una medida de la diferencia de cambio sinónimo entre genes que apoyara la hipótesis de selección sobre la precisión en la traducción.

Afortunadamente existen estudios sobre el sesgo en UCS en un amplio rango de organismos procariotas, que abarca diferentes formas de vida y características genómicas. Esto nos permite comparar nuestros resultados con los obtenidos en casos de organismos que comparten características con Pelagibacteriaceae. A raíz de esto se puede concluir que hay dos características de los organismos que influyen fuertemente sobre la variación en el UCS, estas son el tamaño poblacional efectivo y la forma de vida del organismo, que están, en la mayoría de las especies, fuertemente vinculadas (Foerstner et al., 2005; Rocha & Danchin, 2002).

A diferencia de la mayoría de los endosimbiontes, con los cuales comparte la composición nucleotídica y el pequeño tamaño del genoma, Pelagibacteriaceae, mostró variación intragenómica en el uso de codones, asociada incluso al nivel de expresión de los genes. Esta diferencia puede tener su origen en presiones selectivas sobre Pelagibacteriaceae que aunque débiles, se evidencian gracias a su gran tamaño poblacional efectivo estimado (Charlesworth, 2009; Luo et al., 2014).

Ya se ha planteado que la forma de vida libre influye en la evolución de los genomas en Pelagibacteriaceae, por ejemplo la hipótesis de *genome streamlining* está fuertemente supeditada a esto argumentando que la reducción genómica no puede haber sucedido como en organismos parásitos, debido a presiones selectivas relajadas (Giovannoni et al., 2005c).

Conclusiones generales

Pelagibacteriaceae es una familia peculiar debido a la coexistencia en sus especies de características usualmente no asociadas en otros clados. Los genomas de esta familia son en extremo pobres en G+C y de pequeño tamaño, pero el clado presenta forma de vida libre y un tamaño poblacional efectivo grande. Además a diferencia del enriquecimiento en A+T de endosimbiontes en el cual parece ser muy influyente la falta de mecanismos de reparación, en el clado en estudio estos mecanismos están presentes.

La principal fuente de variación y materia prima para la evolución es la mutación, esto es particularmente válido para la composición nucleotídica de los genomas. Durante mucho tiempo se pensó que la mutación y la selección negativa eran los principales responsables de la composición genómica de organismos procariotas, pero cada vez se torna más evidente que la variabilidad en la composición nucleotídica entre estos organismos no se debe a un sesgo diferencial en el proceso de mutación y posterior selección purificadora. Por el contrario, se ha acumulado evidencia a favor de que procesos como la selección darwiniana o la recombinación, tienen consecuencias sobre el contenido en G+C de procariotas modificando la probabilidad de fijación de las mutaciones.

Cuando estudiamos el patrón de sustituciones en el clado se observa un fuerte desbalance hacia A+T que a medida que pasa el tiempo se vuelve más suave. Observamos también una alta correlación de los patrones de sustitución entre los tipos de sitios (ya sea sinónimos, no sinónimos, o *4fold*) aunque con aspectos singulares de cada uno.

Cuando nos acercamos al patrón mutacional el sesgo es fuerte hacia A+T y en principio la mutación y selección negativa tendrían un rol preponderante, acaso suficiente, en determinar el contenido en G+C de estos genomas. No obstante, existe espacio para que otros eventos selectivos puedan operar modulando suavemente a esas fuerzas.

De acuerdo a estudios realizados en SAR11, este clado tiene adaptaciones importantes a la vida en el ambiente oligotrófico del océano típicamente deficiente en N y P. Esto involucra la eficiente minimización de N en aminoácidos y proteínas, favoreciendo el uso de codones ricos en A+T (Grzymski & DuDussaq, 2012).

Este trabajo mostró que Pelagibacteriaceae usa más frecuentemente codones ricos en A+T, y que en este sentido se pueden encontrar diferencias entre genes. Pero para proponer que estas diferencias sean selectivas faltan evidencias porque no se encontró la clásica relación entre el CAI y la distancia sinónima de los genes. De acuerdo a los resultados, la selección (no se encontró evidencia de recombinación como proceso afectando el contenido en G+C) actúa removiendo las mutaciones ricas en A+T a nivel no sinónimo. Entonces si no existe presión selectiva sobre las terceras posiciones cómo podríamos detectar una correlación de dS con CAI?

