PEDECIBA

Doctorado en Ciencias Biológicas.

Evolución de Sesgos Selectivos en el Uso de Codones Sinónimos y Aminoácidos

Andrés Iriarte Odini

airiarte@fcien.edu.uy

Orientador: Dr. Héctor Musto

hmusto@fcien.edu.uy

Tribunal: Dr. Enrique P. Lessa
Dr. Juan R. Arbiza
Dra. Graciela García

Universidad de la República
Facultad de Ciencias, Instituto de Biología
Departamento de Ecología y Evolución
Laboratorio de Organización y Evolución del Genoma

Marzo 2013

Índice

Resumen

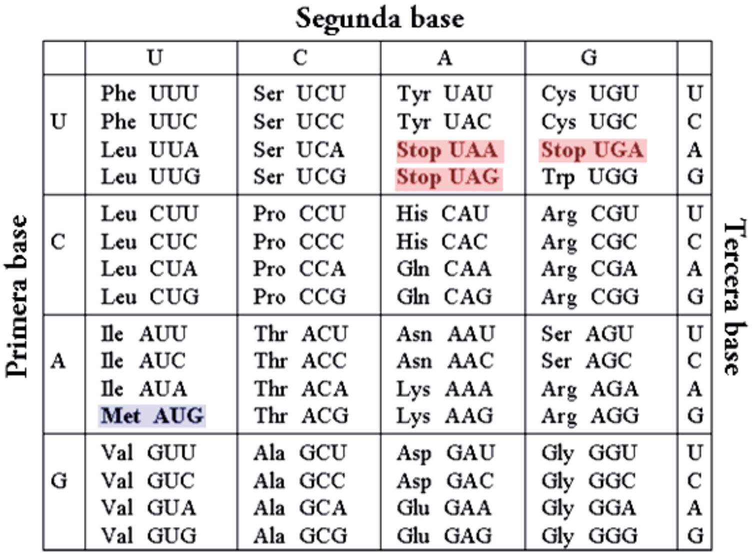
En este trabajo de tesis hemos implementado una combinación de herramientas bioinformáticas para analizar el sesgo en el uso de codones sinónimos (UCS) y aminoácidos (UAA) en grupos de organismos filogenéticamente bien definidos y suficientemente muestreados. En Procariotas: Familia Enterobacteriaceae y Filum Firmicutes con especial énfasis en la clase Mollicutes, Eucariotas unicelulares: hongos del género *Aspergillus* y Virus: *West nile virus* e *Influenza A*. Como estrategia general se han sumado a los análisis clásicos, estudios comparativos y el análisis de los sesgos en un marco filogenético. En relación a esto último, utilizamos técnicas de reconstrucción de estados ancestrales. Se buscó definir tendencias en los distintos linajes, contraponer hipótesis y aproximarnos a entender las dinámicas del cambio en los sesgos y en las preferencias de los codones traduccionalmente óptimos. También se analizaron aspectos vinculados a la inercia filogenética, principalmente en relación a la evolución de los codones óptimos y la presencia de tRNAs. Los estudios en procariotas (Capitulos I y IV), permitieron contraponer las características de los sesgos en organismos parásitos y endosimbiontes. En los estudios presentados en los Capítulos I y II se describieron los sesgos selectivos en el UCS y se contrapusieron los efectos de la selección operando a nivel de la velocidad y la fidelidad. En el Capítulo III se abordó la descripción de los sesgos en el UCS a nivel global de los virus de RNA, entre cepas y en relación al hospedero. Los resultados obtenidos en ese capítulo se discutieron en relación a las dinámicas poblacionales y el rol de la composición del genoma en la evasión del sistema inmune del hospedero. En secciones separadas se discuten las principales contribuciones de los trabajos presentados en el marco del conocimiento actual. Finalmente se presenta una breve descripción de las técnicas de reconstrucción de estados ancestrales utilizadas.

1 Introducción al Uso de Codones Sinónimos y Aminoácidos.

1.1 Uso de Codones Sinónimos (UCS)

El proceso de traducción, la síntesis de proteínas en la célula, es esencial en todos los organismos y extremadamente costoso en términos de energía, materia prima y tiempo. Principalmente para los organismos unicelulares, existe una relación directa entre el tiempo que demoran los procesos celulares y la tasa de crecimiento y división, lo que lleva en muchos casos a una diferencia en la eficacia darwiniana o *fitness* entre los organismos.

En el proceso de traducción, el código genético determina cuál o cuáles de los 61 tripletes o codones corresponde con cada uno de los 20 aminoácidos (Figura 1). Esto es lo que se conoce como redundancia en el código, lo cual implique que existen una enorme cantidad de combinaciones posibles de secuencias nucleotídicas para una misma secuencia proteica. Estos grados de libertad pueden ser "utilizados" en la evolución como materia prima para optimizar el proceso de traducción en cualquiera de sus aspectos. Codones distintos que codifican para un mismo aminoácido se conocen como "codones sinónimos".



Segunda base

Primera base	U	C	A	G	Tercera base
U	Phe UUU Phe UUC Leu UUA Leu UUG	Ser UCU Ser UCC Ser UCA Ser UCG	Tyr UAU Tyr UAC Stop UAA Stop UAG	Cys UGU Cys UGC Stop UGA Trp UGG	U C A G
C	Leu CUU Leu CUC Leu CUA Leu CUG	Pro CCU Pro CCC Pro CCA Pro CCG	His CAU His CAC Gln CAA Gln CAG	Arg CGU Arg CGC Arg CGA Arg CGG	U C A G
A	Ile AUU Ile AUC Ile AUA Met AUG	Thr ACU Thr ACC Thr ACA Thr ACG	Asn AAU Asn AAC Lys AAA Lys AAG	Ser AGU Ser AGC Arg AGA Arg AGG	U C A G
G	Val GUU Val GUC Val GUA Val GUG	Ala GCU Ala GCC Ala GCA Ala GCG	Asp GAU Asp GAC Glu GAA Glu GAG	Gly GGU Gly GGC Gly GGA Gly GGG	U C A G

Figura 1. Código genético "universal". En rojo se muestran los codones que indican el termino de la traducción (codones *stop*) y en azul el codón codificante para metionina que es utilizado como codón de inicio de la traducción.

