CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS AVANZADOS DEL INSTITUTO POLITÉCNICO NACIONAL

UNIDAD IRAPUATO

Evolución de la eficiencia traduccional debido a la selección del uso de codones en procariotas

Tesis que presenta:

Francisco Maximiliano González Serrano

para obtener el grado de

Maestro en Biología Integrativa

Director de Tesis:

Luis José Delaye Arredondo

Irapuato, Guanajuato **[Agosto 2020]**

El presente trabajo de tesis "**Evolución de la eficiencia traduccional debido a la selección del uso de codones en procariotas**" fue llevado a cabo en el laboratorio de Genómica Evolutiva del Departamento de Ingeniería Genética del CINVESTAV-Unidad Irapuato, bajo la dirección del Dr. Luis José Delaye Arredondo.

Francisco Maximiliano González Serrano (CVU: 856429) agradece el apoyo obtenido por parte de CONACYT.

Agradecimientos

A mi asesor de tesis, el Dr. Luis José Delaye Arredondo, por darme la oportunidad de formar parte de su grupo, apoyarme y guiarme en mi proyecto y permitirme involucrarme en otros proyectos para seguir aprendiendo.

Al Dr. Cei Gastón Abreu-Goodger, por su asesoría, valiosa retroalimentación y disponibilidad de tiempo.

Al Dr. Octavio Martínez de la Vega por las valiosas contribuciones y sugerencias al proyecto.

Al LANGEBIO y Araceli Fernández Cortés por brindarme soporte de software y auxiliarme con respecto al uso de MAZORKA.

Al Consejo Nacional de Ciencia y Tecnología (CONACyT) por brindarme el apoyo económico necesario para continuar mis estudios en México.

A mi familia, en particular a mis padres, por sus cimientos y aportes invaluables que servirán para toda mi vida.

Y a quienes me apoyaron de alguna forma durante esta etapa de mi vida con su compañía, paciencia, consejos y amistad, por mencionar algunos, a Eduardo González, Gonzalo Córdova, María González y Carlos Barradas.

Dedicatoria

A Toyo

enero 2009 – marzo 2018

“Dogs are not our whole life, but they make our lives whole”
Roger Andrew Caras

y a mis padres

“El mejor legado de un padre a sus hijos es un poco de su tiempo cada día”

Leon Battista Alberti

Índice

Resumen

La tasa de elongación de la traducción es influenciada por el sesgo en el uso de codones bajo selección (CUB_s, por sus siglas en inglés *codon usage bias under selection*). Esto se explica porque la selección natural favorece que el sesgo en el uso de codones refleje la uniformidad en la disponibilidad de los distintos ARN de transferencia (ARNt) en el citoplasma. Se ha observado que los procariontes que presentan CUB_s en una porción importante de sus genes tienden a mostrar tiempos generacionales más cortos que aquellos que no presentan CUB_s. También se ha sugerido que estos pueden adaptarse a ambientes más diversos. No obstante, la relación entre el CUB_s y los estilos de vida de los organismos sigue siendo tema de debate. No es del todo claro cuál es el objeto de la selección. Además, estudios recientes que han tratado de cuantificar el CUB_s, no han considerado adecuadamente el contenido de G + C o han utilizado a los genes codificantes para proteínas ribosomales (PR) como estándares en la cuantificación del CUB_s sin verificar que dicho criterio aplique siempre. En este trabajo exploramos el fenómeno del CUB_s desde una perspectiva evolutiva tomando en cuenta las consideraciones mencionadas. Nuestro objetivo fue entender en cuales estilos de vida se ve favorecido el CUB_s y cómo se relaciona esto con el tiempo mínimo de duplicación. Contrariamente a lo esperado, la regresión filogenética mostró una asociación relativamente débil entre los valores CUB_s del genoma y los tiempos de duplicación ($r = 0.4$ p-valor < 0.01). Se observó que el CUB_s de los genes codificantes de proteínas ribosomales (PR) no siempre es el más elevado con respecto al resto de los genes en los genomas donde parece haber una co-adaptación entre el uso de codones y el sesgo de los codones. Sin embargo, los valores de CUB_s en PR correlacionaron mejor con los tiempos de duplicación que los valores del CUB_s del genoma completo. La exploración de las funciones génicas y sus valores de CUB_s revelaron que las funciones vinculadas a la traducción, la expresión de genes y el metabolismo de los carbohidratos poseen valores altos de CUB_s, principalmente en los organismos de crecimiento rápido. Inclusive, el CUB_s de estas funciones parece estar asociado con tiempos de duplicación cortos. Algunas funciones específicas también exhibieron altos valores de CUB_s como son la fotosíntesis en cianobacterias, la metanogénesis en arqueas metanógenas, y el metabolismo de antibióticos en organismos productores y resistentes a antibióticos tales como Actinobacterias, Pseudomonadaceae y Bacillales. Interpretamos nuestros resultados para sugerir que el tiempo de duplicación no es la característica que la selección

natural está modificando cuando se favorece el CUB, sino un efecto secundario de la selección en otra. El CUB_s está presente en algunos organismos para mantener funciones fundamentales relacionadas con sus estilos de vida. Sin embargo, el CUB_s parece tener una mayor presencia en la maquinaria de traducción, posiblemente debido a su impacto global en la optimización de la traducción del genoma. Finalmente, con la información generada, propusimos un modelo para explicar cómo el CUB_s evoluciona (el *modus operandi*): en nuestro modelo el CUB_s evoluciona primero en los genes que codifican para la maquinaria de traducción y posteriormente en otros procesos biológicos.

Abstract

Translation elongation rate in Prokaryotes is influenced by natural selection on codon usage bias (CUB_s). This is explained by the co-adaptation between the codon usage and the availability of cognate transfer RNAs (tRNAs) in the cytoplasm. It has been observed that cells showing a strong CUB_s tend to show short generation times. Furthermore, it has been suggested that species living in diverse environments tend to have genomes with strong CUB_s. The implications of CUB_s on organism lifestyles, nevertheless, are still under debate. Trying to unveil them recent studies, while quantifying CUB_s, have not properly considered the nucleotide background or have taken ribosomal protein genes (RP) as gold standards, without verifying if this criterion always applies. Here, we explored the CUB_s phenomenon at the genome and gene-function levels, with an ecological and evolutionary perspective, given the aforementioned considerations. We aim to understand in which lifestyles and biological process CUB_s is favored and how it is related to generation time. Contrary to expectations, phylogenetic regression showed a weak correlation between genome CUB_s values and generation time ($r = 0.4$ p-valor < 0.01). Looking at CUBs on RP demonstrated that these proteins not always have the greatest values in all genomes, nevertheless, RP's CUB_s values correlated better with generation times than CUB_s genomes values. Examination of the genetic functions and their CUB_s values reveal that the functions linked to translation, gene expression, and carbohydrate metabolism have high CUBs values, mainly in fast-growing organisms. Furthermore, the CUB_s on these functions seem to be associated with short generation times. Some specific functions also exhibited high values of CUB_s, such as photosynthesis in cyanobacteria, methanogenesis in methanogenic archaea, and metabolism of antibiotics in antibiotic-producing and resistant organisms as Actinobacteria, Pseudomonadaceae y Bacillales. Our results suggest that generation time is not a determinant factor in leading CUB selection but a secondary effect in fast-growing organisms. Translation elongation efficiency by CUB is under selection in some organisms to maintain fundamental functions linked with their lifestyles. CUB_s appear to be more present on translation machinery, likely because of its impact on optimizing genome translation globally. Additionally, with these data, we propose a model to explain how CUB_s evolves *(modus operandi*) where CUB_s change first in the translation machinery genes and eventually in genes involved in other biological processes.

Introducción

El código genético

Posteriormente al descubrimiento de la doble hélice en 1953, George Gamow empezó a formular lo que hoy conocemos como el código genético en donde combinaciones de cuatro pares de bases conformarían los 20 principales aminoácidos. Más tarde, esta propuesta fue descartada con evidencia experimental usando una secuencia de ARN poli-uracilo por Marshall Nirenberg y Heinrich J. Matthaei en 1961. Con dicho experimento observaron que por cada triplete de nucleótidos se sintetizaba un aminoácido. Dos años más tarde Nirenberg y Leader en 1964 descubrieron la correspondencia para cada uno de los tripletes dando lugar a la base del código genético como lo es hoy en día.

Actualmente, sabemos que el código genético consiste en un grupo de interacciones moleculares usadas por todas las células para traducir la información codificada en secuencias de nucleótidos en proteínas. El principio de estas interacciones recae en la asignación de tripletes de las cuatro bases nitrogenadas (Adenina, Timina/Uracilo, Citosina y Guanina), llamados codones, en los principales 20 aminoácidos que conforman a las proteínas. Dicho código tiene diversas características, una de ellas es que es degenerado, es decir, un mismo aminoácido puede ser asignado a más de un codón. Estos últimos son nombrados codones sinónimos (Figura 1).

Codón	Aminoácido			Codón	Aminoácido			Codón	Aminoácido			Codón	Aminoácido		
TTT	*F*	*Phe*	Fenilalanina	**TCT**	*S*	*Ser*	Serina	**TAT**	*Y*	*Tyr*	Tirosina	**TGT**	*C*	*Cys*	Cisteína
TTC	*F*	*Phe*	Fenilalanina	**TCC**	*S*	*Ser*	Serina	**TAC**	*Y*	*Tyr*	Tirosina	**TGC**	*C*	*Cys*	Cisteína
TTA	*L*	*Leu*	Leucina	**TCA**	*S*	*Ser*	Serina	**TAA**	*	*Ter*	Terminación	**TGA**	*	*Ter*	Terminación
TTG	*L*	*Leu*	Leucina	**TCG**	*S*	*Ser*	Serina	**TAG**	*	*Ter*	Terminación	**TGG**	*W*	*Trp*	Triptófano
CTT	*L*	*Leu*	Leucina	**CCT**	*P*	*Pro*	Prolina	**CAT**	*H*	*His*	Histidina	**CGT**	*R*	*Arg*	Arginina
CTC	*L*	*Leu*	Leucina	**CCC**	*P*	*Pro*	Prolina	**CAC**	*H*	*His*	Histidina	**CGC**	*R*	*Arg*	Arginina
CTA	*L*	*Leu*	Leucina	**CCA**	*P*	*Pro*	Prolina	**CAA**	*Q*	*Gln*	Glutamina	**CGA**	*R*	*Arg*	Arginina
CTG	*L*	*Leu*	Leucina	**CCG**	*P*	*Pro*	Prolina	**CAG**	*Q*	*Gln*	Glutamina	**CGG**	*R*	*Arg*	Arginina
ATT	*I*	*Ile*	Isoleucina	**ACT**	*T*	*Thr*	Treonina	**AAT**	*N*	*Asn*	Asparagina	**AGT**	*S*	*Ser*	Serina
ATC	*I*	*Ile*	Isoleucina	**ACC**	*T*	*Thr*	Treonina	**AAC**	*N*	*Asn*	Asparagina	**AGC**	*S*	*Ser*	Serina
ATA	*I*	*Ile*	Isoleucina	**ACA**	*T*	*Thr*	Treonina	**AAA**	*K*	*Lys*	Lisina	**AGA**	*R*	*Arg*	Arginina
ATG	*M*	*Met*	Metionina	**ACG**	*T*	*Thr*	Treonina	**AAG**	*K*	*Lys*	Lisina	**AGG**	*R*	*Arg*	Arginina
GTT	*V*	*Val*	Valina	**GCT**	*A*	*Ala*	Alanina	**GAT**	*D*	*Asp*	Aspartato	**GGT**	*G*	*Gly*	Glicina
GTC	*V*	*Val*	Valina	**GCC**	*A*	*Ala*	Alanina	**GAC**	*D*	*Asp*	Aspartato	**GGC**	*G*	*Gly*	Glicina
GTA	*V*	*Val*	Valina	**GCA**	*A*	*Ala*	Alanina	**GAA**	*E*	*Glu*	Glutamato	**GGA**	*G*	*Gly*	Glicina
GTG	*V*	*Val*	Valina	**GCG**	*A*	*Ala*	Alanina	**GAG**	*E*	*Glu*	Glutamato	**GGG**	*G*	*Gly*	Glicina

Figura 1. Código genético universal. 18 de los 20 aminoácidos son codificados por más de un codón. Tres codones son específicos para concluir con la síntesis de proteínas y dos codones codifican sólo a un aminoácido (W y M).

El sesgo en el uso de codones (CUB)

La no homogeneidad del uso de codones sinónimos entre genes de un mismo organismo y entre diferentes organismos, comenzó a estudiarse a partir de los 80s por Grantham e Ikemura. Ellos estudiaron unos pocos genes (90 genes en Grantham *et al.*, 1980 y 161 en Grantham *et al.*, 1981), disponibles en aquella época, de organismos eucariontes, procariontes y fagos. Sorpresivamente, encontraron que cada organismo presentaba su propio uso de codones. Los estudios de Ikemura mostraron resultados similares y también una posible correspondencia entre los ARN de transferencia (ARNt) de las células y el uso de los codones, lo cual lo llevó a pensar en una posible explicación para este fenómeno descrito como el sesgo en el uso de codones (CUB, por sus siglas en inglés de *codon usage bias*) (Plotkin & Kudla, 2011).

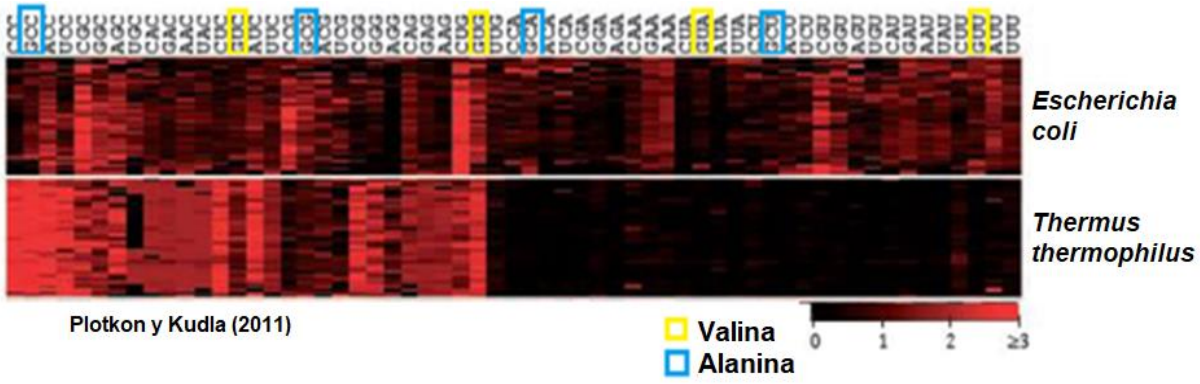

Figura 2. Uso de codones no homogéneo entre genes y genomas. Se muestra el uso de codones sinónimos relativos (RSCU, por sus siglas en inglés) de 50 genes aleatorios de dos organismos: *E. coli* y *T. thermophilus*. Los colores reflejan valores que van desde el 0 (indicando que un codón sinónimo determinado no es usado en lo absoluto); pasando por 1(indicando que un codón sinónimo es usado en igual frecuencia que cualquiera de los otros codones sinónimos); y llegando hasta >=3 y <=6 (indicando que solo uno de los codones sinónimos es usado para codificar a un aminoácido). Los genes de cada organismo se muestran en *phyla* mientras que los codones en columnas (Los codones codificantes para metionina y triptófano son omitidos, así como los codones de paro). Los codones están ordenados según el nucleótido en la tercera posición: C, G, A y U. Para ejemplificar se resaltan dos aminoácidos, Valina en amarillo y Alanina en azul.

Las razones que rigen la existencia del CUB han cautivado el interés científico. La teoría de la selección-mutación-deriva propuesta por Bulmer (1991) es la más aceptada a la fecha. En esta teoría, el uso de codones resulta del balance en una población finita entre la selección por eficiencia traduccional (favoreciendo codones óptimos para cada aminoácido) y la mutación acompañada de la deriva génica (la cual se rige por las frecuencias alélicas iniciales y procesos aleatorios) que permite la persistencia de codones no óptimos.

Eficiencia traduccional y el sesgo en el uso de codones bajo selección (CUB_s)

Desde una perspectiva minimalista, la eficiencia del proceso de traducción está determinada por dos tasas: i) la del inicio de la traducción y ii) la de la elongación. La primera, cuya importancia es mayor, afecta la frecuencia con la que un transcrito es traducido; mientras que la segunda se refiere a la tasa de traducción en sí (Quax *et al*., 2015). Se ha observado que el CUB_s puede tener un impacto en la tasa de inicio de la traducción debido a posibles conformaciones estructurales más estables en la región 5' que facilitan el reconocimiento de las secuencias Shine-Dalgarno (Bhattacharyya *et al*., 2018). Sin embargo, se ha planteado que el mayor aporte del CUB_s ocurre durante la elongación. Este efecto es más notorio en los genes que tienen altos niveles de expresión (es decir, de transcripción) y, en algunos casos, en genes largos (Salim, H. M., & Cavalcanti, A. R., 2008). En estos escenarios los codones sinónimos más frecuentes corresponden a los tipos de ARNt más abundantes en el citoplasma de la célula. De este modo, la probabilidad de que un codón determinado sea reconocido por su ARNt afín aumenta, dando lugar a una síntesis de proteínas más rápida (Quax *et al*., 2015). Que el efecto sea más notorio en genes altamente expresados se explica por el gasto energético. Para una célula es conveniente que un gen que se transcribe rápido se traduzca rápido y así evitar el acumulamiento de sus transcritos que conlleva a su degradación antes de ser traducidos. Por otro lado, en los genes largos el sesgo aumenta a medida que incrementa la longitud del gen. Esto se explica porque a medida que un gen se hace más largo se requiere más energía para la traducción. Entonces cada vez es más importante evitar errores sin sentido en el extremo 3' que terminarían la traducción prematuramente sintetizando péptidos no funcionales (Salim, H. M., & Cavalcanti, A. R., 2008),

Este efecto no sólo impacta en la traducción del gen en particular, sino indirectamente en la del resto de los genes. Esto se debe a que se liberan a mayor velocidad ribosomas listos para recibir nuevos transcritos (Frumkin *et al*., 2018). Se estima que, en células de mamíferos, la tasa de traducción con codones optimizados es de 4.9 codones por segundo, mientras que el de los no optimizados es 3.1 codones por segundo, es decir. Si consideramos que una proteína promedio mide alrededor de 1kb y una célula necesita sintetizar cientos de ellas por unidad de tiempo, esta diferencia es considerable (Hanson, G., & Coller, J. 2018). Cabe mencionar que indudablemente la aplicación de codones óptimos ha traído consigo excelentes resultados en la síntesis de

proteínas heterólogas, técnica que se sigue aplicando en los últimos años (Gustafsson *et al.*, 2004; Pellizza *et al*., 2018).