Sin embargo, el clado mostró preferencia en genes de alta expresión por los codones UUC (Phe), AUC (Ile),

UAC (Tyr) y AAC (Asn), a los que se suman los codones GCU (Ala), GAC (Asp), GGU (Gly), CAC (His), AUU (Ile), CUU (Leu), CCU (Pro), AGA (Arg), UCA (Ser) y GUU (Val), muchos de ellos ricos en G+C en la tercera posición, por lo tanto los sitios sinónimos no deberían ser tan permisivos, por lo menos no en genes HEG. La cantidad de cambio permitido en las secuencias de HEG con respecto a las demás podría estimarse por ejemplo comparando el número de veces que un codón óptimo cambió en una secuencia de alta expresión *versus* el número de veces que sufrió una sustitución en el resto del genoma. De esta forma obtendríamos una medida de la diferencia de cambio sinónimo entre genes que apoyara la hipótesis de selección sobre la precisión en la traducción.

De acuerdo a los resultados obtenidos, el tamaño poblacional es una característica clave en la evolución de la composición de bases de una especie, y dado que los tamaños poblacionales varían con los estilos de vida de los organismos, se puede decir que la forma de vida, parásita o libre, tiene influencia en este carácter.

A su vez, el hallazgo de indicios de selección operando a nivel de la traducción en Pelagibacteriaceae, es peculiar dada la composición de bases, y es otro indicador de que el tamaño poblacional permite que opere la selección manteniendo codones óptimos en genes de alta expresión.

Bibliografía

Akashi, H., & Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proc. Natl. Acad. Sci. U. S. A. *99*, 3695–3700.

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

Andersson, S.G.E., and Sharp, P.M. (1996). Codon usage and base composition inRickettsia prowazekii. J. Mol. Evol. *42*, 525–536.

Andersson, S.G., and Kurland, C.G. (1998). Reductive evolution of resident genomes. Trends Microbiol. *6*, 263–268.

Arenas, M., and Posada, D. (2010a). The effect of recombination on the reconstruction of ancestral sequences. Genetics *184*, 1133–1139.

Arenas, M., and Posada, D. (2010b). The effect of recombination on the reconstruction of ancestral sequences. Genetics *184*, 1133–1139.

Belozersky, A.N., and Spirin, A.S. (1958). A Correlation between the Compositions of Deoxyribonucleic and Ribonucleic Acids. Nature *182*, 111–112.

Bentley, S.D., and Parkhill, J. (2004). Comparative genomic structure of prokaryotes. Annu. Rev. Genet. *38*, 771–792.

Bininda-Emonds, O.R.P. (2005). transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. BMC Bioinformatics *6*, 156.

Blanchette, M., Diallo, A.B., Green, E.D., Miller, W., and Haussler, D. (2008). Computational reconstruction of ancestral DNA sequences. Methods Mol. Biol. Clifton NJ *422*, 171–184.

Botzman, M., and Margalit, H. (2011). Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. Genome Biol. *12*, R109.

Boussau, B., Karlberg, E.O., Frank, A.C., Legault, B.-A., and Andersson, S.G.E. (2004). Computational inference of scenarios for α-proteobacterial genome evolution. Proc. Natl. Acad. Sci. U. S. A. *101*, 9722–9727.

Brown, M.V., Lauro, F.M., DeMaere, M.Z., Muir, L., Wilkins, D., Thomas, T., Riddle, M.J., Fuhrman, J.A., Andrews-Pfannkoch, C., Hoffman, J.M., et al. (2012). Global biogeography of SAR11 marine bacteria. Mol. Syst. Biol. *8*, 595.

Bruen, T.C., Philippe, H., and Bryant, D. (2006). A Simple and Robust Statistical Test for Detecting the

Presence of Recombination. Genetics *172*, 2665–2681.

Carini, P., Steindler, L., Beszteri, S., and Giovannoni, S.J. (2013). Nutrient requirements for growth of the extreme oligotroph “Candidatus Pelagibacter ubique” HTCC1062 on a defined medium. ISME J. *7*, 592–602.

Carlson, C.A., Morris, R., Parsons, R., Treusch, A.H., Giovannoni, S.J., and Vergin, K. (2009). Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. ISME J. *3*, 283–295.

C E McEwan, D.G. (1998). Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. Hereditas *128*, 173–178.

Charif, D., Thioulouse, J., Lobry, J.R., and Perrière, G. (2005). Online synonymous codon usage analyses with the ade4 and seqinR packages. Bioinforma. Oxf. Engl. *21*, 545–547.