En un principio se podría asumir que el intercambio entre dos codones sinónimos no tiene efecto o es silencioso, sin embargo ha sido demostrado en una gran variedad de organismos que los diferentes codones son utilizados en diferente frecuencia, fenómeno que se conoce como sesgo en el uso de codones sinónimos (UCS). Aunque el código genético es mayormente conservado en la evolución (código "universal"), la dirección y la intensidad de los sesgos varían entre organismos (variación inter-genómica), entre genes dentro de un organismo (variación intra-genómica) y entre regiones dentro de un gen (variación intra-génica) (Akashi, 1994; Andersson & Sharp, 1996; Cannarozzi *et al.*, 2010; Duret, 2002a, b; Gouy & Gautier, 1982; Hershberg & Petrov, 2008; Ikemura, 1985; Novoa & Ribas de Pouplana, 2012; Retchless & Lawrence, 2011; Shabalina *et al.*, 2013; Sharp & Li, 1986; Sharp *et al.*, 2005; Sharp *et al.*, 1988; Supek *et al.*, 2010).

Las explicaciones a la existencia de estos sesgos caen en dos grupos o clases: las seleccionistas y las neutralistas. Las primeras sugieren que el sesgo en el UCS contribuye de alguna forma a la eficiencia y/o a la exactitud de la traducción y/o regulación, y por lo tanto es generada y mantenida por la selección natural. La explicación neutralista, por su parte, argumenta que el sesgo existe por un patrón mutacional no aleatorio. Este apartamiento de la aleatoriedad puede estar dado porque algunos codones son más "mutables" que otros o porque distintos organismos experimentan distintos sesgos mutacionales que llevan a sesgos composicionales, los que finalmente repercuten en el UCS.

1.1.1 Variación del UCS a nivel inter-genómico

Algunos trabajos han demostrado que las diferencias en el UCS entre distintos organismos están dadas por los diferentes contenidos en guanina y citosina (GC) en los diversos genomas (Chen *et al.*, 2004; Kanaya *et al.*, 1999). La variación en el UCS observada entre genomas, puede explicarse mayormente por un proceso mutacional genómico global mas que por un proceso dirigido por la selección natural. En este sentido, se ha probado claramente que el sesgo en el UCS puede predecirse a partir de cálculos estimados en base a las secuencias inter-génicas, no codificantes (Chen *et al.*, 2004). Sin embargo, la presión mutacional no puede explicar por qué en muchos casos los codones más frecuentemente utilizados a nivel global, son aquellos reconocidos por los ARN de transferencia (tRNA) isoaceptores más abundantes (Ikemura, 1985; Kanaya *et al.*, 1999; Kanaya *et al.*, 2001; Yamao *et al.*, 1991). Es importante tener en cuenta que existen explicaciones selectivas para los patrones composicionales observados en algunos casos, por ejemplo; en relación al aumento de temperatura optima de crecimiento y/o en relación al nicho aerobio *vs.* anaerobio o simbionte *vs*. vida libre (Musto *et al.*, 2004; Naya *et al.*, 2002; Rocha & Danchin, 2002), pero en estos casos el UCS sería una consecuencia del sesgo selectivo composicional y no la causa.

1.1.2 Variación del UCS a nivel intra-genómico

De acuerdo a varios trabajos el sesgo en el UCS se correlaciona fuertemente con el nivel de expresión de los genes (Gouy & Gautier, 1982; Ikemura, 1985). Esta correlación ha sido observada en procariotas y eucariotas (Akashi, 1994; Botzman & Margalit, 2011; Castillo-Davis & Hartl, 2002; Duret & Mouchiroud, 1999; Supek *et al.*, 2010), incluso se encontró que genes con similares niveles de expresión tienen sesgos correlacionados en el uso de codones sinónimos (Fraser *et al.*, 2004; Lithwick & Margalit, 2003). Aunque este patrón puede ser explicado por sesgos mutacionales particulares asociados al nivel de transcripción, esto fue descartado con estudios en *Drosophila melanogaster* y *Caenorhabditis elegans*, en donde se vio que las terceras posiciones de los codones (sinónimas) y las secuencias intrónicas no son afectadas en la forma prevista por el propuesto proceso mutacional asociado a la expresión (Duret, 2002a; Duret & Mouchiroud, 1999). Las observaciones encajan bien con los modelos seleccionistas propuestos. De acuerdo con estos, los genes que utilizan los codones reconocidos por los tRNAs isoaceptores más abundantes (codones óptimos, ver adelante), serán traducidos más eficientemente y con menos errores. De esta forma, se espera que la presión de selección sea mayor en genes que son expresados en altos niveles. La selección, operando a nivel intra-genómico, puede ir en la dirección de aumentar la velocidad de elongación y/o incrementar la concentración de ribosomas libres, y/o reducir errores. Ha sido demostrado experimentalmente que la elección de los codones sinónimos afecta, entre 4 y 9 veces, la frecuencia de incorporaciones aminoacídicas erróneas a la proteína (Precup & Parker, 1987). También se comprobó que la exactitud (fidelidad) es una propiedad asociada con la traducción de codones óptimos en *D. melanogaster* y *Escherichia coli* (Akashi, 1994; Stoletzki & Eyre-Walker, 2007). También se confirmó que la elección de los codones afecta la velocidad de elongación en la traducción (Curran & Yarus, 1989; Sorensen *et al.*, 1989). Al manipular el gen de la enzima alcohol deshidrogenasa (*Adh*) en *Drosophila* se demostró que unos pocos cambios en el uso de codones sinónimos afectaban la tolerancia al alcohol de las moscas, demostrando directamente que el uso de codones sinónimos puede afectar la función de un gen y el valor adaptativo del individuo (Carlini, 2004; Carlini & Stephan, 2003).