Explicaciones del CUB_s a través de una perspectiva ecológica y evolutiva

En el sentido de la optimización de la traducción, la selección del CUB cobra sentido. Pero en el contexto de las interacciones de organismos con su ambiente, sigue sin ser del todo claro cuáles condiciones ecológicas favorecen su evolución. Entender el fenómeno del CUB_s desde una perspectiva ecológica y evolutiva puede brindar una noción más amplia para entender mejor por qué la eficiencia traduccional se ve favorecida por la selección. Y una forma de comenzar a explorar este fenómeno complejo es en los procariontes, donde la selección puede ser más visible dado que sus poblaciones efectivas son grandes y los procesos de regulación en la expresión de genes son menos complicados que en eucariontes.

Diversos estudios han asociado el CUB_s al tiempo mínimo de duplicación celular (Sharp *et al*., 2005; Vieira-Silva & Rocha, 2010) sugiriendo que la habilidad de crecer rápido o lento está ligada a las condiciones ambientales en las cuales se desarrollan los organismos (Botzman & Margalit, 2011). A la fecha, la hipótesis más aceptada es que la selección ha favorecido el CUB_s en los genes altamente expresados y que este fenómeno es particularmente notorio en las bacterias de crecimiento rápido, ya que una consecuencia de esta optimización es la reducción en los tiempos de duplicación de la célula (Vieira-Silva & Rocha, 2010; Vieira-Silva *et al*., 2011). Las bacterias de crecimiento rápido tienden a ser copiótrofas, las cuales viven en ambientes ricos en nutrientes o en ambientes en donde la abundancia de nutrientes es fluctuante, en contraste con las bacterias oligótrofas,que viven en ambientes pobres en nutrientes (Arthur L. Koch, 2001). Aunado a esto, se ha observado una asociación entre la tasa de crecimiento y la variabilidad metabólica. Siendo que el grado de especificidad de los organismos por el ambiente se asocia con la intensidad de CUB_s en su genoma, donde los organismos más especializados (que generalmente viven en un ambiente constante) tienden a tener genomas con un CUB_s menos intenso y organismos más versátiles (que viven en ambientes fluctuantes), como algunos patógenos que pueden colonizar diversos ambientes, tienden a tener genomas en donde el CUB_s es más notorio (Botzman & Margalit, 2011).

Por otro lado, se ha propuesto que para que la selección pueda optimizar el CUB en una especie dada, existen limitantes tales como la diversidad de los tipos de ARNt y el tamaño del genoma (dos Reis *et al*., 2004). Así, en genomas reducidos que codifican para una diversidad pobre de ARNt, la selección no tendrá "espacio" para optimizar la traducción de genes altamente expresados en relación con otros genes que tienen niveles más bajos de expresión.

Finalmente, las implicaciones del CUB_s en las funciones celulares en una amplia variedad de organismos procariontes han sido poco exploradas, siendo el estudio de Supek (2010) el único publicado a la fecha. Él utilizó una metodología que emplea un clasificador *Random Forest* entrenado con PR (proteínas ribosomales) para detectar genes con altos niveles de CUB_s, y pruebas de Fisher para evaluar las categorías funcionales enriquecidas, destacando por mencionar algunas: procesos relacionados con el metabolismo de ATP, traducción, fotosíntesis y metanogénesis. Una posible desventaja de este estudio, así como otros (Botzman & Margalit, 2011; Vieira-Silva *et al*., 2010) es que sus métodos consideran a las PR como estándares de oro; o que utilizan dichos genes, así como otras proteínas altamente expresadas, para entrenar sus algoritmos sin verificar que estos genes siempre sean altamente expresados en todas las especies. Además, algunos de estos estudios no consideran adecuadamente el contenido de G+C (posiblemente causado por la deriva) de una forma rigurosa. En el presente trabajo se puso a prueba el uso *a priori* de las PR u otros genes altamente expresados como estándares de oro.

Cuantificación del CUB_s sin el uso de estándares de oro

Uno de los trabajos más emblemáticos sobre la cuantificación del CUB_s sin el uso de genes "estándar de oro" es el de dos Reis (2004). Él utilizó dos métricas para evaluar la intensidad de la selección en el CUB, y también midió el coeficiente de correlación de Pearson entre ellas. Estas métricas fueron el índice de adaptación de ARNt (tAI) [EC. 1] y el dNc [EC. 2].

EC. 1

$$W_i = \sum_{j=1}^{n_i}(1 - s_{ij})tGCN_{ij} \qquad w_i = \begin{cases} W_i/Wmax \; si \; W_i \neq 0 \\ \overline{w} \; si \; W_i = 0 \end{cases}$$

$$tAI_g = (\prod_{k=1}^{I_g} w\, i_{kg})^{1/l_g}$$

n_i =número de los diferentes ARNt que reconocen el codón i

$tGCN_{ij}$ =número de copias del ARNtij

s_{ij} =Restricción selectiva de apareo codón-anticodón, donde 0 cuando el apareo codón-anticodón en la posición tambaleante respetan regla Watson-Crick (A-T y C-G), y 1 cuando no hay interacción.

w_i =Valor de adaptación relativa del codón i

W_i =Valor de adaptación absoluta del codón i

l_g = Longitud del codón sin considerar el codón de paro

Este último es la diferencia entre el número efectivo de codones observado (Nc) y el número efectivo de codones esperado debido al contenido de G + C (Nc_e). Por otro lado, tAI, como su nombre lo indica, muestra qué tan bien adaptado está un gen determinado a la diversidad citoplasmática de ARNt que posee una célula, usando como una aproximación de la diversidad citoplasmática el número de copias de los diferentes ARNt en el genoma. Esta métrica estima cómo la eficiencia del proceso de elongación se ve afectada por la diversidad y abundancia de los ARNt. Toma en cuenta el apareamiento entre pares de bases (codón y anticodón). tAI puede tomar valores entre 0 y 1.

EC. 2

$$dNc = Nc_e - Nc$$

$$Nc_e(\Phi) = 2 + \Phi + 29 \div (\Phi + (1 - \Phi)^2)$$

$$Nc = 2 + 9/F2 + 1/F3 + 5/F4 + 3/F6$$

$$F_a = (n_a \sum_{i=1}^{k} {p_i}^2 - 1)/(n_a - 1) \qquad p_i = n_i/n_a$$

Φ =GC3s Contenido de GC en la tercera posición del codón tomando valores entre 0, ausencia y 1 todos los codones tienen GC.

Nc_e =Número de codones efectivo esperado

Nc =Número efectivo de codones observado

Fa =Promedio de F para la familia a (2,3,4 ó 6)

n_i = Frecuencia del codón i

n_a =Suma de todas las frecuencias de los codones sinónimos

k =Número de codones sinónimos diferentes para un aminoácido.

Un valor cercano a 1 de tAI indica que la frecuencia de los codones, dado un gen, corresponde a los tipos de ARNt más abundantes en la célula. Aunque originalmente tAI utiliza el número de copias de los genes para ARNt, Wei (Wei, Y. *et al*., 2019) ha utilizado los *tpm* (por sus siglas en inglés *transcripts per kilobase million*), derivadas de experimentos RNA-seq, en su lugar para una cuantificación más fiable del CUB_s redefiniendo a tAI como tAI'. No obstante, existen limitantes para el empleo de dichos datos cómo su disponibilidad para la mayoría de los organismos y cuáles fueron las condiciones donde los experimentos de RNA-seq se llevaron a cabo.

Nc cuantifica el grado en el que el uso de codones sinónimos de un gen se aparta del uso equitativo de los mismos. En el caso de sesgo extremo, por ejemplo, cuando se usa un solo codón sinónimo por aminoácido, Nc toma el valor de 20 (es decir, un solo codón para cada uno de los 20 aminoácidos). Por otro lado, cuando todos los codones son usados por un gen dado, Nc toma

el valor de 61. Para obtener la métrica dNc, el valor observado de Nc se resta del valor esperado de Nc_e (estimado a partir del contenido de G+C).

Finalmente, la correlación de Pearson, entre tAI y dNc, mide en qué grado la selección ha moldeado el CUB, con base en la co-adaptación entre los codones y los ARNt de los genes de un genoma. Dado que estas dos métricas (y su correlación) tienen una interpretación biológica directa y no emplea genes estándares de oro, se decidió utilizarlas en el presente trabajo.

No obstante, debido a que Nc_e se calcula a través de una ecuación que no toma en cuenta el contenido de G + C adecuadamente, se decidió utilizar el Nc', que es una versión mejorada de Nc (Novembre 2002) ya que corrige el sesgo introducido por el contenido de G+C de una forma más certera (Liu *et al*., 2018) (EC. 3).

EC. 3

$$X_a{}^2 = \sum_{i=1}^{k} n_a (p_i - e_i)^2 / e_i$$

$$F_a{}' = (X_a{}^2 + n_a - k)/k(n_a - 1)$$

$$Nc' = 2 + 9/F'2 + 1/F'3 + 5/F'4 + 3/F'6$$

n_a =Suma de todas las frecuencias de los codones sinónimos

k =Número de codones sinónimos diferentes para un aminoácido.

e_i =Uso esperado para el codón i dado el contexto nucleotídico.

Fa =Promedio de F para la familia a (2,3,4 ó 6)

Justificación

El fenómeno del sesgo en el uso de codones debido a la selección se ha estudiado desde hace poco más de dos décadas. Uno de sus principales atractivos es su aplicación biotecnológica como una estrategia para optimizar la expresión de genes heterólogos con el objetivo de mejorar las tasas de producción de proteínas y los rendimientos de estas en organismos modelo. Sin embargo, nuestro interés en este trabajo es de carácter evolutivo. Cuando el sesgo en el uso de codones de un gen se debe a la selección (CUB_s), en particular cuando este sesgo está co-adaptado a la abundancia de ARNt, la traducción de la proteína codificante ocurre más eficientemente. En principio, una traducción más eficiente permite un metabolismo más acelerado y un tiempo de duplicación menor. Estudios previos han sugerido una correlación evolutiva entre aquellos procariontes con un mayor CUB_s y una tasa de duplicación celular menor. También se ha sugerido una correlación entre la diversidad de ambientes y el CUB_s (procariontes que tienden a habitar en ambientes más diversos tienden a mostrar mayor CUB_s) (Botzman & Margalit, 2011). Sin embargo, los trabajos anteriores no se han realizado utilizando métodos filogenético-comparativos estrictos. En este estudio pretendemos explorar la relación entre el sesgo en el uso de codones debido a la co-adaptación con la abundancia de ARNt (CUB_s) y el tiempo mínimo de duplicación en un contexto filogenético. Pretendemos arrojar datos más fiables, desde el punto de vista estadístico, sobre las posibles causas del CUB_s y su relación con variables fisiológicas (como el tiempo mínimo de duplicación celular) y evolutivas (la filogenia). Consideramos que, aunque no es el objetivo principal del trabajo, a futuro nuestros resultados podrían tener impacto en aplicaciones biotecnológicas.

Objetivo General

Explorar el fenómeno del sesgo en el uso de codones bajo selección, en el linaje de los procariontes, en la búsqueda de explicaciones para entender las causas que lo originan.

Objetivos Particulares

- Revisitar el modelo para medición del CUB_s propuesto por Mario dos Reis (2004) y evaluar la incorporación de Nc' (Novembre, 2002) para una mejor estimación.
- Analizar posibles asociaciones entre el nivel de sesgo en el uso de codones, como señal de eficiencia traduccional, y el tiempo mínimo generacional d(h), bajo un enfoque filogenético.
- Determinar qué categorías funcionales se enriquecen en genes que muestran un uso optimizado de codones en diversos linajes de procariotas y su relación con sus estilos de vida.
- Analizar las dependencias entre el CUB_s sobre la maquinaria de traducción y distintos procesos celulares donde hay presencia del CUB_s.
- Contrastar cambios en la cuantificación del CUB_s cuando se emplean datos de RNA-seq para la estimación de las abundancias de los diferentes ARNt.

Materiales y métodos

Datos

Se emplearon 1800 genomas completos y anotados de procariontes los cuales se descargaron de NCBI (https://www.ncbi.nlm.nih.gov/genome/browse/?report=5#!/prokaryotes/). De estos 1800, se separó un subconjunto integrado por 210 genomas para los cuales se conocen sus tiempos mínimos generacionales (Vieira-Silva *et al*., 2010). Las anotaciones de *Gene Ontology* (GO) para estos 210 genomas se obtuvieron de la base de datos de UniProt (https://www.uniprot.org/).

Métricas para la cuantificación del CUB_s a nivel de gen

Se construyó un *pipeline* (CUBs_max.sh) para obtener los datos requeridos para la estimación del CUB bajo selección (CUBs). Dicho *pipeline* emplea diferentes programas a su vez, así como scripts en lenguaje Perl para limpiar y ordenar los datos. Nc se calcula con el software CodonW (J Peden, versión 1.4.2 http://codonw.sourceforge.net/) y Nc' se calcula usando ENCprime package (Novembre, 2002). Este último toma como entrada las secuencias nucleotídicas codificantes en formato fasta (CDS). Para tAI, dado que no se cuentan con datos de expresión de los diferentes ARNt de todos los genomas utilizados, CUBs_max.sh utiliza tRNAscan-SE (Lowe & Eddy, versión 2.0, http://lowelab.ucsc.edu/tRNAscan-SE/) para buscar todos los posibles genes de ARNt. Este software toma como entrada el genoma completo en formato fasta. En el caso de tAI' se emplearon *tpm* estimadas por Wei (Wei, Y. *et al*., 2019) de los experimentos de RNA-seq: SRX020805, SRX515181, SRX515174, SRX2448246, SRX1372108, SRX1638989, SRX347145; que corresponden a los organismos: *Bacteroides thetaiotaomicron*, *Bacillus subtilis*, *Escherichia coli*, *Leptospira interrogans*, *Mycobacterium tuberculosis*, *Salmonella enterica*, *Synechocystis sp.*; respectivamente. Los *tpm* sustituyeron a el número de copias de los diferentes ARNt directamente en la EC. 1. Estos experimentos se llevaron a cabo durante la fase logarítmica de crecimiento bacteriano en cepas silvestres. Finalmente CUBs_max.sh, utilizando CodonM (dos Reis *et al*., 2004), estima a partir del archivo fasta conteniendo los CDS, la frecuencia de los codones por gen (observada) y las frecuencia esperada de los codones con base en la de los nucleótidos (esperado). tAI y dNc se calcularon utilizando funciones del *script*

RENC_super.r escrito en lenguaje R que usa como entrada los archivos generados por CUBs_max.sh.

Cuantificación del CUB_s a nivel del genoma completo

Como previamente fue descrito, la S de dos Reis y R ENC' son coeficientes de correlación de Pearson entre tAI y dNc, tAI y Nc', respectivamente (dos Reis, 2004). En el caso de R ENC' se multiplica por -1 para que conserve el mismo sentido que S dos Reis. Estos miden en qué grado el sesgo bajo selección natural se explica por la co-adaptación de la abundancia de los de los ARNt. Las dos métricas fueron cuantificadas mediante funciones implementadas en R (RENC_super.r) en el set de los 1800 genomas. Y posteriormente contrastadas mediante correlaciones de Spearman con respecto al contenido de G + C y entre ellas, también en R. Para calcular el p-valor de S dos Reis y R ENC', se empleó una prueba de Monte Carlo. Para realizar esta prueba se aleatorizó la matriz de los pesos para calcular tAI (más detalles en el *script* RENC_super.r). De la misma forma se calculó R ENC' utilizando tAI' para los siete organismos con datos de RNA-seq. Finalmente, se calcularon otras métricas, las cuales utilizan a los genes ribosomales como estándares de oro. Estas fueron ENCr' (la media de Nc' del total de los genes ribosomales y posibles genes ribosomales) (Rocha, 2004), tAIrz (la media de tAI, ajustada a una distribución Z con media 0 y desviación estándar 1, del total de los genes ribosomales y posibles genes ribosomales) y ΔENC' (la diferencia de media de Nc' del total de genes y la media de los genes ribosomales y posibles genes ribosomales, dividida entre la media de Nc' del total de genes, EC. 4) (Rocha, 2004).

EC. 4

$$\Delta ENC' = \frac{NC' - ENCr'}{NC'}$$

Exploración del CUB_s en diversos linajes de procariontes

Para identificar los linajes con mayor o menor intensidad de CUB_s se visualizaron los valores de R ENC' por *phylum* y *subphylum*. Después se realizó una prueba de Krustal-Wallis de una cola y finalmente pruebas de medianas Wilcoxon *signed-rank test* entre *phyla/subphyla* y el total de los organismos. Se empleó R para realizar las pruebas estadísticas y los gráficos necesarios mediante la librería de ggplot2 (H. Wickham, 2016).

Reconstrucción filogenómica

Se infirió un árbol filogenómico de los 210 organismos para los cuales se conoce su tiempo mínimo generacional. Se detectaron diferentes marcadores filogenéticos empleados para la reconstrucción del árbol universal en el trabajo de Nicola Segata *et al*. (2014): ARNt valina ligasa *valS*, Factor de elongación G *fusA*, Metaloproteinasa ATP-dependiente de zinc *FtsH*, ARN polimerasa subunidad beta *rpoC*, Factor de elongación Tu *tufA*. Se alinearon estos genes usando el pipeline PhyloPhlAn (Nicola Segata *et al*., 2014). El mejor modelo de sustitución fue elegido mediante SMS (Vincent Lefort *et al*., 2017), y el árbol fue inferido con PhyML (Guindon S. *et al*., 2010) usando SH para el soporte de ramas.