Charlesworth, B. (2009). Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat. Rev. Genet. *10*, 195–205.

Chen, S.L., Lee, W., Hottes, A.K., Shapiro, L., and McAdams, H.H. (2004). Codon usage between genomes is constrained by genome-wide mutational processes. Proc. Natl. Acad. Sci. U. S. A. *101*, 3480–3485.

Corander, J., Waldmann, P., and Sillanpää, M.J. (2003). Bayesian analysis of genetic differentiation between populations. Genetics *163*, 367–374.

Crick, F.H. (1966). Codon--anticodon pairing: the wobble hypothesis. J. Mol. Biol. *19*, 548–555.

Daubin, V., and Perrière, G. (2003). G+C3 structuring along the genome: a common feature in prokaryotes. Mol. Biol. Evol. *20*, 471–483.

Duret, L., and Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. Annu. Rev. Genomics Hum. Genet. *10*, 285–311.

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

Eiler, A., Hayakawa, D.H., Church, M.J., Karl, D.M., and Rappé, M.S. (2009). Dynamics of the SAR11 bacterioplankton lineage in relation to environmental conditions in the oligotrophic North Pacific subtropical gyre. Environ. Microbiol. *11*, 2291–2300.

Ermolaeva, M.D. (2001). Synonymous codon usage in bacteria. Curr. Issues Mol. Biol. *3*, 91–97.

Fitch, W.M. (1970). Distinguishing homologous from analogous proteins. Syst. Zool. *19*, 99–113.

Foerstner, K.U., von Mering, C., Hooper, S.D., and Bork, P. (2005). Environments shape the nucleotide composition of genomes. EMBO Rep. *6*, 1208–1213.

Fourment, M., and Gibbs, M.J. (2006). PATRISTIC: a program for calculating patristic distances and graphically comparing the components of genetic change. BMC Evol. Biol. *6*, 1.

Giovannoni, S.J., Britschgi, T.B., Moyer, C.L., and Field, K.G. (1990). Genetic diversity in Sargasso Sea bacterioplankton. Nature *345*, 60–63.

Giovannoni, S.J., Bibbs, L., Cho, J.-C., Stapels, M.D., Desiderio, R., Vergin, K.L., Rappé, M.S., Laney, S., Wilhelm, L.J., Tripp, H.J., et al. (2005a). Proteorhodopsin in the ubiquitous marine bacterium SAR11. Nature *438*, 82–85.

Giovannoni, S.J., Tripp, H.J., Givan, S., Podar, M., Vergin, K.L., Baptista, D., Bibbs, L., Eads, J., Richardson, T.H., Noordewier, M., et al. (2005b). Genome Streamlining in a Cosmopolitan Oceanic Bacterium. Science *309*, 1242–1245.

Giovannoni, S.J., Tripp, H.J., Givan, S., Podar, M., Vergin, K.L., Baptista, D., Bibbs, L., Eads, J., Richardson, T.H., Noordewier, M., et al. (2005c). Genome streamlining in a cosmopolitan oceanic bacterium. Science *309*, 1242–1245.

Goldman, N., and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. *11*, 725–736.

Gouy, M., and Gautier, C. (1982). Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. *10*, 7055–7074.

Grantham, R., Gautier, C., Gouy, M., Mercier, R., and Pavé, A. (1980). Codon catalog usage and the genome hypothesis. Nucleic Acids Res. *8*, r49–r62.

Grote, J., Thrash, J.C., Huggett, M.J., Landry, Z.C., Carini, P., Giovannoni, S.J., and Rappé, M.S. (2012). Streamlining and core genome conservation among highly divergent members of the SAR11 clade. mBio *3*.

Grzymski, J.J., and Dussaq, A.M. (2012). The significance of nitrogen cost minimization in proteomes of marine microorganisms. ISME J. *6*, 71–80.

Guindon, S., and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. *52*, 696–704.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307–321.

Hanson-Smith, V., Kolaczkowski, B., and Thornton, J.W. (2010). Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. Mol. Biol. Evol. *27*, 1988–1999.

Hasegawa, M., Kishino, H., and Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. *22*, 160–174.

Heddi, A., Charles, H., Khatchadourian, C., Bonnot, G., and Nardon, P. (1998). Molecular characterization of the principal symbiotic bacteria of the weevil Sitophilus oryzae: a peculiar G + C content of an endocytobiotic DNA. J. Mol. Evol. *47*, 52–61.