El modelo que parece haber ganado aceptación para explicar las variaciones a nivel intra-genómico es conocido como el modelo de balance mutación-selección-deriva (Bulmer, 1991). Este propone que la selección favorece los codones óptimos (también llamados mayores o preferidos) sobre los no-óptimos (menores o no preferidos). La presión mutacional y la deriva genética, por su parte, permiten que los codones menores persistan. Entonces, los factores característicos de cada gen, como el nivel de expresión o las restricciones funcionales, deberían determinar la intensidad de selección operando sobre los sitios sinónimos para mantener los codones mayores. Dicho de otro modo, el sesgo en el uso de codones es el resultado de la selección a favor de mutaciones que incrementan la frecuencia de codones óptimos. Postula, además, que la selección operando sobre el sesgo en el uso de codones sinónimos es relativamente débil. Este modelo, que tiene una base en la genética de poblaciones, predice una dinámica evolutiva vinculada a las diferentes clases de mutaciones sinónimas si hay selección positiva operando en un determinado gen. Mutaciones hacia un codón óptimo desde un codón no-óptimo serán beneficiosas mientras que cambios en la dirección opuesta serán deletéreos. Por lo tanto, la probabilidad de que una población aumente la frecuencia de esta mutación dependerá del tamaño poblacional efectivo (Ne), del coeficiente de selección (s) y del sesgo mutacional. El coeficiente de selección relativamente bajo, hace que los cambios en el tamaño poblacional efectivo tengan importantes efectos en las frecuencias observadas (Hershberg & Petrov, 2008).

1.1.3 Codón óptimo

El concepto de codón óptimo, como aquel codón sinónimo favorecido por la selección, se entiende generalmente bajo el modelo de selección-mutación-deriva (ver arriba). Incluso la metodología usualmente aceptada para identificar dichos codones tiene una perspectiva intra-genómica y por lo tanto se basa en la comparación de distintos grupos de genes dentro de un genoma. Más en detalle, se comparan genes conocidos por ser de alta expresión (ej. genes codificantes para proteínas ribosomales y factores de elongación) con el resto de los genes codificados en el genoma (Sharp *et al.*, 2005; Supek *et al.*, 2010; Wang *et al.*, 2011), aunque recientemente han surgido algunas discusiones sobre cuál es el mejor método para identificar dichos codones (Hershberg & Petrov, 2012; Wang *et al.*, 2011). En contraposición a este concepto, últimamente se han encontrado evidencias y se han propuesto modelos que establecen que los codones "no-óptimos" pueden ser favorecidos y mantenidos por la selección. Particularmente en ciertas regiones de ciertos genes (nivel intra-génico) (ver Sección *Variación a nivel intra-genico*).

A nivel intra-genómico, es posible discutir algunos aspectos históricos que hacen a la identificación de codones óptimos y sus propiedades. En este sentido existen modelos que intentan predecir de forma general la causa de que ciertos codones sean preferidos u óptimos.

En bacterias, en genes altamente expresados, los cuartetos y los duetos poseen preferencias opuestas para C y U en la tercera base. Una explicación para este fenómeno es la estabilización del nivel de energía para el apareamiento codón-anticodón a través de la elección en la tercera base del codón, conocido como modelo de estabilidad (Grosjean & Fiers, 1982). Las primeras dos posiciones de los codones que forman cuartetos son ricas en G y C (tal como muestra el código genético, ver Figura 1), y se observa que existe la tendencia a utilizar U en la tercera posición, lo que contribuiría a estabilizar el nivel de energía del apareamiento codón-anticodón. En el caso de los duetos, el código genético muestra que las dos primeras posiciones de los codones son pobres en G y C, observándose una preferencia al uso de C como tercera base. En cuanto a las purinas, la elección para la tercera posición es más compleja: no se observan muchas preferencias. En estos casos, al igual que para las pirimidinas, los codones con energía de unión al anticodón intermedia serían favorecidos. En el caso particular de los sextetos, los codones que forman los duetos de Arginina, Serina y Leucina no son utilizados prácticamente en genes de alta expresión. Lo que podría explicarse en base al modelo de estabilidad, ya que las energías de unión de dichos codones deberían ser demasiado débiles para una traducción rápida y de alta fidelidad. Se ha observado que si las primeras dos bases de un codón son A o U, la tercera posición preferida será C; mientras que si las primeras dos son C o G, la tercera posición será U (Grosjean & Fiers, 1982). Estos sesgos se observan en forma global en ciertos genomas pero se hacen más evidentes en los genes altamente expresados, es decir que el efecto no puede ser restringido a la fidelidad de la traducción (en el sentido de optimizar la energía de unión codón-anticodón haciéndola intermedia), sino que además está relacionado con la expresión génica. Así se planteó la existencia de una modulación de la estrategia de codificación en base al grado de expresión del mensajero.