Señal y contrastes filogenéticos

La señal filogenética del CUB_s y del tiempo mínimo generacional se estimó con la lambda de Pagel, el índice de Moran y la K de Blomberg usando el paquete phytools (Revell, 2012) y picante (S.W. Kembel *et al*. 2010) en R.

Para contrastar los tiempos mínimos generacionales y el CUB_s, se realizaron correlaciones de Spearman entre R ENC', ΔENC', tAIrz, y ENCr', y el tiempo mínimo generacional. Sin embargo, dichas correlaciones se pueden explicar por la alta señal filogenética de las métricas y la sobrerrepresentación de algunos grupos de procariontes. Por ello se realizaron contrastes filogenéticos tomando como variable dependiente al tiempo mínimo generacional d(h), usando el método de PGLS del paquete caper (Orme D. *et al*., 2013) en R.

Análisis de enriquecimiento de genes con uso optimizado de codones

Para saber qué categorías funcionales presentaban genes con altos niveles de CUB_s (medido a partir de tAI y Nc') se emplearon tres diferentes estrategias. Estas tres metodologías están codificadas en la función CUB_gsea del script RENC_super.r que hace uso de la librería topGO (Alexa A & Rahnenfuhrer J , 2019). Utiliza como entradas: i) la tabla con las métricas de los genes; ii) las anotaciones de los genes del genoma (en el formato especificado por topGO); y iii) una tabla con los ID de los genomas de los cuales se desea realizar la búsqueda y la ubicación de los archivos correspondientes. En un principio los valores de tAI y Nc', de los genes de un genoma, se ajustaron a una distribución Z con media de 0 y desviación estándar de 1, previamente fueron descartados los valores atípicos. La primera estrategia consistió en realizar dos pruebas de Kolmogorov-Smirnov (o GSEA por sus siglas en ingles *Gene Set Enrichment Analysis*), una para cada métrica (tAI y Nc') y las categorías GO compartidas en ambos resultados fueron consideradas como enriquecidas (Figura 3A). La segunda estrategia consistió en emplear la distribución hipergeométrica en una tabla de contingencia (prueba exacta de Fisher). En este caso, para definir cuáles genes tenían valores altos de CUB_s, se empleó como umbral los valores de los genes ribosomales de *E. coli* (Figura 3B). La tercera estrategia consistió en emplear la prueba exacta de Fisher, para definir un grupo del ~25% de genes con valores más altos de CUB_s. Para determinar el ~25% se recorrieron de manera recursiva aumentando la ventana de 0.1 en 0.1 ambos vectores que contenían los valore por gen de tAI y Nc' (Figura 3C). Para todas las pruebas, dado que no se tomó en cuenta la jerarquía GO, se empleó un ajuste del p-valor para pruebas múltiples usando el método BH mediante la función *p.adjust* en R. Se tomó un FDR < 0.05 para referirse a un resultado con significancia estadística.

-

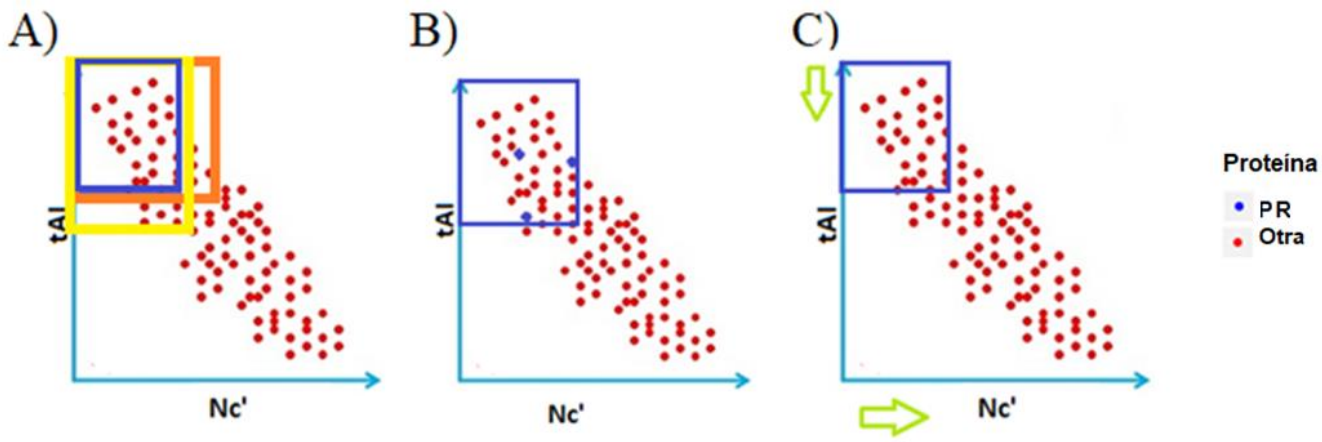

Figura 3. Estrategia para detección de categorías de genes con enriquecimiento de CUB_s. Se marcan las proteínas ribosomales en azul y el resto en rojo. El recuadro azul representa los genes significativos. **A**pruebas de Kolgomorov-Smirnov, el recuadro amarillo representa los genes significativos tras aplicar la prueba a tAI, el naranja a Nc' y el azul el subconjunto de genes significativos en ambas pruebas (A). **B** prueba exacta de Fisher tomando PR de *E. coli* como estándares. **C** prueba exacta de Fisher considerando como significativo el top ~25% de los genes. Las flechas verdes representan el proceso recursivo sobre los vectores Nc' y tAI para asignar el top.

Las tres metodologías dieron resultados similares, pero la primera estrategia parece menos arbitraria (no se necesita información previa para definir a los genes con altos niveles de CUB_s). Entonces, los resultados de sólo esta metodología empleados para los subsecuentes análisis.

Para contrastar diferencias en la optimización del CUB_s detectado por tAI' y tAI también se realizó GSEA empleando las cuentas de tAI' de cada uno de los siete organismos con datos de RNA-seq. Finalmente, para su visualización de las categorías funcionales empleó la librería ComplexHeatmap (Gu Z, *et al.*, 2016) en R y iTOL (Letunic, I., & Bork, P., 2019).

Funciones enriquecidas y tiempos mínimos generacionales

Para buscar asociaciones entre los tiempos mínimos generacionales y las categorías GO enriquecidas, se realizaron pruebas de rangos Wilcoxon para cada categoría GO. Se compararon

las distribuciones de los tiempos generacionales entre los organismos con la categoría significativa para ambas métricas (tAI y Nc') y los que no tuvieron ninguna métrica significativa.

Frecuencias dependientes y reconstrucción de caracteres ancestrales

Con el fin de probar posibles dependencias evolutivas entre la optimización del CUB_s en la maquinaria de traducción y el resto de las categorías funcionales de genes, se realizó el siguiente análisis. Primero, se definieron los eventos "T" y "G". El evento "T" indica si la categoría GO de Traducción fue significativa en ambas métricas (tAI y Nc') para un genoma dado; y el evento "G" indica si una categoría GO determinada (distinta a la de Traducción), fue significativa en ambas métricas (tAI y Nc') para el mismo genoma. Si una categoría dada (ya sea la de Traducción o alguna otra) fue significativa en ambas métricas (tAI y Nc') entonces, el evento correspondiente (ya sea "T" o "G") adquiere el valor de 1, de lo contrario es 0. Las categorías GO que solo presentaron eventos significativos para una métrica (tAI o Nc') tanto en "T" como en "G" se descartaron. De esta forma, se obtuvo una matriz en donde la primera columna contiene a cada uno de los genomas aquí estudiados, la segunda columna el evento "T" y las demás columnas los eventos "G_i". El subíndice "i" representa el resto de las categorías GO (excluyendo la categoría de Traducción). Como ya fue mencionado, las columnas "T" y "G_i" pueden tomar los valores de 0 ó 1 (Figura 4). De estas matrices se calcularon las frecuencias relativas esperadas por dependencia entre los eventos "T" con respecto al "G_i" y las esperadas por independencia.

Sean las frecuencias relativas:

$$Freq(T) = \frac{\text{Número de organismos donde el evento T ocurrió}}{\text{Número total de organismos}}$$

$$Freq(G_i) = \frac{\text{Número de organismos donde el evento } G_i \text{ ocurrió}}{\text{Número total de organismos}}$$

Frecuencia relativa esperada dependiente:

$$Freq(G_i \cap T) = \frac{\text{Número de organismos donde los eventos } G_i \text{ y T ocurrieron simultáneamente}}{\text{Número total de organismos}}$$

Frecuencia relativa esperada independiente:

$$Freq(G_i)\ x\ Freq(T)$$

Para la reconstrucción de caracteres ancestrales se definió una segunda matriz a partir de la matriz anterior, en donde las columnas "G_i" se sumaron para obtener una sola columna "G". Ahora bien, si la suma total de las "G_i" era mayor o igual a 1, entonces esta única columna "G" tomaría un valor de 1, de lo contrario sería 0 (Figura 4). Previamente, las categorías GO relacionadas con la traducción fueron eliminadas usando REVIGO (Supek F *et al*., 2011) y de manera manual usando la jerarquía GO. De esta matriz, utilizando el árbol filogenómico previamente inferido, se calcularon los estados ancestrales marginales empleando el método *re-rooting* de Yang *et al*. (1995) con una matriz "Q" simétrica, a través de la función rerootingMethod implementada en R del paquete phytools (Revell, 2012). También se realizó la reconstrucción de los estados ancestrales usando máxima parsimonia a través de la función asr_max_parsimony del paquete *castor* (Louca S *et al*., 2017) en R.

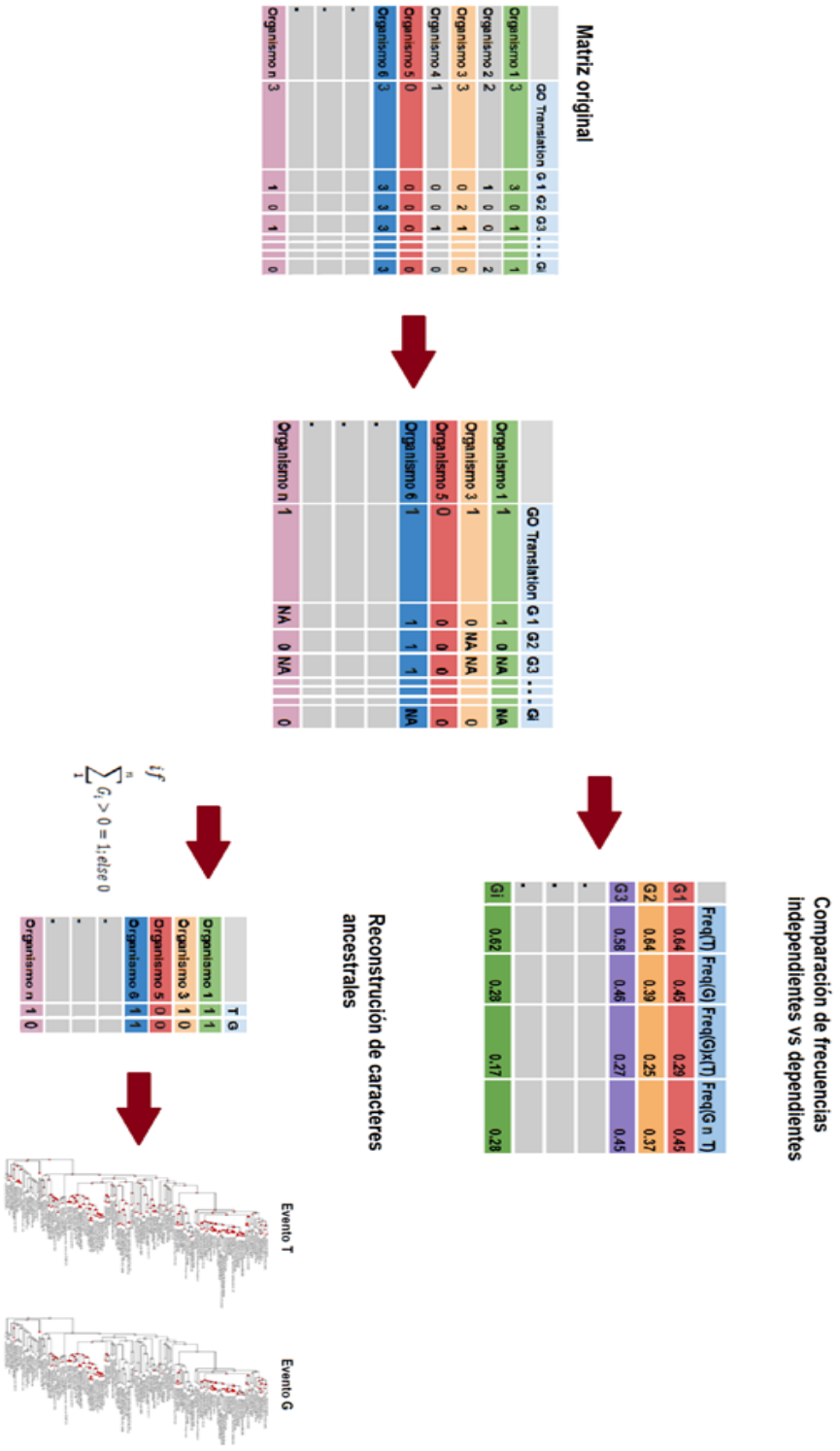

Figura 4. Diagrama donde se muestra cómo fue definido el evento t y el evento g para ambos análisis: la comparación de frecuencias (flechas superiores) y la reconstrucción de caracteres ancestrales (flechas inferiores).

Resultados

Exploración de datos

Al explorar la diversidad del conjunto de los 1800 genomas, dado que algunos *phyla* se han estudiado más que otros, se observó una sobre representación de los siguientes grupos biológicos: Firmicutes, Proteobacterias y Actinobacterias (Figura 5A). Esto mismo se observó en el subconjunto de 210 genomas (Figura 5B).

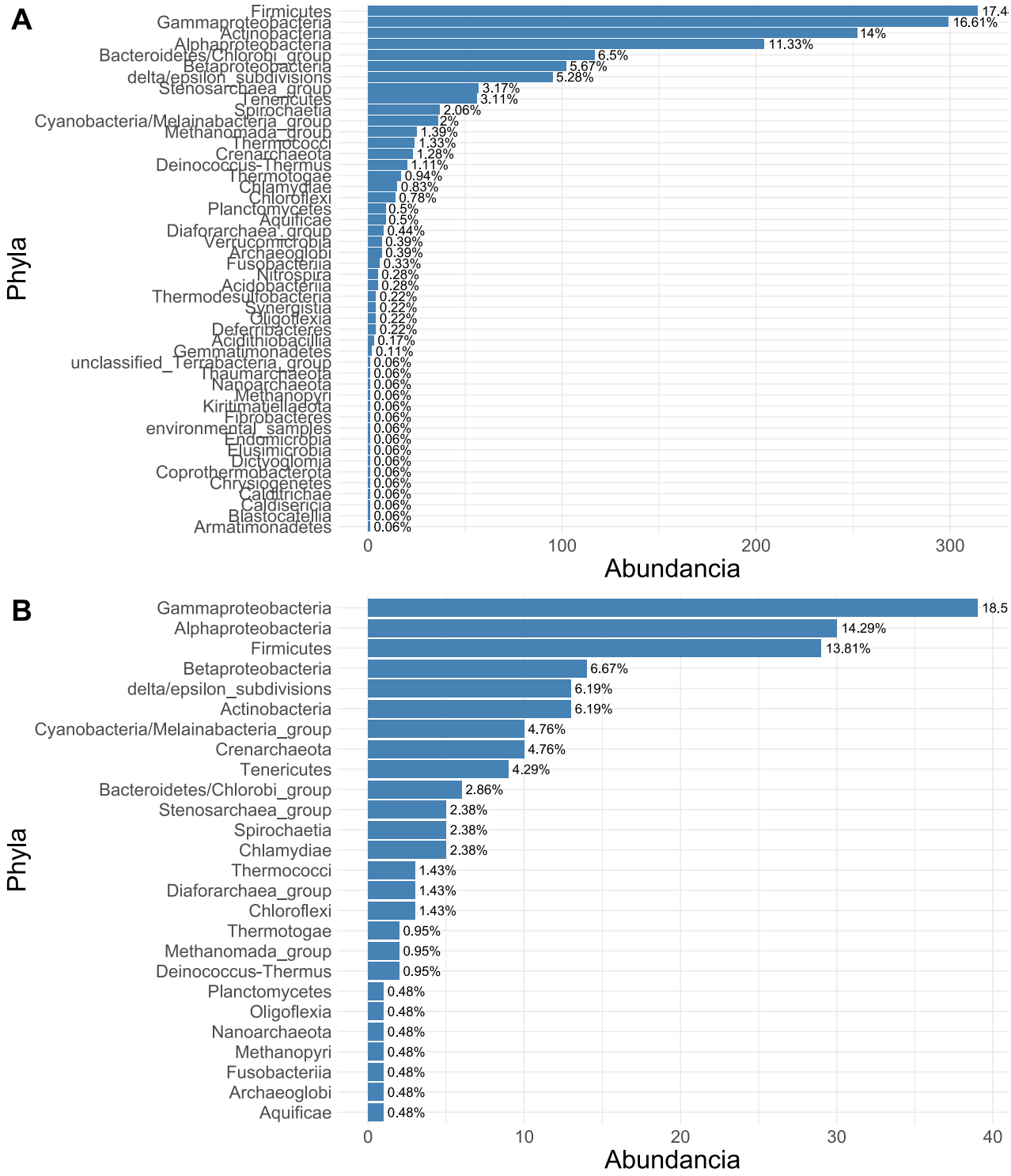

Figura 5. Distribución de los genomas estudiados en este trabajo y organizados por *phylum*. **A** conjunto de 1800 genomas; y **B** conjunto de 210 genomas (para este conjunto se conoce su tiempo mínimo generacional).