Hershberg, R., and Petrov, D.A. (2008). Selection on codon bias. Annu. Rev. Genet. *42*, 287–299.

Hershberg, R., and Petrov, D.A. (2010a). Evidence that mutation is universally biased towards AT in

bacteria. PLoS Genet. *6*.

Hershberg, R., and Petrov, D.A. (2010b). Evidence That Mutation Is Universally Biased towards AT in Bacteria. PLoS Genet. *6*.

Hershberg, R., and Petrov, D.A. (2012). On the limitations of using ribosomal genes as references for the study of codon usage: a rebuttal. PloS One *7*, e49060.

Higgs, P.G., and Ran, W. (2008). Coevolution of Codon Usage and tRNA Genes Leads to Alternative Stable States of Biased Codon Usage. Mol. Biol. Evol. *25*, 2279–2291.

Hildebrand, F., Meyer, A., and Eyre-Walker, A. (2010). Evidence of selection upon genomic GCcontent in bacteria. PLoS Genet. *6*.

Holder, M., and Lewis, P.O. (2003). Phylogeny estimation: traditional and Bayesian approaches. Nat. Rev. Genet. *4*, 275–284.

Huson, D.H., and Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. *23*, 254–267.

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. *2*, 13–34.

Iriarte, A., Baraibar, J.D., Romero, H., Castro-Sowinski, S., and Musto, H. (2013). Evolution of optimal codon choices in the family Enterobacteriaceae. Microbiol. Read. Engl. *159*, 555–564.

Kimura, M. (1977). Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature *267*, 275–276.

Lassalle, F., Périan, S., Bataillon, T., Nesme, X., Duret, L., and Daubin, V. (2015). GC-Content Evolution in Bacterial Genomes: The Biased Gene Conversion Hypothesis Expands. PLoS Genet *11*, e1004941.

Legendre, P., and Legendre, L.F.J. (1998). Numerical Ecology (Elsevier).

Lightfield, J., Fram, N.R., and Ely, B. (2011). Across Bacterial Phyla, Distantly-Related Genomes with Similar Genomic GC Content Have Similar Patterns of Amino Acid Usage. PLoS ONE *6*, e17677.

Lind, P.A., and Andersson, D.I. (2008). Whole-genome mutational biases in bacteria. Proc. Natl. Acad. Sci. U. S. A. *105*, 17878–17883.

Luo, H., Swan, B.K., Stepanauskas, R., Hughes, A.L., and Moran, M.A. (2014). Comparing effective population sizes of dominant marine alphaproteobacteria lineages. Environ. Microbiol. Rep. *6*, 167–172.

Martin, D.P., Lemey, P., and Posada, D. (2011). Analysing recombination in nucleotide sequences. Mol. Ecol. Resour. *11*, 943–955.

McCutcheon, J.P., McDonald, B.R., and Moran, N.A. (2009). Origin of an Alternative Genetic Code in the Extremely Small and GC–Rich Genome of a Bacterial Symbiont. PLoS Genet *5*, e1000565.

McDonald, J.H. (2006). Apparent Trends of Amino Acid Gain and Loss in Protein Evolution Due to Nearly Neutral Variation. Mol. Biol. Evol. *23*, 240–244.

Miyata, T., and Yasunaga, T. (1980). Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J. Mol. Evol. *16*, 23–36.

Morris, J.J., Lenski, R.E., and Zinser, E.R. (2012). The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss. mBio *3*, e00036–12.

Morris, R.M., Rappé, M.S., Connon, S.A., Vergin, K.L., Siebold, W.A., Carlson, C.A., and Giovannoni, S.J. (2002). SAR11 clade dominates ocean surface bacterioplankton communities. Nature *420*, 806–810.

Mugal, C.F., Wolf, J.B.W., and Kaj, I. (2014). Why Time Matters: Codon Evolution and the Temporal Dynamics of dN/dS. Mol. Biol. Evol. *31*, 212–231.

Musto, H., Romero, H., and Zavala, A. (2003). Translational selection is operative for synonymous codon usage in Clostridium perfringens and Clostridium acetobutylicum. Microbiol. Read. Engl. *149*, 855–863.

Musto, H., Naya, H., Zavala, A., Romero, H., Alvarez-Valín, F., and Bernardi, G. (2004). Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. FEBS Lett. *573*, 73–77.