El papel de los ARNs de transferencia (tRNAs) presentes en el genoma y la distribución de los anticodones presentes es otro aspecto que se ha considerado y se acepta que juega un papel en la definición de los codones óptimos (dos Reis *et al.*, 2004; Gingold & Pilpel, 2011; Hershberg & Petrov, 2008; Ikemura, 1985; Rocha, 2004). En este sentido se genera un modelo de predicción de codones óptimos que puede, de alguna forma, ser complementario al anterior. Con respecto a la posición de balanceo del codón, los codones reconocidos por un único tRNAs, en *E. coli* y *S. cerevisiae*, siguen determinadas restricciones en la elección del codón, las cuales están inmersas dentro de las cuatro reglas que plantea para el sesgo en el uso de codones (Ikemura, 1985). La primera regla se refiere a la disponibilidad de tRNAs (o sea la población de ARNs de transferencia isoaceptores en la célula) estableciendo que este factor restringe la elección de codones en *E. coli* y *S. cerevisiae*. En otras palabras, la primera causa de la existencia de diferentes codones óptimos estaría dada por la diferente población de tRNAs en los distintos organismos. La segunda regla está relacionada con la existencia de la uridina tiolada o la 5-carboximetil uridina (Weissenbach & Dirheimer, 1978). En genes de *E. coli* y *S. cerevisiae* se vio *in vitro* que esta base, en la posición de tambaleo del anticodón, es reconocida preferentemente por codones que terminan en A frente a codones que terminan en G. Esto es debido a que la uridina tiolada o la 5-carboximetil uridina pueden reconocer tanto A como G en la tercera posición del codón, pero reconoce mejor a la primera ya que A es el la base que aparea Watson-Crick con U. La tercera regla propone que la Inosina en la posición de tambaleo del anticodón produce a la preferencia por codones terminados en U y C frente a codones terminados en A. La Inosina (I), un derivado de la adenina, es una base modificada que puede reconocer tres bases en la posición de balanceo del codón: U, C y A, pero reconoce U y C mucho mejor que A, de ahí que la tercera regla de Ikemura establezca que la Inosina provoca una preferencia en el uso de codones de U y C sobre A. Estos serían casos particulares del efecto sobre el uso de codones de la utilización diferencial de bases modificadas en los tRNAs en distintos genomas, lo que recientemente ha cobrado mayor interés (Chan *et al.*, 2010; Novoa & Ribas de Pouplana, 2012). Existirá, entonces, una selección en contra de los codones terminados en A, particularmente en genes de alta expresión. Por último, la cuarta regla plantea que un codón del tipo (A/U)-(A/U)-Y[1], que tiene intrínsecamente una interacción débil con un anticodón en las dos primeras posiciones, tendrá una interacción óptima con el mismo si la pirimidina es C. Esta regla está establecida y adaptada del modelo de estabilidad planteado en esta misma sección (ver arriba). Anotando A o U como W (del inglés, *weak*, por enlace débil) y G o C como S (*strong*, enlace fuerte), la regla establece que codones del tipo WWY prefieren C con respecto a U y codones del tipo SSY prefieren U sobre C y los codones WSY y SWY quedan intermedios. Según este modelo, las modificaciones de nucleótidos en la posición de balanceo del anticodon de los tRNA, que difieren entre los organismos, contribuye a generar un patrón específico del uso de codones de cada organismo.

Los codones que son preferidos, en base a la combinación de estas cuatro reglas, serían los mejores para el sistema de traducción del organismo, y se conocen como codones óptimos o codones mayores. En este sentido es ampliamente aceptado que la elección de codones, específica de cada organismo, puede en parte atribuirse a las diferencias en los nucleótidos modificados en la posición de tambaleo y en la disponibilidad de tRNAs (Higgs & Ran, 2008; Novoa & Ribas de Pouplana, 2012; Rocha, 2004; Supek *et al.*, 2010).

1.1.4 Variación del UCS a nivel intra-génico

Como ya se ha mencionado anteriormente, un aspecto importante e históricamente aceptado de la selección operando a nivel de la traducción es la fidelidad, es decir la selección puede ir en la dirección de reducir errores. Ocasionalmente los emparejamientos tambaleantes o imperfectos

[1] Y = Pirimidina, C o T.

entre el codón y el anticodón pueden ocurrir, lo que puede llevar a la incorporación errónea de un aminoácido en la secuencia. Como resultado, tenemos una pérdida de energía y eficiencia. Además esto puede ser aún más negativo si la proteína que se produce es completamente no funcional o si está mal plegada.

En este sentido se ha demostrado experimentalmente que la elección de los codones sinónimos afecta la frecuencia de incorporaciones aminoacídicas erróneas a la proteína (Precup & Parker, 1987). Por lo tanto la fidelidad (exactitud o "*accuracy*") es una propiedad asociada con la traducción de codones óptimos (Akashi, 1994). Siguiendo este razonamiento es posible predecir que el sesgo en el UCS de un gen, a lo largo de su secuencia codificante, no es constante, ya que las regiones del gen que codifican para los residuos funcionalmente más importantes de las proteínas (más conservados) tenderían a presentar un sesgo hacia codones que minimicen la posibilidad de error (Akashi, 1994; Drummond & Wilke, 2008; Stoletzki & Eyre-Walker, 2007). Este aspecto de la selección a nivel de la traducción es aceptado, aunque no está clara cuál es su importancia relativa frente a la selección para aumentar la velocidad de la traducción (Ver sección Velocidad *vs*. Fidelidad en la traducción). Resulta evidente que la importancia relativa de ambos aspectos podría cambiar al analizar distintos genes y organismos (Drummond & Wilke, 2008; Ran & Higgs, 2012; Stoletzki & Eyre-Walker, 2007).

En general se acepta que los codones óptimos son al mismo tiempo aquellos que minimizan la tasa de errores y maximizan la velocidad. Por otro lado, también se acepta que las proteínas de alta expresión presentan un sesgo mayor en las regiones conservadas respecto a las no conservadas debido al efecto de la selección operando a nivel de la fidelidad de la traducción (Drummond & Wilke, 2008; Supek *et al.*, 2010). Estos aspectos son discutidos en los artículos presentados en el Capitulo I (Iriarte *et al.*, 2013a) y II (Iriarte *et al.*, 2012).

Los avances más recientes en el tema de UCS extienden el alcance de los modelos clásicos a la variación del uso de codones sinónimos dentro de un gen, haciendo innovadoras predicciones sobre las posibles fuerzas evolutivas y los mecanismos que podrían operar para definir el UCS; aunque por supuesto los nuevos modelos no limitan sus predicciones al nivel intra-génico.

Más allá de la fidelidad en la traducción, se han abordado otros aspectos o factores asociados al sesgo en el UCS. Estos otros aspectos unirían el sesgo en el UCS y la abundancia en los tRNA a la regulación de la expresión génica (Novoa & Ribas de Pouplana, 2012). Por ejemplo podemos nombrar: la autocorrelación de codones (Cannarozzi *et al.*, 2010), el agrupamiento de codones raros (Parmley & Huynen, 2009), la estructura secundaria del ARN mensajero (Shabalina *et al.*, 2013), la densidad de ribosomas (Tuller *et al.*, 2010), la presencia de regiones del tipo *Shine-Dalgarno* en la secuencia codificante (Li *et al.*, 2012), la composición a nivel de átomos (Bragg *et al.*, 2012) o las interacciones con modificaciones específicas de los tRNAs (Novoa *et al.*, 2012), entre otros.