El contenido de G + C mostró ser muy variable. En el conjunto de 1800 tuvo una media de 50.1, un mínimo de 20.2 en *Buchnera aphidicola BCc* y un máximo de 74.7 en *Anaeromyxobacter dehalogenans*, *Cellulomonas fimi* y *Corynebacterium sphenisci*. En el subconjunto de 210, la media disminuyó a 47.7, el mínimo fue de 22.5 en *Wigglesworthia glossinidia* y el máximo se mantuvo. Algunos *phyla* fueron más variados en su contenido de G + C que otros y en general (sin considerar la distancia filogenética), se observa una gran diversidad en esta variable. De los *phya/subphyla* más abundantes, las Gammaproteobacterias, la subdivisión delta/epsilon y las Cyanobacterias mostraron medianas centradas al 50% de G + C, las Actinobacterias, Betaproteobacterias y Alfaproteobacterias mostraron medianas mayores al 50% de y los Firmicutes mostraron una mediana menor al 50% (Figura 6).

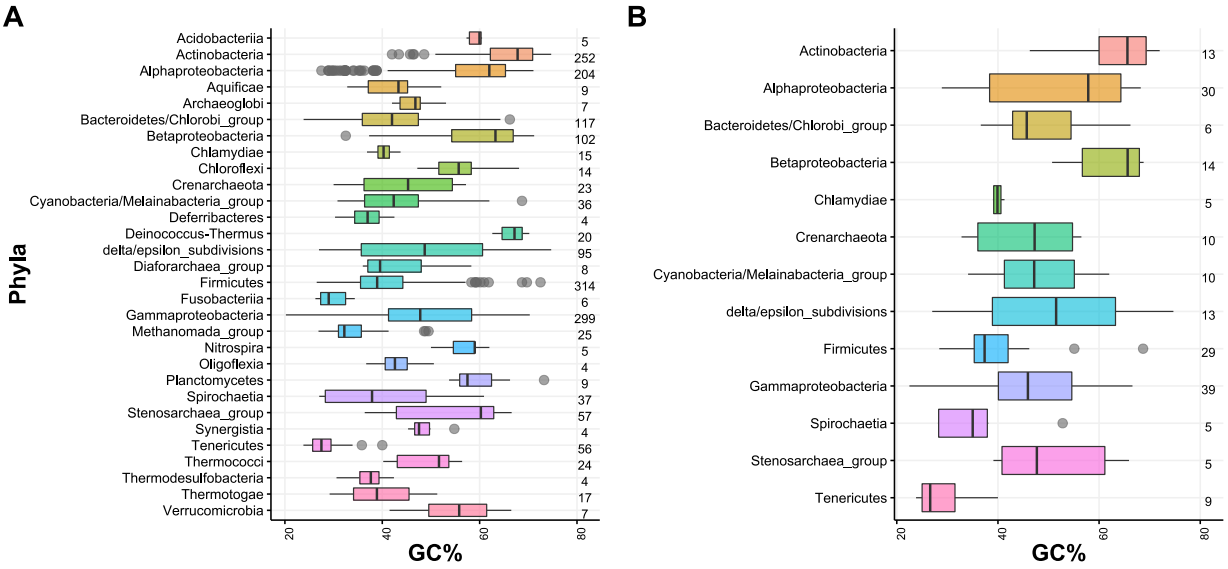

Figura 6. Distribución del contenido de G+C en porcentaje de los dos conjuntos por *phylum*. **A** conjunto de 1800 genomas; y **B** conjunto de 210 genomas.

El tamaño de los genomas también mostró un amplio rango. En el conjunto de 1800 genomas se encontró una media de 3.7 M.b., un mínimo de 0.3 M.b. en un endosimbionte secundario de la cochinilla *Trabutina mannipara*, y un máximo de 13 M.b. en *Sorangium cellulosum*. En el subconjunto de 210 genomas se encontró una media de 3.4 un mínimo de 0.49 M.b. en *Nanoarchaeum equitans*, el máximo fue el mismo. Dentro de los *phyla/subphyla* más abundantes

las Cyanobacterias fueron quienes mostraron el rango intercuartil más amplio. En general, los rangos intercuartiles de los *phyla/subphyla* más abundantes fueron entre 3 y 5 M.b. (Figura 7). Es importante considerar que a pesar de que un objetivo de este estudio fue trabajar con la mayor diversidad de genomas posibles, la sobrerrepresentación de determinados *phyla*, pueden afectar negativamente la generalización de los resultados obtenidos. Sin embargo, el conjunto de genomas muestra una alta diversidad en el tamaño y el contenido de G + C en los *phyla* más abundantes.

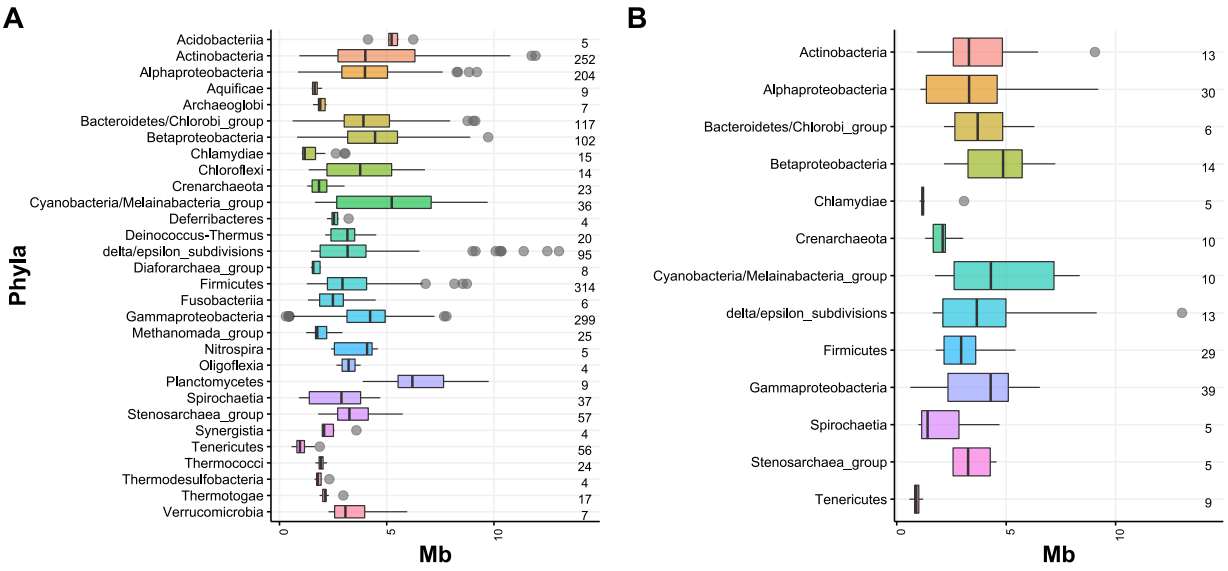

Figura 7. Distribución del tamaño de genoma en Mb de los dos conjuntos por *phyla*. **A** 1800 genomas; y **B** 210 genomas.

Mejoras en el índice S dos Reis

Posteriormente a la revisión de la literatura sobre las diferentes formas de medir la intensidad del sesgo en el uso de codones y en particular de la métrica “S” propuesta por dos Reis (2004) que mide dicha intensidad debida a la co-adaptación entre los ARNt y los codones, se planteó realizar modificaciones a la propuesta de dos Reis (2004). Dichas modificaciones consistieron en cambiar la forma en que se mide el número efectivo de codones bajo selección (dNc) por la fórmula propuesta por JA Novembre (2002) (Nc’). Esto debido a que Nc’ es una métrica que considera el fondo nucleotídico de manera directa (el contenido de GC/AT) (Liu *et al*., 2018).

Comparando ambas métricas contra el contenido de G + C, se observó que, en efecto, la nueva métrica (R ENC') que se obtiene a partir de la correlación entre Nc' y tAI, se ve menos influenciada por el contenido de G + C (Figura 8) (rho= 0.116 p-valor = 7.3e-07, vs rho=0.48 p-valor < 2.2e-16 para la métrica "S dos Reis").

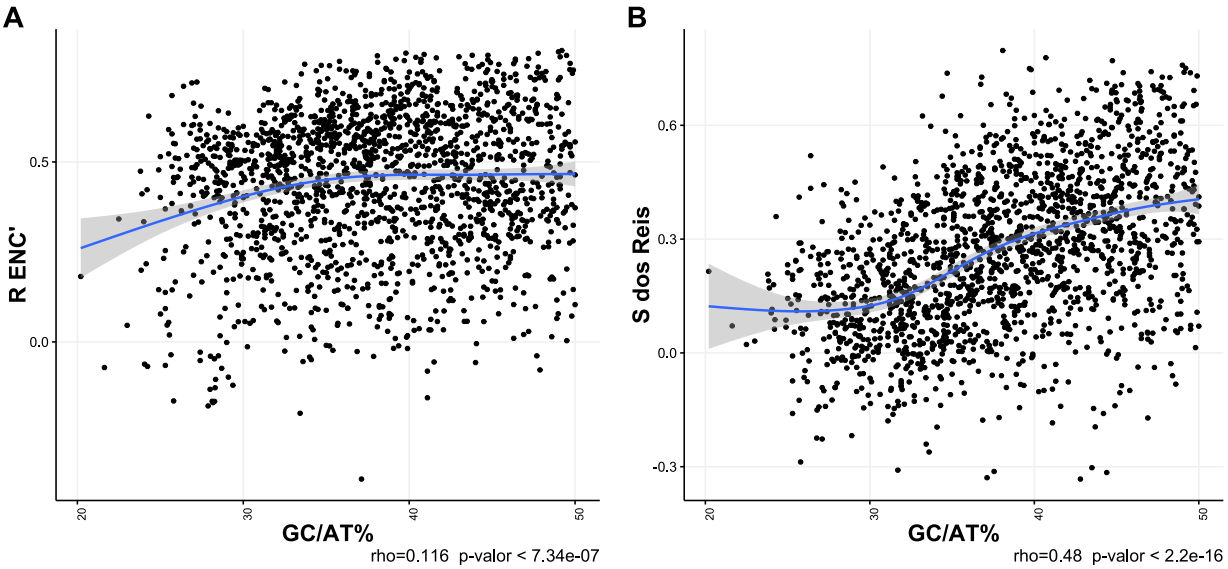

Figura 8. R ENC' y S dos Reis versus el contenido de G+C. **A** correlación entre R ENC' con respecto al contenido de G+C; **B** correlación entre S dos Reis y G+C.

A continuación, se contrastó el ajuste de ambas métricas con el panorama genómico, definido por el número de copias de ARNt y el tamaño del genoma (dos Reis, 2004). El panorama genómico define el espacio en donde actúa la selección con respecto al CUB. Los modelos de regresión lineal mostraron que el panorama genómico explica ligeramente mejor el CUB_s genómico cuantificado usando la métrica R ENC' que usando S dos Reis (Figura 9) (r^2 = 0.27 p-valor <2.2e-16, y r^2 = 0.25 p-valor <2.2e-16, respectivamente). Estos resultados demuestran que R ENC' se ve menos influenciado por el contenido de G + C que la S de dos Reis (2004). También se encontró que R ENC' detecta la señal de CUB_s en organismos donde la S dos Reis no lo hacía. Por ejemplo, aquí se estimó un valor de R ENC' cercano a uno para el genoma de *B. subtilis* (R ENC ' 0.426, p-valor < 0.05), lo que indica que este genoma está sujeto a CUB_s, caso contrario con la S de dos Reis donde se obtuvo un valor significativo, pero bajo (S dos Reis 0.27,

p-valor < 0.05). Con otras metodologías incluso experimentales se ha observado que dicho genoma sí está sujeto a un CUB_s intenso (McHardy, A. C. *et al*., 2004; Supek F *et al*., 2010), siendo este resultado más consistente con la métrica R ENC'.

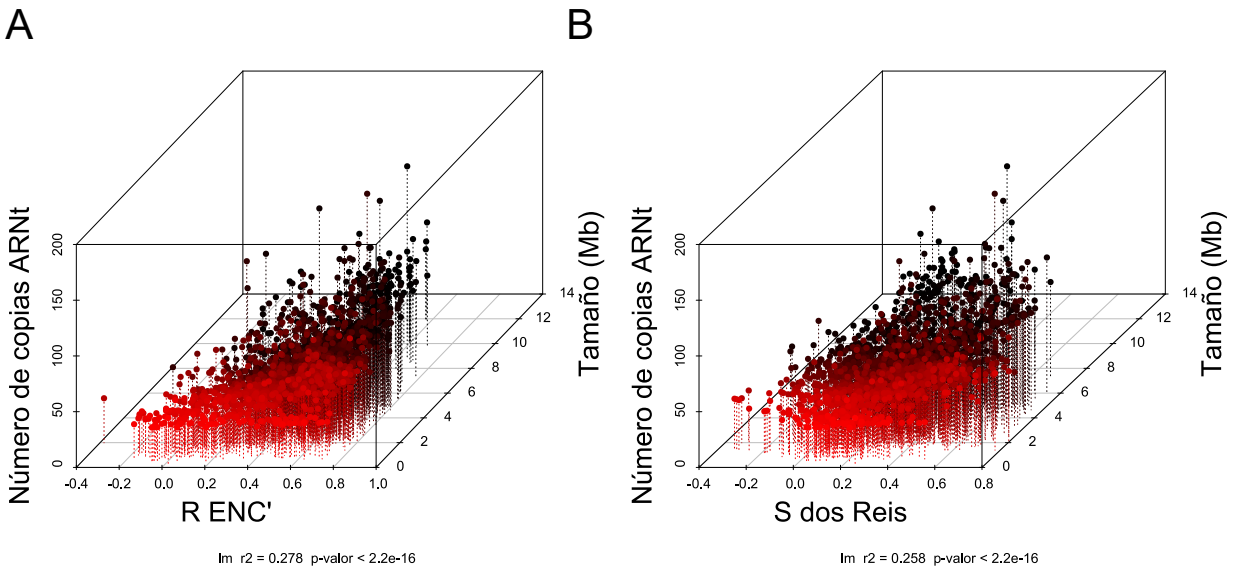

Figura 9. Correlación de R ENC' (**A**) y S dos Reis (**B**) con respecto al panorama genómico (*Genomic landscape*). El color rojo indica el tamaño del genoma en Mb (color rojo claro denota genomas pequeños).

Por otro lado, se exploró la distribución de los valores de R ENC' en el conjunto de 1800 genomas (Figura 10). Cuatro linajes, Gammaproteobacteria, Firmicutes, Actinobacteria y Betaproteobacteria, mostraron valores de CUB_s genómicos significativamente más intensos (FDR <0.05) que el resto de los taxones. También se encontró una tendencia de los taxones filogenéticamente cercanos a poseer valores similares en contraste con taxones distantes. Esto fue confirmado por una elevada señal filogenética (lambda de Pagel = 0.99, p-valor <0.001) que indica que el CUB_s a nivel de genoma tiende a estar conservado a lo largo de los linajes evolutivos.

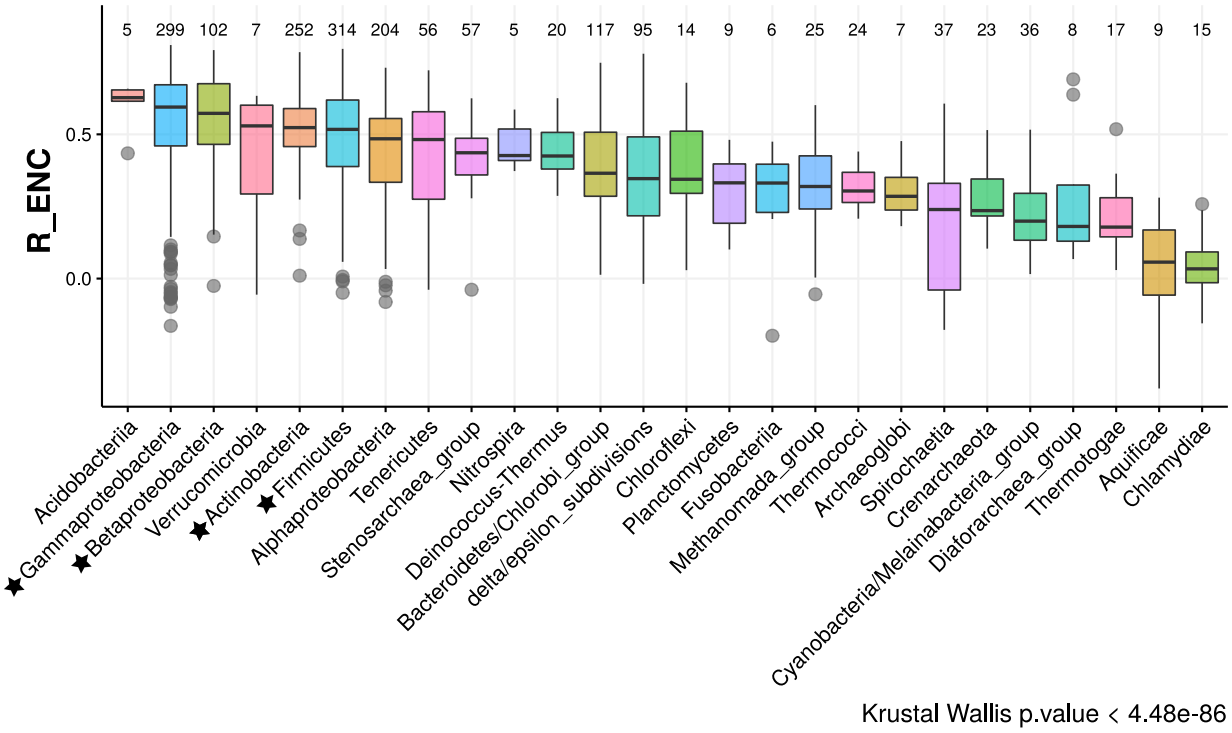

Figura 10. Distribución de los valores de R ENC' por *phylum*. Los gráficos de cajas muestran la distribución de los valores de por *phylum*. Los números en la parte superior indican el número de organismos que conforman el *phylum* en el conjunto de datos. Las estrellas representan los *phyla* con valores de R ENC' significativamente (p-valor < 0.05) mayores.