Muto, A., and Osawa, S. (1987). The guanine and cytosine content of genomic DNA and bacterial evolution. Proc. Natl. Acad. Sci. U. S. A. *84*, 166–169.

Naya, H., Romero, H., Zavala, A., Alvarez, B., and Musto, H. (2002). Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. J. Mol. Evol. *55*, 260–264.

Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. *3*, 418–426.

Ogle, J.M., and Ramakrishnan, V. (2005). Structural insights into translational fidelity. Annu. Rev. Biochem. *74*, 129–177.

Ohama, T., Muto, A., and Osawa, S. (1990). Role of GC-biased mutation pressure on synonymous codon choice in Micrococcus luteus a bacterium with a high genomic GC-content. Nucleic Acids Res. *18*, 1565–1569.

Ohkubo, S., Muto, A., Kawauchi, Y., Yamao, F., and Osawa, S. (1987). The ribosomal protein gene cluster of Mycoplasma capricolum. Mol. Gen. Genet. MGG *210*, 314–322.

Ohta, T. (2000). Mechanisms of molecular evolution. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *355*, 1623–1626.

Ottesen, E.A., Young, C.R., Eppley, J.M., Ryan, J.P., Chavez, F.P., Scholin, C.A., and DeLong, E.F. (2013). Pattern and synchrony of gene expression among sympatric marine microbial populations. Proc. Natl. Acad. Sci. U. S. A. *110*, E488–E497.

Plotkin, J.B., and Kudla, G. (2011). Synonymous but not the same: the causes and consequences of

codon bias. Nat. Rev. Genet. *12*, 32–42.

Poptsova, M.S., and Gogarten, J.P. (2007). BranchClust: a phylogenetic algorithm for selecting gene families. BMC Bioinformatics *8*, 120.

Ran, W., and Higgs, P.G. (2012). Contributions of speed and accuracy to translational selection in bacteria. PloS One *7*, e51652.

Ratnakumar, A., Mousset, S., Glémin, S., Berglund, J., Galtier, N., Duret, L., and Webster, M.T. (2010). Detecting positive selection within genomes: the problem of biased gene conversion. Philos. Trans. R. Soc. Lond. B Biol. Sci. *365*, 2571–2580.

Retchless, A.C., and Lawrence, J.G. (2012). Ecological adaptation in bacteria: speciation driven by codon selection. Mol. Biol. Evol. *29*, 3669–3683.

Rocha, E.P. (2008). Evolutionary patterns in prokaryotic genomes. Curr. Opin. Microbiol. *11*, 454–460.

Rocha, E.P.C. (2004). Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res. *14*, 2279–2286.

Rocha, E.P.C., and Danchin, A. (2002). Base composition bias might result from competition for metabolic resources. Trends Genet. TIG *18*, 291–294.

Rocha, E.P.C., and Feil, E.J. (2010). Mutational patterns cannot explain genome composition: Are there any neutral sites in the genomes of bacteria? PLoS Genet. *6*.

Rocha, E.P.C., Smith, J.M., Hurst, L.D., Holden, M.T.G., Cooper, J.E., Smith, N.H., and Feil, E.J. (2006). Comparisons of dN/dS are time dependent for closely related bacterial genomes. J. Theor. Biol. *239*, 226–235.

Rodríguez-Ezpeleta, N., and Embley, T.M. (2012). The SAR11 group of alpha-proteobacteria is not related to the origin of mitochondria. PloS One *7*, e30520.

Roller, M., Lucić, V., Nagy, I., Perica, T., and Vlahovicek, K. (2013). Environmental shaping of codon usage and functional adaptation across microbial communities. Nucleic Acids Res. *41*, 8842–8852.

Romero, H., Zavala, A., and Musto, H. (2000). Codon usage in Chlamydia trachomatis is the result of strand-specific mutational biases and a complex pattern of selective forces. Nucleic Acids Res. *28*, 2084–2090.

Seligmann, H. (2003). Cost-minimization of amino acid usage. J. Mol. Evol. *56*, 151–161.

Sharp, P.M., Bailes, E., Grocock, R.J., Peden, J.F., and Sockett, R.E. (2005). Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. *33*, 1141–1153.

Sharp, P.M., Emery, L.R., and Zeng, K. (2010). Forces that influence the evolution of codon bias. Philos. Trans. R. Soc. Lond. B. Biol. Sci. *365*, 1203–1212.