1.1.5 Evolución del UCS

Con la introducción de la secuenciación a finales de los años 70's resultó evidente que el patrón del UCS variaba entre especies, y dentro de los genomas de las especies (ver arriba). Desde entonces se ha establecido por regla general (por lo menos en procariotas y eucariotas unicelulares) que la selección juega un rol fundamental en la persistencia de los sesgos observados en el UCS en genes de alta expresión. Se ha debatido mucho acerca de las causas por las cuales los codones traduccionalmente óptimos son seleccionados (Sharp *et al.*, 2010). En general se acepta que estos codones incrementan la eficiencia del proceso traduccional pero no está claro cómo y en qué medida (ver arriba en sección *Codon óptimo*).

Dos de las revisiones recientes más importantes han planteado la necesidad de abordar nuevas interrogantes sobre el proceso evolutivo y las causas de los sesgos en el UCS. Por ejemplo cómo es la dinámica de cambios de codones óptimos a lo largo de la evolución o cuál es el papel preponderante de los codones óptimos: fidelidad o velocidad (Hershberg & Petrov, 2008; Sharp *et al.*, 2010). Varios estimadores y trabajos han intentado cuantificar la magnitud del sesgo selectivo en el uso de codones (Botzman & Margalit, 2011; dos Reis *et al.*, 2004; Retchless & Lawrence, 2011; Sharp *et al.*, 2005; Supek *et al.*, 2010). Los resultados sugieren que existe una relación entre la tasa de duplicación poblacional (velocidad de crecimiento) y el estilo de vida con el sesgo selectivo que experimenta el organismo. Por ejemplo, la escasa variación a nivel intra-genómico en el UCS en especies parásitas y/o simbiontes y por consiguiente los pocos o nulos indicios de sesgos selectivos serían producto de los sesgos mutacionales extremos en combinación con las tasas evolutivas rápidas y la deriva génica. Esto sería suficiente para deprimir el efecto de la selección sobre el UCS, aunque no completamente (Herbeck *et al.*, 2003; Iriarte *et al.*, 2011; Moran & Wernegreen, 2000; Rispe *et al.*, 2004; Supek *et al.*, 2010). Por otro lado existe evidencia que muestra que aquellos organismos que viven en una amplia gama de hábitats tienden a presentar mayores sesgos en el uso de codones y mayor variabilidad interna (Botzman & Margalit, 2011). En los estudios desarrollados a lo largo de este doctorado hemos intentado abordar estas preguntas en *taxa* de gran importancia que cuentan, al mismo tiempo, con suficientes genomas secuenciados y suficiente información eco-fisiológica,; por ejemplo la familia *Enterobacteriaceae* y el género de hongos filamentosos *Aspergillus*.

Es claro que la identidad de los codones óptimos varía entre las especies. En general se acepta que las preferencias responden a los cambios en la composición de tRNAs en la célula (ver arriba). Si bien la abundancia ha sido medida en pocas especies, en general el valor de la concentración es estimado a partir del número de genes codificantes para tRNAs presentes en el genoma (Kanaya *et al.*, 1999). Cuando comparamos especies tan lejanas como *E. coli* o *Clostridium perfingens* parece que hay claros indicios de una co-adaptación de los codones significativamente más utilizados en los genes de alta expresión (codones óptimos) con los tRNAs que se predicen como más abundantes en ambas especies, mostrando distintos estados de preferencia solo para seis aminoácidos (o familias de codones sinónimos) (Sharp *et al.*, 2010). Esto sugiere conservación pero también un proceso de divergencia. En esto sentido no es claro cómo ocurre la divergencia entre las especies. Se ha propuesto que el sesgo mutacional condiciona el conjunto de codones óptimos en un organismo (Hershberg & Petrov, 2009; Shields, 1990), aunque no todos los autores comparten esta idea (Hershberg & Petrov, 2012; Wang *et al.*, 2011).

Es probable que el patrón de preferencias de los codones en genes de alta expresión varíe en forma relativamente lenta (Sharp *et al.*, 2010) y que alcance estados de estabilidad, co-evolucionando con el contenido de genes de tRNA en el genoma (Bulmer, 1987; Higgs & Ran, 2008). También es probable que las preferencias actuales representen mejor un estado de adaptación al *pool* de tRNA ancestral y mayormente conservado, como ha sido demostrado en *E. coli* (Withers *et al.*, 2006). De ser así se espera que las inserciones o deleciones recientes de los genes de tRNA tengan efectos menores sobre el patrón de codones óptimos observados.

El papel de los cambios en las dinámicas poblacionales podría también ser importante. Períodos con prolongados cuellos de botella resultarían en una reducción del efecto de la selección, incluso en genes de alta expresión; seguido por un aumento en el tamaño poblacional, se daría lugar a un incremento del sesgo selectivo con un nuevo patrón de preferencias. Esto sería posible, siempre y cuando exista un cambio en los niveles de expresión o en la cantidad de los tRNA (anticodones) presentes en el genoma (Hershberg & Petrov, 2008).

Los mismos autores plantean que es posible que la adquisición de un gen clave para la sobrevivencia o el aumento de la eficacia darwiniana dirija el cambio en el patrón de codones óptimos. La adquisición de este nuevo gen, que presenta diferente uso de codones, podría generar una presión de selección para cambiar el nivel de expresión de ciertos genes de tRNA. En este caso el resto de los genes acompañaría el cambio en forma posterior (Hershberg & Petrov, 2008).

Al momento de escribir esta tesis no se tiene una idea clara de con qué frecuencia cambia la preferencia por codones óptimos, ni el cómo, ni el por qué.

1.1.6 Breve introducción al UCS en virus

Los virus son parásitos intracelulares que co-evolucionan con y se adaptan al hospedero. Varios son los mecanismos responsables de esta adaptación. Se ha propuesto que el UCS es uno de ellos. El UCS ha sido estudiado en virus, principalmente analizando el sesgo global del genoma y de genes específicos (Wong *et al.*, 2010). Los virus son completamente dependientes de la maquinaria de traducción de la célula hospedera para sintetizar sus propias proteínas. De esta forma se espera que el UCS de los transcriptos del virus reflejen el *pool* de tRNAs del hospedero (Bahir *et al.*, 2009) al igual que el propio UCS del hospedero (Ikemura, 1981, 1985). Sin embargo varios estudios han demostrado que esto no parece ser la regla general para los virus, ya que en algunos casos los sesgos parecen estar pobremente adaptados o no adaptados en absoluto a la abundancia de tRNAs en la célula (Ahn & Son, 2012; Goni *et al.*, 2012; Moratorio *et al.*, 2013; Roychoudhury *et al.*, 2011; Wong *et al.*, 2010).