La relación entre los tiempos mínimos generacionales y el CUB_s

Uno de los principales objetivos de este trabajo fue dilucidar a qué se le atribuye la existencia de la selección en el uso de codones desde una perspectiva ecológica y evolutiva. Como anteriormente se mencionó, se ha sugerido que existe una correlación entre el tiempo mínimo generacional con el CUB_s (Rocha, 2004; Sharp, 2005). Aquí se buscó, considerando la inercia filogenética, si existe una correlación del tiempo mínimo generacional con el CUB_s estimado a partir del genoma completo (utilizando R ENC' y ΔENC') y exclusivamente a partir de las proteínas ribosomales, utilizando ENCr' y tAIrz.

Los dos modelos mejor ajustados resultaron ser las correlaciones de ΔENC' y ENCr' (PGLS' r^2 = 0.15 p-valor 6.9e-9, PGLS' r^2 = 0.24 p-valor 4.6e-14, respectivamente) (Figura 11, Tabla S1).

ENCr’ es la media de los valores de Nc’ de las PR mientras que ΔENC’ es una estandarización de la media de todos los valores Nc’ de los genes de un genoma con respecto a la media de las PR. Es decir, ambas métricas consideran el sesgo de las PR de forma directa o indirecta. A pesar de que los modelos fueron significativos (p-valor < 0.05), la asociación entre los fenómenos no fue muy cercana (Figura 11). Esto nos hizo pensar que la velocidad de duplicación quizá no sea la única variable que explica CUB_s. Por otro lado, que el sesgo en las PR se asocie de manera más sólida al tiempo generacional podría indicar que la disminución en el tiempo de duplicación se deba a un efecto subsecuente de reducir el tiempo sobre la traducción de la propia maquinaria de traducción.

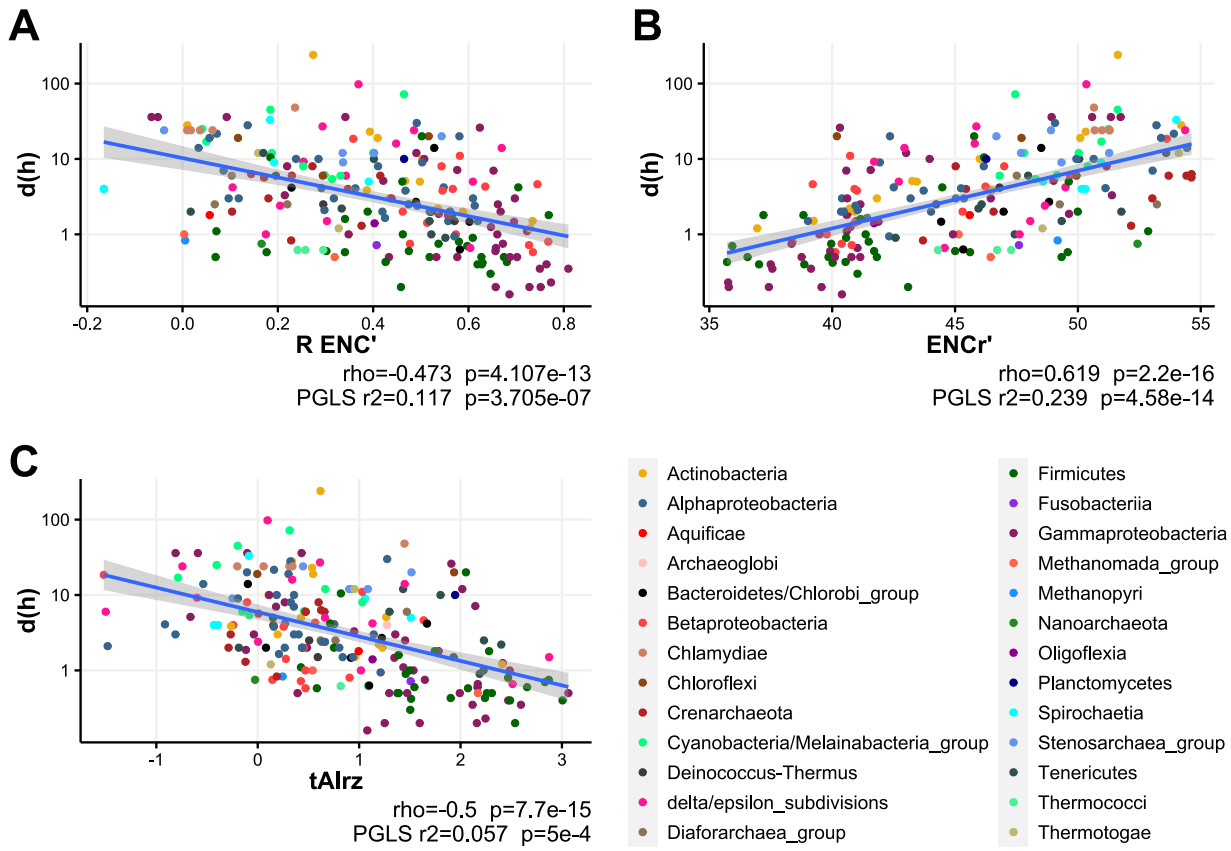

Figura 11. Tiempos mínimos generacionales y el CUB_s. Se muestran las correlaciones entre el tiempo mínimo generacional o tasa de crecimiento mínima d(h) con respecto a R ENC’ (**A**), ENC’ r (**B**) y tAIrz (**C**). Los colores indican los *phyla* a los cuales corresponde cada uno de los 210 genomas empleados para la correlación.

Inspección del CUB_s en las proteínas ribosomales

Las proteínas ribosomales han sido utilizadas como estándares de oro para cuantificar el CUB_s, puesto que son genes altamente expresados. Por otro lado, como se observó aquí, las asociaciones entre las métricas que consideran el CUB_s de las PR y el tiempo mínimo generacional fueron las que presentaron un índice de correlación mayor (Figura 11). De ambos hechos surgió la pregunta de cómo se comporta el CUB_s de las PR con respecto al resto de los genes. En el caso de que el sesgo de las PR fuese intenso, estas tendrían valores altos de tAI y bajos valores de Nc', es decir, aparecerían en el cuadrante superior izquierdo cuando ambas métricas tienen una correlación significativa e intensa cercana a 1 (Figura 12). Si fuese siempre de este modo, nos indicaría que el CUB_s actúa principalmente sobre las proteínas ribosomales de manera independiente al resto del genoma y que dichas proteínas funcionan como referencia para buscar genes con intensos niveles de CUB_s, justificando así el uso de las PR como estándares de oro. Si, por el contrario, hay genomas que claramente muestran CUB_s pero las PR no aparecen en el cuadrante superior izquierdo de la gráfica, entonces no se justifica que las PR se usen siempre para medir el sesgo en el uso de codones bajo selección.

En el conjunto de los 210 genomas anotados, 164 mostraron sus PR en el cuadrante superior izquierdo (Kolmogorov Smirnov Nc' cola superior p-valor < 0.05, tAI cola inferior p-valor < 0.05), tal como el caso de *Escherichia coli* (Figura 12A). Sólo dos organismos mostraron sus proteínas en el cuadrante inferior derecho (Kolmogorov Smirnov Nc' cola inferior p-valor < 0.05, tAI cola superior p-valor < 0.05), *Nitrosomonas_europaea_ATCC_19718* y *Syntrophus_aciditrophicus_SB*. En este último se muestra que su genoma está sujeto a CUB_s (R ENC' 0.486, p-valor < 0.05) (Figura 12B), y curiosamente, la proteína GroEL, otra importante proteína altamente expresada, aparece en la parte superior izquierda. El resto de los organismos no mostraron evidencia significativa de que las PR estuvieran enriquecidas en alguno de los dos cuadrantes. Y se observó un comportamiento similar al de *Trichormus variabilis* (Figura 12C) y *Buchnera aphidicola* (Figura 12D), donde las PR parecen estar dispersas.

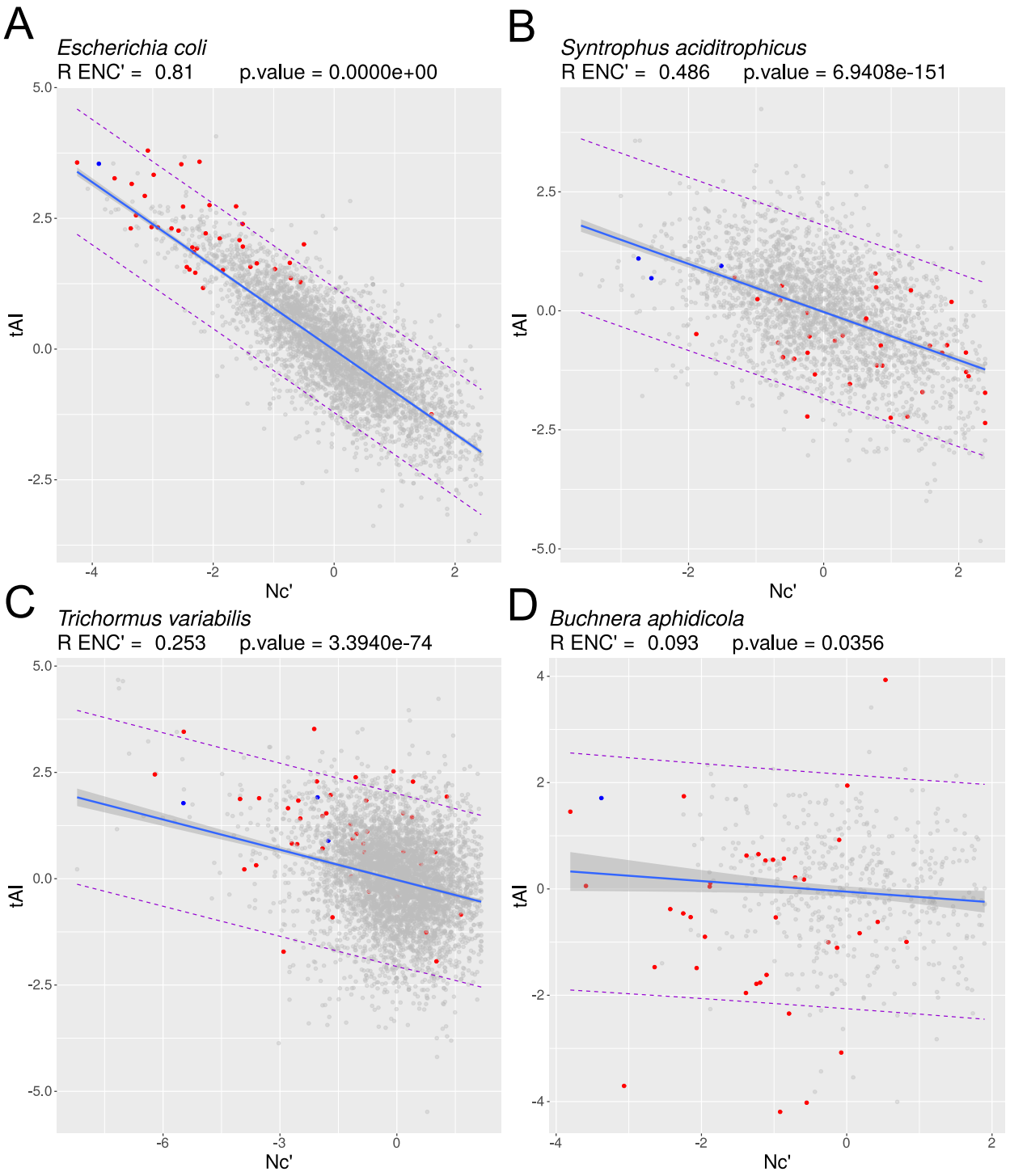

Figura 12. CUB_s a nivel de genoma y las proteínas ribosomales. Se muestran las correlaciones entre tAI y Nc' que describen el grado de selección sobre CUB a nivel de genoma, y en dónde se encuentran las PR para cuatro genomas representativos: *E. coli* (**A**); *S. aciditrophicus* (**B**); *T. variabilis* (**C**); y *B. aphidicola* (**D**). Las líneas azules muestran la pendiente de la correlación, los puntos azules corresponden a la proteína *groEL*, los rojos a RPs y los grises al resto de los genes.

Al evaluar la asociación entre el CUB_s de las PR (ENCr') y la intensidad de selección del sesgo en todo el genoma (R ENC'), mediante una correlación de Pearson, se encontró una asociación significativa ($r^2 = 0.58$, p-valor <2.2e-16). También se observó lo mismo en las proteínas GroEL, pero con menor intensidad (r^2 0.43, p-valor <2.2e-16). Estos resultados sugieren que, aunque las PR no siempre son los genes con los valores de CUB_s más intensos, la traducción es uno de los principales procesos en donde el CUB_s actúa. Y por tanto en términos generales las PR pueden servir como estándares de oro. Sin embargo, probablemente la selección del CUB está actuando en muchos procesos diferentes dependiendo del estilo de vida de los organismos.

Procesos biológicos con genes enriquecidos en CUB_s

Con el objetivo de observar en qué procesos celulares el CUB_s es más frecuente, se realizó un estudio de enriquecimiento de genes con elevado CUB_s. Para esta parte, probamos tres diferentes estrategias (ver materiales y métodos) con las cuales se obtuvieron resultados similares. Sin embargo, se decidió centrar el trabajo en la estrategia A (la pruebas de Kolmogorov Smirnov o GSEA) dado que los resultados fueron similares y la estrategia A parece ser la que menos suposiciones *a priori* requiere. En general los procesos relacionados con la traducción, expresión de genes y metabolismo de carbohidratos y de aminoácidos, y procesos relacionados con el ATP fueron las categorías mayormente enriquecidas (Figura 13). Y en menor medida fueron las categorías para ácidos tricarboxílicos, el ciclo del citrato, respiración aeróbica, plegamiento de proteínas y otros procesos metabólicos.

Figura 13. Procesos enriquecidos en genes que mostraron elevados niveles de CUB_s. Se muestran a nivel de *phylum/subphylum* las categorías GO enriquecidas. Las celdas de color café indican que la categoría se mostró enriquecida para ambas métricas (tAI y ENC') en alguno de los organismos que integran el *phylum* respectivo. El color gris indica que no hubo enriquecimiento en alguna de las dos métricas o en ninguna. Por último, el color rosa se señala que no hay genes con anotaciones correspondientes a las categorías.

Los genes codificantes para proteínas que participan en la traducción, que se ven altamente enriquecidas aquí, son cruciales para los procariotas de crecimiento rápido (Karlin, 2011, Klumpp, 2013). Esto explica por qué las métricas relacionadas con los genes ribosomales CUB (ΔENC', ENCr') se correlacionan mejor con el tiempo generacional que con la intensidad de la selección en todo el genoma (R ENC').

Los *phyla* Gammaproteobacteria y Firmicutes tuvieron más categorías enriquecidas que el resto de los taxones (Figuras 13, S1). Aunque son filogenéticamente distantes, comparten muchas de estas categorías enriquecidas. Lo que sugiere que estas bacterias se han adaptado convergentemente debido a presiones de selección similares. La fotosíntesis y la metanogénesis también fueron parte de las categorías con mayor enriquecimiento, pero sólo están representadas en cianobacterias y arqueas metanógenas, respectivamente (Figuras 13, S1). Esto muestra que la selección no sólo actúa sobre el sesgo de genes involucrados en procesos generales (como la traducción) y en procariotas de rápido crecimiento.

La mayoría de las arqueas, cianobacterias y bacterias endosimbióticas como *Buchnera aphidicola*, *Wigglesworthia glossinidia* y bacterias especializadas como *Syntrophus aciditrophicus* no mostraron categorías enriquecidas. Sin embargo, curiosamente, *Syntrophus aciditrophicus* mostró una señal intensa de CUB_s en su genoma (Figura 12B), aunque sus PR no tenían los valores de sesgo elevados, pero las proteínas GroEL sí. Al analizar el 20% superior de los genes ordenados de mayor intensidad de sesgo a menor intensidad, notamos que varios de estos eran proteínas hipotéticas y otros estos estaban involucrados en oxidorreducción, unión de metales, transferencia de ADN, traducción y plegamiento de proteínas.

Otros procesos menos enriquecidos nos llamaron la atención, como son el metabolismo de antibióticos en organismos productores o resistentes: Actinobacterias como *Mycobacterium* y *Streptomyces* (Hopwood, D. A., 2007), *Pseudomonadaceae* (Pang, Z. *et al.*, 2019) y Bacillales como *Staphylococcus* y *Bacillus* (Livermore, D. M, 2000; Bernhard, K. *et al*., 1978) y procesos relacionados con la motilidad presente en algunos Rhizobiales y *Yersinia*. Fue interesante encontrar que la mayoría de los organismos que mostraron procesos biológicos enriquecidos también mostraron enriquecimiento en el proceso de la traducción.

CUB_s y la maquinaria de traducción

En vista de encontrar una posible dependencia entre el CUB_s en el proceso de traducción y el CUB_s en otros procesos celulares, se realizó una exploración a detalle. Se compararon las frecuencias relativas de los eventos T y G_i con las frecuencias relativas esperadas dependientes y las independientes. (ver Materiales y métodos). En la mayoría de los casos las frecuencias relativas del enriquecimiento de otros procesos y la intersección de las dos frecuencias fueron iguales (67 casos de los 77 totales). Ningún proceso presentó frecuencias relativas iguales a las independientes, excepto el proceso de recombinación de ADN (GO:0006310) que sólo mostró enriquecimiento en un organismo. Esto hace sentido en el contexto de que, si el sesgo del uso de codones tiene un efecto en la eficiencia de la traducción sobre la maquinaria de traducción, habrá un impacto mayor en la eficiencia de la traducción en general del genoma. Siendo así, puede que ocurra la selección del sesgo en la maquinaria de traducción antes que en otras maquinarias celulares.