Shimodaira, H., and Hasegawa, M. (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinforma. Oxf. Engl. *17*, 1246–1247.

Singer, C.E., and Ames, B.N. (1970). Sunlight ultraviolet and bacterial DNA base ratios. Science *170*, 822–825.

Sowell, S.M., Wilhelm, L.J., Norbeck, A.D., Lipton, M.S., Nicora, C.D., Barofsky, D.F., Carlson, C.A., Smith, R.D., and Giovanonni, S.J. (2009). Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. ISME J. *3*, 93–105.

Sueoka, N. (1961). Variation and heterogeneity of base composition of deoxyribonucleic acids: A compilation of old and new data. J. Mol. Biol. *3*, 31–IN15.

Sun, J., Steindler, L., Thrash, J.C., Halsey, K.H., Smith, D.P., Carter, A.E., Landry, Z.C., and Giovannoni, S.J. (2011). One carbon metabolism in SAR11 pelagic marine bacteria. PloS One *6*, e23973.

Suzuki, H., Brown, C.J., Forney, L.J., and Top, E.M. (2008). Comparison of correspondence analysis methods for synonymous codon usage in bacteria. DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes *15*, 357–365.

Teira, E., Martínez-García, S., Lønborg, C., and Alvarez-Salgado, X.A. (2009). Growth rates of different phylogenetic bacterioplankton groups in a coastal upwelling system. Environ. Microbiol. Rep. *1*, 545–554.

Thrash, J.C., Boyd, A., Huggett, M.J., Grote, J., Carini, P., Yoder, R.J., Robbertse, B., Spatafora, J.W.,

Rappé, M.S., and Giovannoni, S.J. (2011). Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. Sci. Rep. *1*, 13.

Toft, C., and Andersson, S.G.E. (2010). Evolutionary microbial genomics: insights into bacterial host adaptation. Nat. Rev. Genet. *11*, 465–475.

Travers, A.A. (2004). The structural basis of DNA flexibility. Philos. Transact. A Math. Phys. Eng. Sci. *362*, 1423–1438.

Tripp, H.J. (2013). The unique metabolism of SAR11 aquatic bacteria. J. Microbiol. Seoul Korea *51*, 147–153.

Tripp, H.J., Kitner, J.B., Schwalbach, M.S., Dacey, J.W.H., Wilhelm, L.J., and Giovannoni, S.J. (2008). SAR11 marine bacteria require exogenous reduced sulphur for growth. Nature *452*, 741–744.

Tripp, H.J., Schwalbach, M.S., Meyer, M.M., Kitner, J.B., Breaker, R.R., and Giovannoni, S.J. (2009). Unique glycine-activated riboswitch linked to glycine-serine auxotrophy in SAR11. Environ. Microbiol. *11*, 230–238.

Vergin, K.L., Tripp, H.J., Wilhelm, L.J., Denver, D.R., Rappé, M.S., and Giovannoni, S.J. (2007). High intraspecific recombination rate in a native population of Candidatus pelagibacter ubique (SAR11). Environ. Microbiol. *9*, 2430–2440.

Vergin, K.L., Beszteri, B., Monier, A., Thrash, J.C., Temperton, B., Treusch, A.H., Kilpert, F., Worden, A.Z., and Giovannoni, S.J. (2013). High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. ISME J. *7*, 1322–1332.

Viklund, J., Ettema, T.J.G., and Andersson, S.G.E. (2012). Independent genome reduction and phylogenetic reclassification of the oceanic SAR11 clade. Mol. Biol. Evol. *29*, 599–615.

Vos, M., and Didelot, X. (2009). A comparison of homologous recombination rates in bacteria and archaea. ISME J. *3*, 199–208.
Ward, N., and Moreno-Hagelsieb, G. (2014). Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss? PloS One *9*, e101850.

Whelan, S., Liò, P., and Goldman, N. (2001). Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet. TIG *17*, 262–272.

Xia, X., Xie, Z., and Li, W.-H. (2003). Effects of GC content and mutational pressure on the lengths of exons and coding sequences. J. Mol. Evol. *56*, 362–370.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

Yang, null, and Bielawski, null (2000). Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. *15*, 496–503.

Yang, Z., and Nielsen, R. (2000). Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. Mol. Biol. Evol. *17*, 32–43.

Yang, Z., and Rannala, B. (2012). Molecular phylogenetics: principles and practice. Nat. Rev. Genet. *13*, 303–314.