La variación en el UCS está regida por la mutación, la selección y la deriva genética, aunque con distinta forma en los distintos organismos y genes (Sharp *et al.*, 2010; Sharp *et al.*, 2005). En este sentido, entender la relación entre estas fuerzas, las causas y consecuencias de los sesgos observados en virus es importante para comprender la evolución de estos organismos, particularmente la interrelación entre los virus y hospedero (Shackelton *et al.*, 2006). En algunos grupos de virus se ha establecido que la selección operando en el UCS jugaría un papel importante, considerando por ejemplo la interacción del virus con el ambiente del hospedero o con la estructura secundaria del RNA (Aragones *et al.*, 2010; Carbone, 2008; Firth & Brierley, 2012; Jenkins & Holmes, 2003; Karlin *et al.*, 1994; Lucks *et al.*, 2008; Pavon-Eternod *et al.*, 2013; Stedman *et al.*, 2013; Young *et al.*, 2013). Si bien parece haber un consenso general con la idea de que la selección juega un rol en el sesgo que observamos en el UCS en virus, no esta claro como la selección contribuye al mismo en relación a la deriva y, principalmente, al sesgo mutacional (Bishal *et al.*, 2013). El sesgo (o presión) mutacional ocurre como consecuencia de errores en la replicación que ocurren de forma asimétrica entre los posibles cambios. En el caso de los virus de RNA, la tasa de sustitución por base en cada ronda de replicación se estima entre 10^{-4} y 10^{-5} (Domingo, 1997). En vista de esto, es lógico pensar en un rol preponderante para el sesgo mutacional, como ocurre en otros parásitos (Rocha & Danchin, 2002).

Por otro lado, resulta interesante que algunos genomas virales contengan genes de tRNAs, los cuales serían agregados al *pool* de tRNAs de la célula. Algunos trabajos han sugerido que la presencia de estos genes es el resultado de la selección durante la evolución, probablemente porque su presencia redirigiría la maquinaria celular para aumentar la eficiencia de la transcripción de los genes virales (Bailly-Bechet *et al.*, 2007). Otra alternativa es que estas moléculas le permitan al virus infectar un amplio espectro de hospederos con variados sesgos en el UCS (Gingold & Pilpel, 2011). Por otro lado, un trabajo recientemente publicado ha sugerido que, contrario a lo esperado, algunos virus (incluyendo *Influenza A*) podrían seleccionar los tRNA que optimizan su traducción mas que ajustar su sesgo al *pool* de tRNAs de la célula (Pavon-Eternod *et al.*, 2013).

En el Capítulo III analizamos el sesgo en el UCS en virus con genomas de RNA con polaridades opuestas. Cuantificamos los sesgos y la variabilidad entre cepas. Estudiamos la relación con los tRNAs del hospedero y los factores composicionales asociados a los sesgos encontrados. Se discute el papel de la mutación y de la selección en relación a la posible adaptación al *pool* de tRNAs del hospedero.

1.2 Uso de Aminoácidos (UAA):

Durante 30 años se creyó que solamente las restricciones funcionales de una proteína dictaban el ritmo (o tasa) de evolución de la secuencia, esto es conocido como la “hipótesis funcional”. De acuerdo a esta idea, los genes esenciales evolucionan lentamente en relación al resto del genoma por efecto de la selección purificadora, ya que pequeños cambios en la secuencia proteica resultarían deletéreos para el organismo (Park & Choi, 2009). Se han planteado otras ideas en el tema. Las tasas de evolución proteica son usualmente cuantificadas por el número de cambios nucleotídicos no-sinónimos por sitios no-sinónimos (dN); observándose al mismo tiempo que el nivel de expresión de un gen es el mejor predictor de dN (Duret & Mouchiroud, 2000; Pal *et al.*, 2001, 2003; Rocha & Danchin, 2004; Subramanian & Kumar, 2004). Los resultados sugieren que los genes de alta expresión son sistemáticamente más conservados a nivel no-sinónimo, por lo que el nivel de expresión jugaría un rol a la hora de explicar el nivel de conservación de los genes. Este hecho parece ser universal, desde bacterias a mamíferos. La influencia de este fenómeno en la variación intra-genómica observada, es particularmente importante en aquellos organismos que han experimentado cambios en las frecuencias aminoacídicas globales. Dicho de otro modo, las tendencias globales serían menos notables en genes de alta expresión (más conservados), ocurriendo como respuesta directa o indirecta a fuerzas selectivas o neutrales (Lightfield *et al.*, 2011; Singer & Hickey, 2000).

Actualmente se acepta que el patrón de conservación diferencial podría, en parte, tener una vinculación con el costo bio-sintético de los aminoácidos y la selección operando sobre el UAA (Heizer *et al.*, 2011; Wagner, 2005).

Dado que para la mayoría de los organismos procariotas la biosíntesis de aminoácidos representa una porción significativa del costo total energético y que existe una variación muy importante en cuanto al costo de los distintos aminoácidos (el costo de producir el aminoácido glutamato es 9,5 uniones de fosfato de alta energía contra 75,5 en el caso del Triptofano), no sería sorprendente, entonces, que exista un sesgo selectivo que favorezca el uso preferencial de aminoácidos menos costosos (Akashi & Gojobori, 2002). Esto, a su vez, debería ser más pronunciado en genes de alta expresión y como consecuencia se generaría un fenómeno de variación a nivel intra-genómico.

Según esta hipótesis la selección opera decisivamente en la composición del proteoma, lo que ya se ha comprobado en organismos filogenéticamente no relacionados (Akashi & Gojobori, 2002; Seligmann, 2003; Zavala *et al.*, 2002). Lo que se observó, como regla general, es una correlación inversa entre el costo bio-sintético relativo del producto proteico y el nivel de expresión de los genes. Dichas observaciones se han extendido a otros organismos, procariotas y eucariotas, con variedad de preferencias ecológicas (Heizer *et al.*, 2011; Heizer *et al.*, 2006; Seligmann, 2003).