Lo siguiente que se probó, fue evaluar el supuesto de que el CUB_s sucede primero, a lo largo de la evolución, en la maquinaria de traducción, y luego en otros procesos biológicos. Para esto se hicieron dos reconstrucciones de los estados ancestrales: 1) del evento del CUB_s en el proceso de traducción; y 2) el de cualquier otro proceso definido por genes agrupados en una misma categoría funcional GO (ver materiales y métodos). Para ello se empleó el método de máxima parsimonia y el método *rerooting* con una matriz simétrica, obteniendo resultados similares (Figura 14 y S2). Al comparar las dos reconstrucciones se observó una mayor probabilidad de que se haya seleccionado el sesgo en el proceso de traducción en la parte más basal del árbol en contraste con los restos de los procesos donde es más probable que haya ocurrido cerca de las hojas del árbol (Figura 14 y S2). Esto puede sugerir que el CUB_s en el proceso de traducción es anterior (en términos evolutivos) al CUB_s en el resto de los procesos.

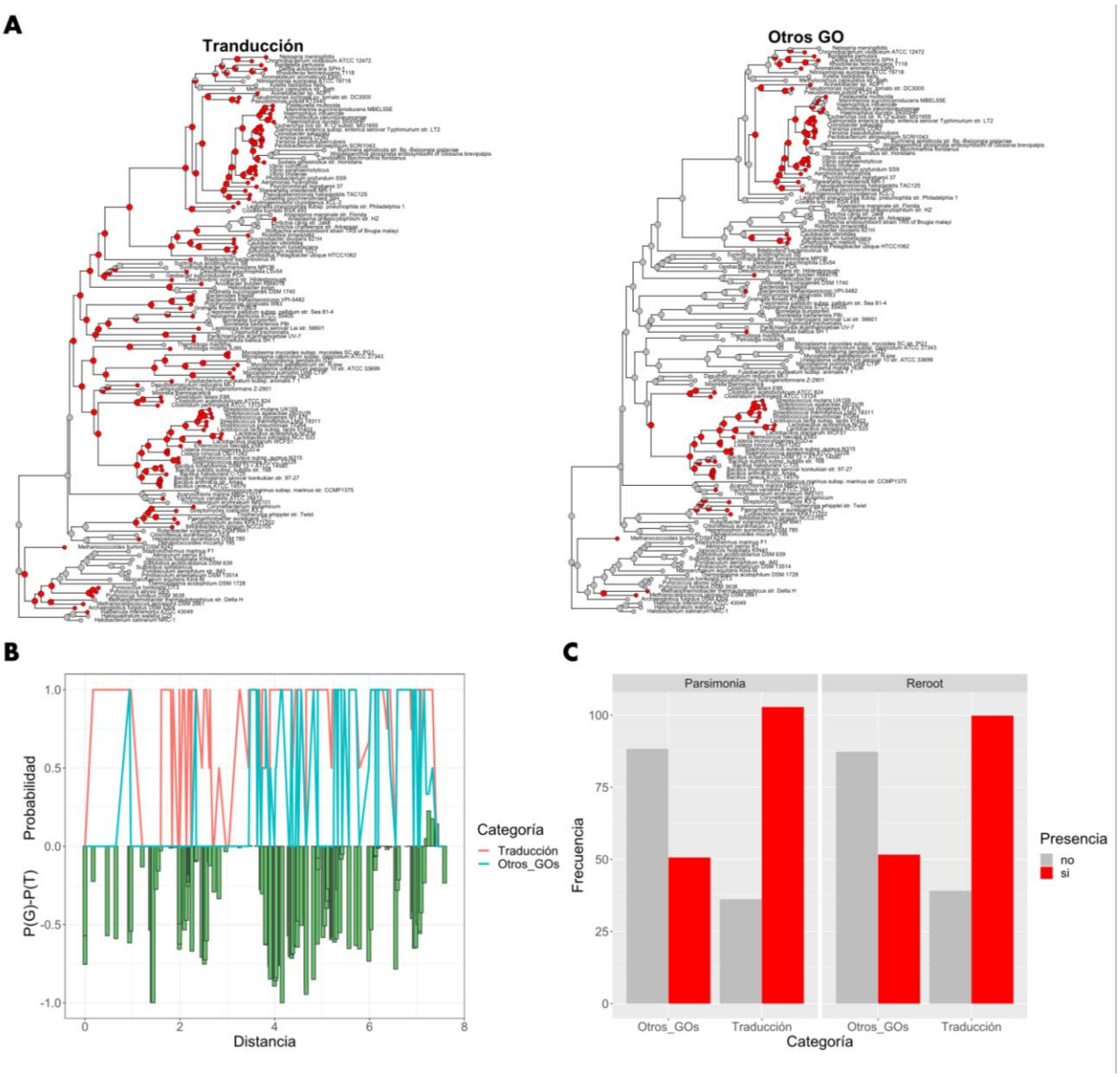

Figura 14. Reconstrucciones de caracteres ancestrales de CUB_S. En **A** se muestran el árbol con las reconstrucciones para la categoría de Traducción y el evento "G" (Otros GOs), utilizando el método de máxima parsimonia. El color rojo representa la probabilidad de que el carácter haya surgido o la posible presencia de este. En **B** se observan las probabilidades, por nodo, del evento "T" (Traducción, color azul) o "G" (Otros GOs, color rojo), con base en el método de máxima parsimonia. Las probabilidades se encuentran ordenadas con respecto a la distancia entre el nodo y la raíz del árbol. Las barras verdes representan la diferencia entre la probabilidad del evento "G" menos el evento "T". En **C** se presentan gráficos de barras donde se representa la suma de las probabilidades por cada nodo de cada evento (color rojo) y la ausencia de estos (color gris) empleando ambos métodos (parsimonia derecha y *reroting* izquierda).

Finalmente, se contrastaron los valores de R ENC', que nos indican el grado de intensidad del sesgo dada la co-adaptación de los ARNt de un genoma, entre organismos enriquecidos en el proceso de traducción y no enriquecidos (Figura 15A). La prueba de Wilcoxon mostró que las distribuciones difieren (p-valor 6.21e-13) y la regresión logística una asociación considerable (MacFadden 0.34) (Figura 15B). Esto podría implicar que en los genomas donde el CUB_s es intenso, se prioriza la eficiencia de la traducción en la maquinaria de traducción. Además, se observó que incluso en genomas con baja intensidad en CUB_s, la maquinaria de traducción puede tener CUB_s (Figura 15).

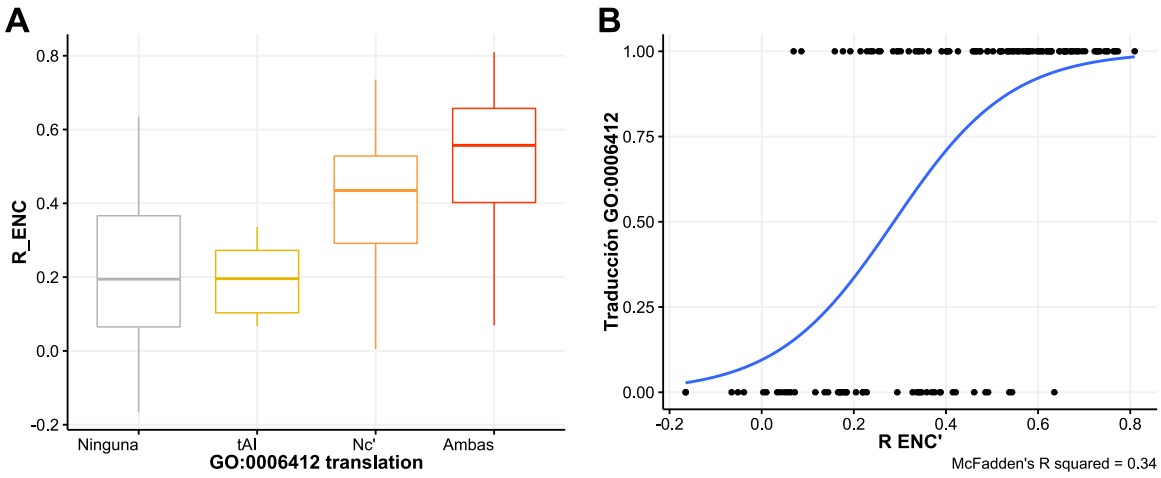

Figura 15. Contraste entre el CUB_s a nivel de genoma y en la maquinaria de traducción. En **A** se muestra la distribución de los valores de R ENC' en los genomas que mostraron alguna métrica significativa en el enriquecimiento del CUB_s en la categoría GO para traducción. En **B** se indica la correlación logística entre la significancia por ambas métricas (tAI y Nc') en el enriquecimiento del CUB_s en la categoría GO para traducción y los valores de R ENC'.

Procesos enriquecidos por CUB_s y el tiempo mínimo generacional

Debido a que el aumento en la tasa de crecimiento de los microorganismos generalmente exhibe una regulación positiva en las proteínas involucradas en la traducción, la expresión génica y la síntesis de proteínas (Scott *et al*., 2010; Molenaar *et al*., 2009; Peebo *et al*., 2015; Zavřel *et al*., 2019; Mori *et al*., 2017), se comparó la distribución de tiempos de mínimos generacionales entre organismos que tenían enriquecimiento del CUB_s en un proceso dado y organismos sin enriquecimiento en el mismo proceso. Los procesos relacionados con la traducción, el plegamiento de proteínas, la expresión génica y el ciclo del ácido tricarboxílico tuvieron diferencias significativas (Tabla S2, FDR < 0.05). Esto indica que, los tiempos mínimos de generación tienden a ser cortos en los organismos que tienen estas categorías enriquecidas.

Uso de datos de expresión de los diferentes ARNt en la estimación del CUB_s

Es importante destacar que las métricas empleadas en este trabajo son aproximaciones y poseen limitantes. En el caso de tAI se utiliza el número de copias de los diferentes ARNt como una aproximación de la abundancia de estas moléculas en el citoplasma. No obstante, se han realizado esfuerzos por mejorar dicha estimación. Recientemente se ha implementado el uso de datos de expresión de los ARNt, provenientes de experimentos RNA-seq, en lugar del número de copias, con lo cual se ha encontrado mayor señal de CUB_s en algunos genomas (Wei, Y. *et al*., 2019).

Puesto que aquí se propuso a R ENC’ como un mejor estimador, surgió la pregunta sobre si esta métrica cambiaría al utilizar datos de la expresión de los ARNt. Para ello se utilizaron los *tpm* de los ARNt, de siete organismos, empleados en el estudio de Wei *et al*., (2019). Los resultados en general no mostraron una gran diferencia entre R ENC’ y R ENC’ usando tAI' ($r^2 = 0.75$, p-valor = 0.01; Tabla 1). No obstante, la señal del CUB_s que se reporta usando tAI’ en *L. interrogans* y *Synechocystis sp*. es mayor, ambos organismos de crecimientos lento. Mientras que en *E. coli* y *S. enterica* de crecimiento rápido se observó lo contrario.

Nombre	Misma cepa	R ENC'	R ENC' tAI'
E. coli str. K-12	Si	0.810	0.739
S. enterica str. LT2	Si	0.764	0.729
B. subtilis str. 168	No	0.426	0.465
Synechocystis sp. PCC 6803	Si	0.400	0.583
B. thetaiotaomicron VPI-5482	Si	0.602	0.579
L. interrogans str. 56601	No	0.192	0.382
M. tuberculosis	No	0.409	0.330

Tabla 1. Estimaciones de S dos Reis y sus derivados incluyendo las que emplean datos de RNA-seq para la estimación de los ARNt. La columna "Misma cepa" indica si el experimento de RNA-seq se realizó con la misma cepa de la cual se estimó el CUB_s.

Lo siguiente que se realizó fue la inspección de las categorías GO enriquecidas utilizando cada métrica. Aunque la mayoría de los procesos biológicos significativos para tAI también lo fueron para tAI' (FDR < 0.05), hubo algunas excepciones. En *Synechocystis sp.* fue significativo el CUB_s del proceso de traducción reportado por tAI' y NC' pero no por tAI (Figura 16), así como otros procesos relacionados con el metabolismo de carbohidratos y aminoácidos que fueron significativos sólo para tAI' (Figura S3). *L. interrogans* mostró resultados similares además de procesos relacionados con motilidad celular y respuesta a estímulos, pero sólo significativos para tAI'. En el caso de *E. coli* y *S. entérica* se detectaron menos procesos relacionados con obtención de energía para tAI' (Figuras 16 y S3).

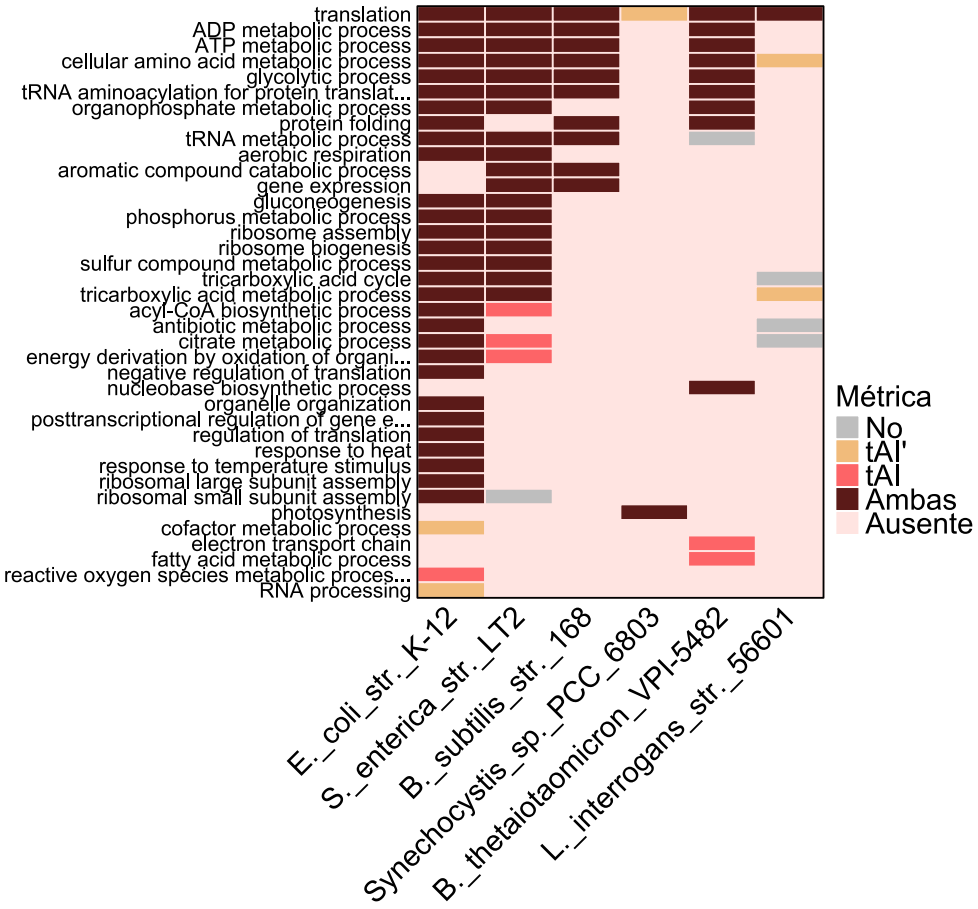

Figura 16. Contraste de procesos biológicos significativos entre tAI y tAI'. Se muestran los procesos biológicos (renglones) que resultaron significativos (FDR < 0.05) tras el análisis GSEA con respecto a la métrica Nc' y, tAI o tAI' para cada organismo (columnas). Esta última una modificación de tAI donde se utilizan datos de RNA-seq para cuantificar la abundancia de los ARNt. Las celdas de color café indican que el proceso se mostró enriquecido para ambas métricas (tAI y tAI'). El color gris indica que no hubo enriquecimiento para ninguna métrica. En amarillo o carmesí que fue significativo para tAI' o tAI, respectivamente. Por último, el color rosa se señala que no fue considerada debido a que no fue significativa para NC' o no hay genes con anotaciones correspondientes al proceso biológico.

Discusión

Contraste entre R ENC’ y S dos Reis

En el presente estudio se abordó la evolución del fenómeno del sesgo en el uso de codones bajo selección en procariontes. Para ello se emplearon dos métricas Nc’ y tAI, y la correlación de estas dos (dos Reis *et al*., 2004) para cuantificar el CUB_s a nivel del genoma. A pesar de que actualmente existen diversas formas de medir el sesgo bajo selección, la mayoría de estas se basan en proteínas altamente expresadas (tales como las proteínas ribosomales) e incluso algunas aplican técnicas de análisis topológicos de datos involucrando modelos previamente entrenados con dichos genes. Dichas métricas son útiles para identificar el fenómeno, pero no siempre tienen una interpretación biológica clara. Es por lo que decidimos emplear una métrica previamente descrita por dos Reis (2004), que se basa en la hipótesis de que el sesgo en el uso de codones se debe a una co-adaptación entre la frecuencia del uso de los codones (presencia y ausencia de estos en los genes codificantes de un genoma) y la abundancia de los diferentes ARNt dentro de un genoma. Sin embargo, como anteriormente se describió, la métrica S dos Reis tiene algunas limitantes, la principal es que considera sólo el contenido G + C en las terceras posiciones de forma teórica y no directamente tomando en cuenta el contexto nucleotídico del gen Esto, como se pudo observar, hace que la métrica continúe influenciada por el contenido de G + C. Dicha influencia disminuye en el índice propuesto, el R ENC’, que utiliza Nc’ de Novembre (2002) en lugar de dNc de dos Reis (2004). Con esta mejora, pudimos detectar CUB_s en un mayor número de genomas. Como en el caso de *Streptomyces* que, dado su alto contenido de G + C, el CUB_s se había subestimado (Sharp *et al.*, 2005). Y en el de *Bacillus subtilis* (S dos Reis 0.27, R ENC' 0.426), del cual se había reportado que tiene un genoma fuertemente sesgado por Sharp *et al.* (2005), observando un uso codones de sinónimos específicos, y por Wei *et al*. (2019), quien calculó la S dos Reis usando datos de expresión de los ARNt.