Entender los principales factores que afectan las fuentes de variación a nivel intra- e inter-genómico en el UAA es un tema principal en evolución molecular. En este sentido, el análisis de correspondencias sobre el uso relativo de aminoácidos (COA-RAAU) ha sido usualmente aplicado a proteomas. Mediante este procedimiento se encontró en *E. coli* que los tres factores más importantes que gobiernan la variaron aminoacídica intra-genómica son la hidrofobicidad, el nivel de expresión y la aromaticidad promedio de las proteínas (Lobry & Gautier, 1994). Los autores proponen que restricciones a nivel de la traducción, que se hacen más patentes en genes de alta expresión, afectan la composición global de las proteínas. En *Giardia lamblia*, un eucariota unicelular, se encontró que los factores más relevantes estaban relacionados con mecanismos particulares de defensa contra especies reactivas del oxígeno (Garat & Musto, 2000). Al mismo tiempo, los autores observaron que los genes de alta expresión presentaban una clara tendencia a utilizar aminoácidos más pequeños, de lo cual se dedujo que la economía celular es otro factor importante (ver arriba). Desde entonces muchos trabajos han explorado los posibles efectos de presiones selectivas sobre el UAA a nivel global e intra-genómico en una variedad de organismos procariotas y eucariotas (Akashi & Gojobori, 2002; Basak *et al.*, 2004; Kahali *et al.*, 2007; Naya *et al.*, 2004; Zavala *et al.*, 2002).

El efecto de la selección operando sobre el uso de aminoácidos es discutido en el caso de las bacterias simbiontes (Das *et al.*, 2005; Herbeck *et al.*, 2003; Palacios & Wernegreen, 2002; Rispe *et al.*, 2004; Schaber *et al.*, 2005; Seligmann, 2003). Algunos autores creen que el sesgo en el uso de aminoácidos en genes de alta expresión ocurre porque estos genes “resisten” el enriquecimiento con residuos aromáticos y/o codificados por familias de codones ricos en A y T. De este modo, la mayor conservación característica de los genes de alta expresión desde la divergencia de sus parientes de vida libre explicaría la resistencia observada al enriquecimiento con residuos ricos en A y T (Banerjee *et al.*, 2004b; Herbeck *et al.*, 2003; Rispe *et al.*, 2004). Otros autores, sin embargo, argumentan que la selección también opera favoreciendo aminoácidos codificados por familias de codones ricos en G y C (Schaber *et al.*, 2005). Paralelamente se ha observado que la cantidad total de tRNA específico para un determinado aminoácido se correlaciona con la composición aminoacídica en proteínas ribosomales de *Mycoplasma capricolum* (Yamao *et al.*, 1991). Esto implicaría que los sesgos observados en el UAA en proteínas de alta expresión esta dirigido por la selección, siendo la cantidad relativa de tRNAs en una célula el factor principal.

Las hipótesis neutralistas proponen que el sesgo global en el UAA es producto del sesgo mutacional principalmente (a nivel inter-genómico), lo que ha sido observado en la mayoría de los organismos (Lightfield *et al.*, 2011; Singer & Hickey, 2000). Sin embargo esto no es cierto en todos los casos, por ejemplo, se ha demostrado que los sicrófilos y termófilos presentan sesgos hacia el uso de ciertos aminoácidos (Saunders *et al.*, 2003; Singer & Hickey, 2003; Tekaia & Yeramian, 2006; Tekaia *et al.*, 2002).

Por otro lado, a nivel intra-genómico predominan las hipótesis seleccionistas de minimización del costo. En el Capítulo V abordaremos, mediante una perspectiva comparativa, el estudio del sesgo selectivo en el UAA, sus causas y consecuencias, y las tendencias de variación intra-genómicas en relación a la hipótesis de minimización del costo a lo largo de la evolución en Mollicutes, un grupo particular de bacterias parásitas (Iriarte *et al.*, 2013b).

2 Objetivos

2.1 Objetivo General

Reconstruir estados ancestrales de sesgos observados a nivel intra-genómico en el UCS y UAA en *taxas* de interés. Analizar las tendencias en un marco filogenético.

2.2 Objetivos Específicos

- Trabajar con *taxas* que presenten suficiente número de especies secuenciadas e información eco-fisiológica. Muestrear los genomas para evitar la sobre representación de ciertos géneros. Reconstruir relaciones filogenéticas robustas en base a información de múltiples genes.
- Definir la existencia de sesgos selectivos e identificar los codones óptimos. Estimar los coeficientes de selección en los organismos actuales.
- Analizar de forma independiente la preferencia por codones sinónimos (y el sesgo selectivo) a nivel de la traducción para velocidad y para fidelidad.
- Mapear los cambios en los codones óptimos y el coeficiente de selección en el UCS por métodos Bayesianos y por Máxima Verosimilitud.
- Analizar el papel de los cambios en la composición de tRNAs en los genomas.
- Reconstruir las secuencias de proteínas ancestrales y estimar tendencias en sesgos observados en los genes de alta expresión de ciertas propiedades importantes: peso molecular, aromaticidad, costo metabólico, etc. Contrarrestar las hipótesis propuestas para explicar la variación a nivel intra-genómico en organismos parásitos o endosimbiontes.

3 Estrategia

Se utilizó una aproximación puramente bioinformática, aprovechando la disponibilidad de genomas completamente secuenciados y de libre acceso. Se implementaran un conjunto de herramientas que se integraron en nuestro laboratorio. Esto incluyó el manejo de varios programas, paquetes de programas y lenguajes de programación. Se adaptaron y desarrollaron índices simples para medir y comparar la magnitud de los sesgos. Se trabajó sobre genes ortólogos, reconstruyendo estados ancestrales sobre relaciones filogenéticas robustas inferidas en base a secuencias aminoacídicas.