Como se mencionó anteriormente, dos Reis *et al*. (2004) propusieron que el CUB_s se encuentra acotado por dos variables genómicas principales que son: el tamaño del genoma y el número de ARNt. El último asociado a su diversidad, es decir, cuanto mayor sea el número de ARNt, mayor será diversidad de los ARNt. En este trabajo estimamos dos modelos de regresión lineal, uno

para S dos Reis y otro para R ENC' con respecto a las dos variables genómicas y observamos que R ENC' fue ligeramente mejor explicado por las variables. Este resultado es interesante porque se ajusta a la teoría formulada donde ambas variables influyen de manera positiva en la intensidad de la selección natural ejercida sobre el sesgo en el uso de codones en procariontes. Es decir, mientras un genoma sea más grande y contenga un mayor número de posibles ARNt la selección podrá actuar más eficientemente sobre el CUB.

Cabe destacar que R ENC' asume que la abundancia y diversidad de los ARNt codificados en el genoma refleja su abundancia y diversidad en el citoplasma. Esta aproximación tiene un grado de error que desconocemos. Se ha sugerido que los transcritos de ARNt son estables y su vida media suele ser más larga que el tiempo de duplicación de una célula promedio (Li, Y. *et al*., 2009). Sin embargo, existe evidencia de que los ARNt están altamente regulados (Wilusz J.E., 2015), incluso en diferentes condiciones la disponibilidad de estos puede cambiar (Svenningsen *et al*., 2016). Hoy en día con técnicas de secuenciación masiva como RNA-seq es posible obtener datos experimentales para una mejor estimación de tAI. Wei ya ha implementado el cálculo de tAI en siete organismos mostrando que, si hay diferencias, aunque no muy grandes, entre ambas aproximaciones (2019). En este trabajo se realizó una reestimación del CUB_s utilizando los *tpm* empleados por Wei para los siete organismos. Se encontraron algunas diferencias que pueden ser explicadas por dos razones: 1) que la expresión de los ARNt varía sin importar la condición, es decir, unos se expresan más que otros, independientemente del número de copias, debido a que tienen, por ejemplo, promotores más eficientes (Hori, H. *et al*., 2014); 2) la expresión de los ARNt es afectada por las condiciones a las cuales se someten los organismos (Svenningsen *et al*., 2016; Torrent, M. *et al*., 2018). Puesto que los muestreos se realizaron durante la fase logarítmica del crecimiento bacteriano puede que los genes optimizados estén implicados en procesos importantes en esa temporalidad y condiciones específicas, es decir, en diferentes condiciones o ambientes los genes optimizados podrían ser otros. Con esto podemos sugerir que usar copias de ARNt para el cálculo de R ENC’ no es una mala aproximación. No obstante, sería interesante considerar el empleo de datos de secuenciación en una mayor cantidad de organismos y condiciones, y observar que tanto los hallazgos de este trabajo cambian.

Estilos de vida y CUB_s

En estudios anteriores se ha vinculado el estilo de vida a la selección del sesgo en el uso de codones (Vieira & Rocha 2010). Para entender a profundidad dicha relación realizamos dos análisis. El primero fue la asociación entre los tiempos mínimos generacionales y el CUB_s a nivel de genoma. Anteriormente se había descrito que en el caso de los organismos copiótrofos, es decir, organismos cuyo hábitat por lo general es rico en nutrientes durante periodos de demasía, la maquinaria de expresión es abundante lo que favorece tiempos de generación cortos (Viera & Rocha 2010). Nosotros observamos que existe una correlación significativa entre el CUB_s y los tiempos mínimos (R ENC' rho = 0.47, PGLS r^2 = 0.12 p-valor < 0.05) y que, a su vez, la asociación entre los tiempos y el sesgo en genes involucrados en la maquinaria de traducción es aún más fuerte (Prueba de rangos Wilcoxon, p-valor < 0.05), como son las PR (ENCr' rho = 0.62, PGLS r^2 = 0.24, p-valor < 0.05). Esto concuerda con la noción de que el CUB_s puede ser más fuerte en organismos copiótrofos y en específico en sus proteínas altamente expresadas. La explicación de esto se debe a que dichas proteínas por lo general se encuentran bajo selección purificadora debido a que las mutaciones no sinónimas pueden tener un impacto deletéreo, por lo que las sinónimas parecen ser la única fuente de variación y a través del CUB_s puede haber cabida para optimizar su funcionamiento (Vieira-Silva *et al*., 2011).

Por otro lado, encontramos otras funciones biológicas relacionadas con bacterias de crecimiento rápido y el estilo de vida copiótrofo (Ho *et al*., 2017) con fuerte CUB_s asociadas con tiempos generacionales cortos. Por mencionar algunas: glucólisis, síntesis de aminoácidos y nucleótidos, metabolismo de ATP (Figura S1). Nosotros sugerimos que la disminución en el tiempo generacional, puede no ser un resultado directo del CUB_s sino subsecuente de la estrategia adoptada por algunos organismos por aprovechar las condiciones propicias acelerando su capacidad metabólica. Y una de las múltiples formas para acelerar su metabolismo es mejorar la eficiencia traduccional por medio del CUB.

Cabe mencionar que el CUB_s puede ser importante para los copiótrofos, pero no exclusivo. Otros organismos como son las Cianobacterias no mostraron un CUB_s fuerte a nivel genómico, sin embargo, sí mostraron algunas funciones biológicas con fuerte sesgo. Una de ellas fue la fotosíntesis. En estudios previos se observó que la mayor parte del proteoma de *Synechocystis*,

del cual depende su crecimiento en condiciones de luz, involucra proteínas implicadas en la traducción, expresión de genes y fotosíntesis. Si bien la traducción de genes involucrados en la fotosíntesis podría ser una limitante para el crecimiento, no observamos preferencia de las que mostraban tiempos generacionales más cortos a tener un sesgo fuerte en esta función. Sería interesante explorar esta asociación en una mayor cantidad de procariontes fotosintéticos para obtener más robustez en nuestro análisis.

Además de la fotosíntesis, otras categorías funcionales cautivaron nuestra atención como la metanogénesis en arqueas metanógenas, el metabolismo de antibióticos en organismos productores o resistentes a antibióticos y la motilidad en bacterias Rhizobiales. Como fue mencionado por Carbone *et al*. (2005), estas categorías funcionales enriquecidas por el CUB_s pueden darnos una noción del estilo de vida de estos organismos y las posibles adversidades que presentan en su ambiente. Para los organismos metanógenos es claro que gracias al metabolismo del metano pueden subsistir en su ambiente, mientras que tanto para las bacterias productoras y resistentes a antibióticos, una deficiencia en dicha función podría repercutir en su adecuación, siendo menos competitivas (Hibbing *et al*., 2010). En el caso de la motilidad en las Rhizobiales, se sabe que en el suelo la mayor abundancia de nutrientes se encuentra en la rizosfera, por lo tanto, la habilidad de sintetizar eficientemente la maquinaria para moverse a dicha zona o a los sistemas de raíces de las plantas sería presuntamente benéfica (Tambalo DD. *et al*., 2015). Al momento, el impacto del CUB_s en los estilos de vida es hipotético puesto que deriva de un análisis genómico, no obstante, en un futuro próximo sería interesante ponerlas a prueba experimentalmente.

En contraste, diversos organismos no mostraron señal de poseer un fuerte CUB_s. La mayoría de estos fueron endosimbiontes, tal es el caso de *Buchnera aphidicola*, *Wigglesworthia glossinidia*, *Coexiella burnetii* y organismos pertenecientes a la familia *Anaplasmataceae*. Este resultado tiene sentido porque los genomas de los endosimbiontes por lo general tienden a reducirse repercutiendo en la diversidad de los ARNt. Además, debido a la reducción del tamaño efectivo de las poblaciones, las presiones de selección tienden a relajarse (Jennifer J. Wernegreen 2015; Sharp *et al*., 2010; Gottlieb, *et al.,* 2015). No obstante, en este estudio estamos asumiendo que las copias de los genes de los ARNt representan una aproximación de la concentración de estos en condiciones estables. Es decir, puede que en condiciones de estrés la regulación de la

expresión de los ARNt cambie y de esta forma la eficiencia traduccional de los genes sea moldeada a favor de las condiciones ambientales. Un ejemplo es el caso de *Buchnera aphidicola* que cuando se induce a estrés nutricional a su hospedero (el áfido), ocurren cambios de expresión en los genes de los ARNt lo que conlleva a un posible impacto en la eficiencia traduccional de ciertos genes (Charles *et al*., 2006).

Otros organismos que tampoco mostraron un sesgo fuerte fueron arqueas pertenecientes al *phylum* Crenarchaeota. Anteriormente se había explorado el uso de codones óptimos en este *phylum*, concluyendo que el CUB no se explica por la diversidad de los ARNt sino por el contenido de G + C en codones óptimos, predichos con base a su presencia en genes altamente expresados (Baruah, *et al*., 2016). Recordando que en este estudio decidimos enfocarnos sólo en el CUB_s explicado por la co-adaptación de los ARNt y descartamos la posible selección sobre el CUB debida al contenido de G + C, es de esperarse que no hayamos encontrado señal del CUB_s en este *phylum*.

Actuación del CUB_s en los genomas de procariotas

Como sabemos, los genes de las proteínas ribosomales se han utilizado en diferentes métricas para cuantificar que tan intensamente está sesgado un gen con respecto al CUB. Esto bajo la premisa de que al ser genes altamente expresados deberán de traducirse rápidamente, de lo contrario la acumulación de sus ARN mensajeros y posterior degradación sería un gasto metabólico innecesario. Sin embargo, diversos estudios han demostrado que los mARN tienen diversos tiempos de vida media lo cual puede afectar las tasas de traducción (Laalami *et al*., 2014), además de que estos tiempos pueden cambiar en diferentes condiciones (Svenningsen *et al*., 2017) e incluso las asociaciones entre los niveles de mARN y proteína están lejos de ser perfectos (coeficiente de correlación $r^2 \sim 0.17–0.47$) (Hanson, G., & Coller, J. 2018; Guimaraes J. C., Rocha, M., & Arkin, A. P. 2014). En este trabajo, utilizando las dos métricas (Nc' y tAI) observamos que los genes de las proteínas ribosomales no siempre muestran el sesgo más elevado, lo que respalda la idea de que el uso de PR puede ser motivo de preocupación para estimar el CUB_s (Hershberg, R., & Petrov, D. A., 2012). Esto podría indicar tres posibles hipótesis: 1) que la aseveración de que un mensajero que se transcribe rápido se traducirá rápido no necesariamente es cierta, quizá basta en algunos casos que se transcriba mucho; 2) que para ciertos organismos con estilos de vida muy particulares las PR no son proteínas tan altamente

expresadas como otros de sus genes; o 3) que dado los estilos de vida muy específicos y los tamaños efectivos poblacionales pequeños en estos organismos, la selección sobre el CUB se ha relajado, olvidándose de la eficiencia traduccional. Al momento no podemos decir con certeza cual o cuales hipótesis son correctas. Sin embargo, sugerimos que, si se toman genes como estándar para cuantificar el CUB_s, estos deberían ser elegidos dependiendo del organismo y en particular de su estilo de vida. Por ejemplo, en cianobacterias se tomarían genes relacionados con la fotosíntesis, mientras que en arqueas metanógenas, genes relacionados con el metabolismo del metano.

No obstante, la mayoría de los genomas que mostraron fuerte CUB_s a nivel genoma y categorías enriquecidas con fuerte CUB_s mostraron también fuerte CUB_s en genes de la maquinaria de traducción. Teóricamente, si se optimiza la eficiencia traduccional de las proteínas altamente expresadas se optimizaría la traducción del resto de los genes, pero en menor medida, debido a la continua liberación de los ribosomas (Hershberg, R., & Petrov, D. A., 2008). De manera general podríamos suponer que estas proteínas al ser altamente expresadas serían cruciales para mantener procesos esenciales de la célula y el estilo de vida de los organismos. Algunas proteínas serían las vinculadas con la traducción, expresión de genes, o, por ejemplo, en cianobacterias las involucradas en la fotosíntesis, en copiótrofos proteínas del metabolismo celular de carbohidratos (como la glucólisis y el ciclo de Krebs). Ahora bien, si nos centramos en los genes vinculados con la maquinaria de traducción, el efecto de la eficiencia traduccional en el genoma sería aún mayor, ya que se tendría una mayor cantidad de maquinaria para traducir todos los transcritos de la célula. Este beneficio general en la célula repercutirá en la adecuación del organismo, permitiéndole responder de manera más rápida a su ambiente. Por otro lado, si se mejora la eficiencia traduccional de otros procesos celulares que no incluyan a la propia maquinaria de traducción, puede que el efecto no sea tan favorable por la falta de suficiente maquinaria.

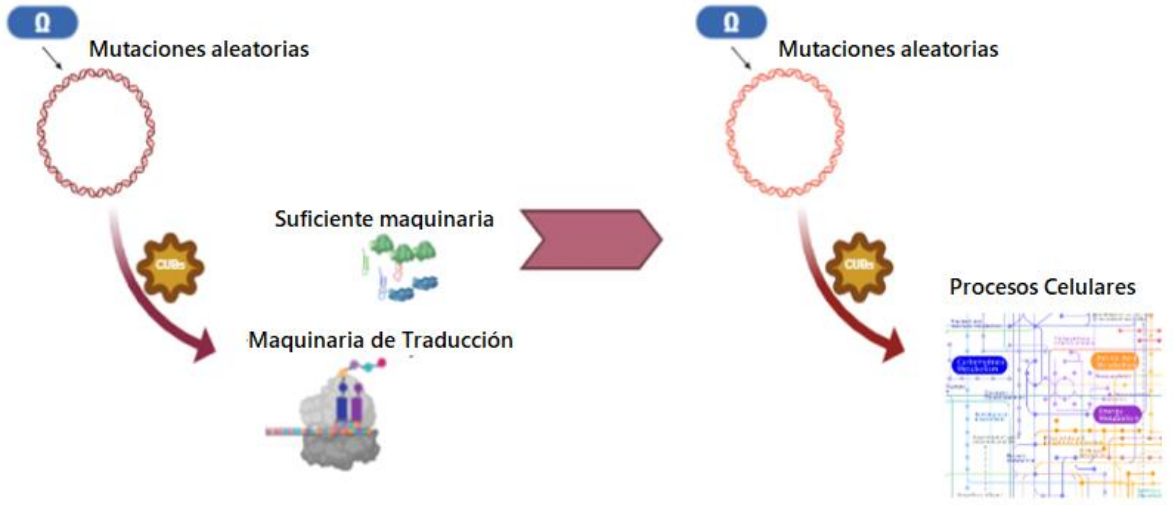

Figura 16. Modelo propuesto de cómo actúa el fenómeno del CUB_s en los genomas de los procariontes. En una primera etapa se generan variantes de manera aleatoria, la selección sobre el CUB actúa principalmente sobre la maquinaria de traducción permitiendo que exista suficiente maquinaria para optimizar la traducción en la célula. En la segunda etapa ahora la selección sobre el CUB actúa sobre otros procesos celulares relacionados con el estilo de vida del organismo en cuestión.

Visto desde esta perspectiva, proponemos un modelo de operación del CUB_s (*modus operandi*) donde la selección sobre el CUB estaría mayormente presente en la maquinaria de traducción, inclusive en un caso extremo podría ser considerada como una limitante para que la selección del CUB se efectúe en otros procesos a lo largo de la historia evolutiva de los distintos linajes. El *modus operandi* consistiría en dos etapas, donde la primera sería la optimización de la eficiencia traduccional en la propia maquinaria de traducción y en la segunda etapa se eficientizarían otros procesos biológicos, particularmente los relacionados con el estilo de vida de los organismos (Figura 16). Nuestros resultados obtenidos suportan este mecanismo en la mayoría de los procesos. Sin embargo, se necesita profundizar más en el estudio de otros procesos como, por ejemplo, la fotosíntesis que no muestra evidencia clara de posible independencia o dependencia entre la evolución del CUB_s de los dos procesos (traducción y fotosíntesis). En este caso se logra observar la señal del CUB_s en la maquinaria fotosintética, pero sólo *Synechocystis* usando datos de RNA-seq se logró observar en la maquinaria de traducción. Probablemente la limitante haya sido el estimado adecuado de la diversidad de los ARNt.

Conclusiones

A pesar de que el fenómeno del CUB_s se ha estudiado durante décadas, muchas preguntas sobre cómo opera y por qué funciona aún no tienen una respuesta clara. En particular no se comprende del todo en qué condiciones ambientales se favorece el CUBs o qué característica fenotípica es objeto de la selección (por ejemplo, la tasa mínima de duplicación). En este trabajo encontramos algunas luces al respecto, sin embargo, la pregunta sigue abierta. Estudiamos la correlación entre algunas de las variables que podrían estar involucradas en el fenómeno como la tasa mínima de duplicación, el tamaño del genoma, la abundancia de los ARNt y el estilo de vida de los organismos. Finalmente, propusimos un modelo del *modus operandi* donde el CUB_s evoluciona primero en la maquinaria de traducción y luego en otros procesos biológicos (Figura 16).

Inequívocamente, se necesita más información para mejorar la cuantificación del CUB_s y comprender mejor el fenómeno. En futuros estudios se podrían incorporar datos más fiables acerca de la disponibilidad de los ARNt para evaluar la co-adaptación con respecto a los codones. No obstante, como prueba de concepto, reafirmamos el potencial del CUB_s para contarnos sobre las características ecológicas que definen parte del estilo de vida de los organismos como son sus funciones metabólicas y tasa de crecimiento (Botzman 2011; Vieira-Silva 2010; Willenbrock, 2006). Por otro lado, encontramos que la traducción es uno de los procesos que se encuentran mayormente afectados por la evolución del sesgo. Posiblemente debido a su impacto global en la optimización de la traducción del genoma. Siendo de este modo, el uso de genes ribosomales una buena aproximación para la cuantificación del CUB_s en la mayoría de los organismos procariontes. Sin embargo, en algunos casos como las cianobacterias y arqueas se recomendarían usar otras proteínas relacionadas directamente con su estilo de vida particular.

Perspectivas

En este trabajo asumimos que las copias de los genes de los ARNt representan una aproximación de la concentración de estos en condiciones estables. Sabemos que la expresión de los ARNt puede estar regulada y posiblemente difiera en ciertas condiciones de estrés. Estudios interesantes que se podrían realizar sería evaluar la expresión de los ARNt en diferentes condiciones similares a las que se enfrentan estos organismos en la naturaleza y cuantificar el CUB_s.

Para cuantificar la métrica tAI se utilizaron valores estándar de apareo entre codón y anticodón (Ecuación 1 variable "s"). Estos datos fueron obtenidos a través de correlaciones entre valores de expresión de proteínas abundantes en *E. coli* y sus codones. Al momento no existen estudios que evalúen propiamente si la afinidad de estos puede cambiar entre especies diferentes a organismos modelo (Sabi *et al*., 2014). Quizá en algún futuro sea posible realizar una inspección más detallada con datos experimentales usando inclusive ensayos de afinidad e incorporar los resultados a dicha variable.

Por otro lado, aquí se realizó un análisis descriptivo general, sin embargo, se encontraron procesos celulares con genes enriquecidos con el CUB_s, podría ser interesante analizar desde un punto de vista más puntual quiénes son esos genes, en qué rutas participan y si existe coocurrencia de CUB_s, es decir, genes que participan en la misma ruta poseen el mismo nivel de CUB_s.

Algo importante a considerar es que por lo general se trata de dar una explicación a lo que tiene un alto CUB_s. Sin embargo, existen otras preguntas sin respuesta como qué ocurre con los genes o procesos celulares que muestran un CUB_s débil en genomas con evidencia de CUB_s. Con base en el conocimiento que se tiene al momento, se podría hipotetizar que dichos genes serían los que se expresan poco, de reciente adquisición (que aún no se han co-adaptado con el CUB del hospedero) o incluso genes accesorios que no son esenciales para el estilo de vida del hospedero. Contestar esta pregunta ayudaría a considerar otras limitantes para entender mejor cómo actúa el CUB_s.

En el caso de endosimbiontes y organismos con estilos de vida particulares, como son los de las arqueas, por lo general no se encontraron procesos celulares enriquecidos. Una de las hipótesis es que sus tamaños efectivos poblacionales son bajos, por lo cual la selección no puede actuar fuertemente sobre el CUB. Algo que se podría incorporar en estudios futuros son los tamaños efectivos poblacionales teóricos de algunos de estos organismos para comprender mejor el efecto de esta variable.

En el modelo del *modus operandi* del CUB_s se mostró como una posible limitante el que hubiese un fuerte CUB_s en los genes que participan en la maquinaria de la traducción para que el CUB_s actuara en otros procesos. Podría ser interesante también inspeccionar qué ocurre con el efecto de los procesos postraduccionales en el modelo. Esto bajo la premisa de que para el correcto funcionamiento de una proteína es necesario un adecuado plegamiento.

Finalmente, en la filogenia construida pudimos observar a simple vista que, en general, en los linajes más recientes el CUB_s a nivel del genoma (cuantificado por R ENC') y en los procesos celulares (categorías enriquecidas) es más fuerte con respecto a los *phyla* que divergieron más temprano. Sin embargo, se ha descrito que en eucariotas complejos el CUB_s es muy débil (dos Reis *et al*., 2004). Esto abre la posibilidad de indagar sobre la antigüedad del fenómeno y si esta estrategia incluso fue empleada por el último ancestro común universal (LUCA, por sus siglas en inglés).

Material Suplementario

Todos los scripts mencionados en Materiales y métodos y el material suplementario se encuentran en el file comprimido CUBs_max o en el sitio https://github.com/PacoMax/CUBs_max

Figura S1

Árbol filogenómico y mapa de calor que muestran las categorías de GO significativas por OTU. Los colores OTU representan diferentes taxones. El tamaño del triángulo verde es el tiempo mínimo de generación en horas, y el tamaño del triángulo rojo representa los CUB del genoma (-R ENC "). El color rojo en el mapa de calor corresponde a las categorías GO donde ambas métricas (tAI y Nc ') fueron significativas, naranja donde solo Nc' fue significativo, amarillo donde solo tAI fue significativo, gris ninguno fue significativo y blanco si la categoría GO está ausente.

Figura S2

Reconstrucciones de caracteres ancestrales de CUB_s utilizando el método *rerooting*. El color rojo representa la probabilidad de que el carácter haya surgido o la posible presencia de este.

Figura S3

Contraste de procesos biológicos significativos entre tAI y tAI'. Se muestran los procesos biológicos (renglones) que resultaron significativos tras el análisis GSEA con respecto a la métrica tAI y tAI' para cada organismo (columnas). Esta última una modificación de tAI donde se utilizan datos de RNA-seq para cuantificar la abundancia de los ARNt. Las celdas de color café indican que el proceso se mostró enriquecido para ambas métricas (tAI y tAI'). El color gris indica que no hubo enriquecimiento para ninguna métrica. En amarillo o carmesí que fue significativo para tAI' o tAI, respectivamente. Por último, el color rosa se señala que no hay genes con anotaciones correspondientes al proceso biológico.

Tabla S1

Correlaciones y PGLS entre las métricas que miden el CUB_s a nivel de genoma y el tiempo mínimo generacional.

Tabla S2

Resultados de la prueba de rangos Wilcoxon de las distribuciones de los tiempos mínimos generacionales por categoría GO.

Bibliografía

- Alexa A, Rahnenfuhrer J (2019). topGO: Enrichment Analysis for Gene Ontology. R package version 2.38.1.
- Baruah, V. J., Satapathy, S. S., Powdel, B. R., Konwarh, R., Buragohain, A. K., & Ray, S. K. (2016). Comparative analysis of codon usage bias in Crenarchaea and Euryarchaea genome reveals differential preference of synonymous codons to encode highly expressed ribosomal and RNA polymerase proteins. *Journal of genetics*, 95(3), 537-549.
- Bernhard, K., Schrempf, H., & Goebel, W. (1978). Bacteriocin and antibiotic resistance plasmids in Bacillus cereus and Bacillus subtilis. *Journal of bacteriology*, 133(2), 897-903.
- Bhattacharyya, S., Jacobs, W. M., Adkar, B. V., Yan, J., Zhang, W., & Shakhnovich, E. I. (2018). Accessibility of the Shine-Dalgarno sequence dictates N-terminal codon bias in E. coli. *Molecular cell*, 70(5), 894-905.
- Botzman, M., & Margalit, H. (2011). Variation in global codon usage bias among prokaryotic organisms is associated with their lifestyles. *Genome biology*, 12(10), R109.
- Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129(3), 897-907.
- Carbone, A., Kepes, F., & Zinovyev, A. (2005). Codon bias signatures, organization of microorganisms in codon space, and lifestyle. *Molecular biology and evolution*, 22(3), 547-561.
- Charles, H., Calevro, F., Vinuelas, J., Fayard, J. M., & Rahbe, Y. (2006). Codon usage bias and tRNA over-expression in Buchnera aphidicola after aromatic amino acid nutritional stress on its host Acyrthosiphon pisum. *Nucleic acids research*, 34(16), 4583-4592.
- dos Reis, M. D., Savva, R., & Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic acids research*, 32(17), 5036-5044.
- Frumkin, I., Lajoie, M. J., Gregg, C. J., Hornung, G., Church, G. M., & Pilpel, Y. (2018). Codon usage of highly expressed genes affects proteome-wide translation efficiency. *Proceedings of the National Academy of Sciences*, 115(21), E4940-E4949.
- Gottlieb, Y., Lalzar, I., & Klasson, L. (2015). Distinctive genome reduction rates revealed by genomic analyses of two Coxiella-like endosymbionts in ticks. *Genome biology and evolution*, 7(6), 1779-1796.
- Grantham, R., Gautier, C., Gouy, M., Mercier, R., & Pave, A. (1980). Codon catalog usage and the genome hypothesis. *Nucleic acids research*, 8(1), 197-197.
- Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., & Mercier, R. (1981). Codon catalog usage is a genome strategy modulated for gene expressivity. *Nucleic acids research*, 9(1), 213-213.
- Gu Z, Eils R, Schlesner M (2016). "Complex heatmaps reveal patterns and correlations in multidimensional genomic data." *Bioinformatics*.
- Guimaraes, J. C., Rocha, M., & Arkin, A. P. (2014). Transcript level and sequence determinants of protein abundance and noise in Escherichia coli. *Nucleic acids research*, 42(8), 4791-4799.

- Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, 59(3), 307-321.
- Gustafsson, C., Govindarajan, S., & Minshull, J. (2004). Codon bias and heterologous protein expression. *Trends in biotechnology*, 22(7), 346-353.
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York*, 2016.
- Hanson, G., & Coller, J. (2018). Codon optimality, bias and usage in translation and mRNA decay. *Nature reviews Molecular cell biology*, 19(1), 20-30.
- Hershberg, R., & Petrov, D. A. (2008). Selection on codon bias. *Annual review of genetics*, 42, 287-299.
- Hershberg, R., & Petrov, D. A. (2012). On the limitations of using ribosomal genes as references for the study of codon usage: a rebuttal. *PloS one*, 7(12), e49060.
- Hibbing, M. E., Fuqua, C., Parsek, M. R., & Peterson, S. B. (2010). Bacterial competition: surviving and thriving in the microbial jungle. *Nature Reviews Microbiology*, 8(1), 15-25.
- Ho, A., Di Lonardo, D. P., & Bodelier, P. L. (2017). Revisiting life strategy concepts in environmental microbial ecology. *FEMS microbiology ecology*, 93(3), fix006.
- Hopwood, D. A. (2007). How do antibiotic-producing bacteria ensure their self-resistance before antibiotic biosynthesis incapacitates them?. *Molecular microbiology*, 63(4), 937-940.
- Hori, H., Tomikawa, C., Hirata, A., Toh, Y., Tomita, K., Ueda, T., & Watanabe, K. (2014). Transfer RNA Synthesis and Regulation. *eLS*.
- Ikemura, T. (1981a). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *Journal of molecular biology*, 146(1), 1-21.
- Ikemura, T. (1981b). Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *Journal of molecular biology*, 151(3), 389-409.
- Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Molecular biology and evolution*, 2(1), 13-34.
- Karlin, S., Mrázek, J., Campbell, A., & Kaiser, D. (2001). Characterizations of highly expressed genes of four fast-growing bacteria. *Journal of Bacteriology*, 183(17), 5025-5040.
- Klumpp, S., Scott, M., Pedersen, S., & Hwa, T. (2013). Molecular crowding limits translation and cell growth. *Proceedings of the National Academy of Sciences*, 110(42), 16754-16759.
- Koch, A. L. (2001). Oligotrophs versus copiotrophs. *Bioessays*, *23*(7), 657-661.
- Lefort, V., Longueville, J. E., & Gascuel, O. (2017). SMS: smart model selection in PhyML. *Molecular biology and evolution*, 34(9), 2422-2424.
- Letunic, I., & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic acids research*, 47(W1), W256-W259.
- Li, Y., & Zhou, H. (2009). tRNAs as regulators in gene expression. *Science in China Series C: Life Sciences*, 52(3), 245-252.

- Liu, S. S., Hockenberry, A. J., Jewett, M. C., & Amaral, L. A. (2018). A novel framework for evaluating the performance of codon usage bias metrics. *Journal of The Royal Society Interface*, 15(138), 20170667.
- Livermore, D. M. (2000). Antibiotic resistance in staphylococci. *International journal of antimicrobial agents*, 16, 3-10.
- Louca S, Doebeli M (2017). "Efficient comparative phylogenetics on large trees." *Bioinformatics*..
- McHardy, A. C., Pühler, A., Kalinowski, J., & Meyer, F. (2004). Comparing expression level-dependent features in codon usage with protein abundance: an analysis of 'predictive proteomics'. *Proteomics*, 4(1), 46-58.
- Molenaar D, van Berlo R, de Ridder D, Teusink B (2009) Shifts in growth strategies reflect tradeoffs in cellular economics. *Mol Syst Biol* 5: 323.
- Mori, M., Schink, S., Erickson, D. W., Gerland, U., & Hwa, T. (2017). Quantifying the benefit of a proteome reserve in fluctuating environments. *Nature communications*, 8(1), 1-8.
- Novembre, J. A. (2002). Accounting for background nucleotide composition when measuring codon usage bias. *Molecular biology and evolution*, 19(8), 1390-1394.
- Orme, D., Freckleton, R., Thomas, G., Petzoldt, T., & Fritz, S. (2013). The caper package: comparative analysis of phylogenetics and evolution in R. R package version, 5(2), 1-36.
- Pang, Z., Raudonis, R., Glick, B. R., Lin, T. J., & Cheng, Z. (2019). Antibiotic resistance in Pseudomonas aeruginosa: mechanisms and alternative therapeutic strategies. *Biotechnology advances*, 37(1), 177-192.
- Pellizza, L., Smal, C., Rodrigo, G., & Arán, M. (2018). Codon usage clusters correlation: towards protein solubility prediction in heterologous expression systems in E. coli. *Scientific reports*, 8(1), 1-12.
- Peebo, K., Valgepea, K., Maser, A., Nahku, R., Adamberg, K., & Vilu, R. (2015). Proteome reallocation in Escherichia coli with increasing specific growth rate. *Molecular BioSystems*, 11(4), 1184-1193.
- Plotkin, J. B., & Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nature Reviews Genetics*, 12(1), 32.
- Quax, T. E., Claassens, N. J., Söll, D., & van der Oost, J. (2015). Codon bias as a means to fine-tune gene expression. *Molecular cell*, 59(2), 149-161.
- Reis, M. D., Savva, R., & Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic acids research*, 32(17), 5036-5044.
- Revell, L. J. (2012) phytools: An R package for phylogenetic comparative biology (and other things). Methods Ecol. Evol. 3 217-223.
- Rocha, E. P. (2004). Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. *Genome research*, 14(11), 2279-2286.
- S.W. Kembel, P.D. Cowan, M.R. Helmus, W.K. Cornwell, H. Morlon, D.D. Ackerly, S.P. Blomberg, and C.O. Webb. 2010. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* 26:1463-1464.
- Salim, H. M., & Cavalcanti, A. R. (2008). Factors influencing codon usage bias in genomes. *Journal of the Brazilian Chemical Society*, 19(2), 257-262.

- Sabi, R. E. N. A. N. A., & Tuller, T. A. M. I. R. (2014). Modelling the efficiency of codon–tRNA interactions based on codon usage bias. *DNA research*, 21(5), 511-526.
- Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T (2010) Interdependence of cell growth and gene expression: origins and consequences. *Science* 330: 1099 – 1102
- Segata, N., Börnigen, D., Morgan, X. C., & Huttenhower, C. (2013). PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. *Nature communications*, 4(1), 1-11.
- Sharp, P. M., Bailes, E., Grocock, R. J., Peden, J. F., & Sockett, R. E. (2005). Variation in the strength of selected codon usage bias among bacteria. *Nucleic acids research*, 33(4), 1141-1153.
- Sharp, P. M., Emery, L. R., & Zeng, K. (2010). Forces that influence the evolution of codon bias. Philosophical Transactions of the *Royal Society B: Biological Sciences*, 365(1544), 1203-1212.
- Supek F, Bošnjak M, Škunca N, Šmuc T (2011). "REVIGO summarizes and visualizes long lists of Gene Ontology terms" *PLoS ONE 2011*. doi: 10.1371/journal.pone.0021800
- Supek, F., Škunca, N., Repar, J., Vlahoviček, K., & Šmuc, T. (2010). Translational selection is ubiquitous in prokaryotes. *PLoS genetics*, 6(6), e1001004.
- Svenningsen, S. L., Kongstad, M., Stenum, T. S., Muñoz-Gómez, A. J., & Sørensen, M. A. (2017). Transfer RNA is highly unstable during early amino acid starvation in Escherichia coli. *Nucleic acids research*, 45(2), 793-804.
- Tambalo DD, Yost CK, Hynes MF. 2015. Motility and chemotaxis in the Rhizobia. In: Biological nitrogen fixation. *New Jersey: John Wiley & Sons*, Ltd. 337-348
- Torrent, M., Chalancon, G., de Groot, N. S., Wuster, A., & Babu, M. M. (2018). Cells alter their tRNA abundance to selectively regulate protein synthesis during stress conditions. *Science signaling*, 11(546).
- Vieira-Silva, S., & Rocha, E. P. (2010). The systemic imprint of growth and its uses in ecological (meta) genomics. *PLoS genetics*, 6(1), e1000808.
- Vieira-Silva, S., Touchon, M., Abby, S. S., & Rocha, E. P. (2011). Investment in rapid growth shapes the evolutionary rates of essential proteins. *Proceedings of the National Academy of Sciences*, 108(50), 20030-20035.
- Wei, Y., Silke, J. R., & Xia, X. (2019). An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria. *Scientific reports*, 9(1), 1-11.
- Wernegreen, J. J. (2015). Endosymbiont evolution: predictions from theory and surprises from genomes. *Annals of the New York Academy of Sciences*, 1360(1), 16.
- Wilusz, J. E. (2015). Controlling translation via modulation of tRNA levels. *Wiley Interdisciplinary Reviews: RNA*, 6(4), 453-470.
- Yang, Z., Kumar, S., Nei, M. (1995) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, 141, 1641-1650.
- Zavřel, T., Faizi, M., Loureiro, C., Poschmann, G., Stühler, K., Sinetova, M., ... & Červený, J. (2019). Quantitative insights into the cyanobacterial cell economy. *Elife*, 8, e42508.