3.1 Uso de codones sinónimos (UCS)

Se estudiaron aspectos de la evolución de los sesgos relacionados a la velocidad y fidelidad de la traducción. Para estimar los componentes de la selección operando a nivel de la velocidad se asumió que los genes de alta expresión (genes conservados codificantes para proteínas ribosomales y factores de elongación) presentan sesgos selectivos más pronunciados con respecto al resto de los genes del genoma y al resto de los genes ortólogos analizados. Se estimaron componentes de la selección operando a nivel de la fidelidad de la traducción asumiendo que las regiones conservadas de los genes presentan sesgos selectivos con respecto a las regiones no conservadas.

Partiendo del grupo o *set* de genes ortólogos y en base a la secuencia de las proteínas codificadas, se reconstruyeron las relaciones filogenéticas en forma robusta. Se utilizaron las secuencias de estos mismos genes para calcular el índice de selección de Sharp (Sharp *et al.*, 2005) para cada organismo. De esta forma eliminamos los efectos de la transferencia horizontal, duplicaciones recientes o pseudogenes, utilizando los mismos genes en los distintos organismos.

Se reconstruyó el índice de selección en los nodos ancestrales como indicador directo del sesgo, y las divergencias sinónimas como indicador indirecto del papel de la selección operando a nivel de la traducción a lo largo de todo el grupo. Como regla general, se espera que, si hay selección purificadora operando manteniendo un sesgo conservado, aquellos genes de alta expresión presenten una divergencia sinónima menor que el resto de los genes en el genoma (Sharp & Li, 1987).

Codones óptimos: La diferencia significativa ($\chi 2$) en el uso de un codón sinónimo por parte de los genes de alta expresión puede ser entendida como un estado del carácter "codón sinónimo". Esta diferencia puede existir (1) o no existir (0), de manera que el carácter codón sinónimo, puede cambiar entre estos dos estados posibles. Obtenemos entonces, para un conjunto de organismos, una tabla con datos binarios (0-1), reflejando la presencia o ausencia de codones óptimos en cada organismo. Dichos cambios de estados en cada carácter fueron mapeados en la filogenia mediante inferencia Bayesiana.

Se predijo la presencia de genes de tRNA y sus anticodones. Se utilizó el número de genes como indicador indirecto de la concentración celular (Kanaya *et al.*, 1999).

Se define que existe un efecto de la selección operando a nivel de la velocidad de traducción cuando: i) la expresión es un factor principal a la hora de explicar la variabilidad en el UCS (factor principal de análisis multivariado), ii) los genes de alta expresión (*set* de referencia) presentan un sesgo mayor en el UCS, iii) los genes de alta expresión presentan una divergencia sinónima menor (medida como dS) y iv) los genes de alta expresión presentan una preferencia por los cuatro codones universalmente óptimos (UUC, AUC, UAC y AAC) definidos por Sharp (Sharp *et al.*, 2005). Los sesgos en regiones conservadas, o las preferencias de codones óptimos en las regiones conservadas de los genes con respecto a las no conservadas son indicativos de selección operando a nivel de la traducción. La correlación de los codones significativamente más utilizados con los tRNAs isoaceptores más abundantes es señal de que la selección esta operando a nivel de la traducción tanto para velocidad como para fidelidad.

3.2 UCS en virus

Se estudiaron los sesgos en el uso de codones sinónimos a nivel global de dos virus de RNA monocatenario, uno con polaridad positiva y otra negativa. Se analizaron el conjunto de genomas completamente secuenciados. Utilizando métodos “clásicos” se estudió la variabilidad inter-genómica entre cepas de la misma especie. Estos métodos incluyen análisis multivariados, predicción de tRNAs en los hospederos en los que hay información disponible, estadísticos composicionales (composición nucleotídica, frecuencia de dinucleótidos) del genoma viral y el hospedero, entre otros.

3.3 Uso de aminoácidos (UAA)

Mediante análisis multivariado se analizaron los factores más importantes para explicar la variabilidad intra-genómica en cada organismo. Se desarrolló un índice simple para calcular y comparar el sesgo en el uso de cada aminoácido en cada organismo. Se compararon los patrones de un grupo de bacterias parásitas y un grupo hermano de vida libre. Mediante reconstrucción de secuencias ancestrales pudieron inferirse los cambios aminoacídicos ocurridos a lo largo de la evolución del grupo. Con estos resultados se resumieron las tendencias promedio de los genes de alta expresión y del resto de genoma para algunas de las propiedades en cuestión: aromaticidad, hidrofobicidad y peso molecular. Las hipótesis alternativas para explicar los sesgos observados en los genes de alta expresión en otros grupos de parásitos y endosimbiontes fueron puestas a prueba.

4 Capítulo I: Evolución del UCS en la familia Enterobacteriaceae

Evolution of optimal codon choices in the family Enterobacteriaceae

Iriarte A, Baraibar J, Romero H, Castro S, Musto H.

Microbiology 2013, 159:554-563

Resumen:

Enterobacteriaceae es una gran familia de Proteobacterias, que incluye algunos de los géneros más conocidos y estudiados dentro de procariotas: *Escherichia*, *Salmonella* y *Yersinia*. Muchas de las ideas más importantes de la evolución del uso de codones sinónimos y selección operando a nivel de la traducción fueron profundamente influenciadas por estudios con estos grupos bacterianos. En este trabajo se describe el análisis del patrón de uso de codones en genomas completamente secuenciados pertenecientes a esta familia. Los efectos de la selección operando sobre la velocidad y precisión en la traducción se describen y comparan, así como las tendencias filogenéticas dentro del grupo. Se identificaron los codones preferidos (u óptimos) para todas las especies y se estudió la dinámica evolutiva de estas preferencias. Siguiendo un enfoque bayesiano se mapea la evolución de los codones óptimos hasta el ancestro común de la familia. Hemos encontrado que hay un cierto nivel de variación en la selección entre los microorganismos analizados que podría estar asociados con tendencias linaje-específicas. Sin embargo se observa que existe una importante conservación en el sesgo en el uso de codones a través de la evolución de la familia, tanto en genes altamente expresados, como en las regiones conservadas de proteínas, lo que sugiere un rol importante de la selección negativa. En este sentido, los resultados del análisis de los tRNAs apoyan la idea de que los sesgos observados se han mantenido, en muchos casos, ajustados al conjunto ancestral de tRNAs.

Microbiology (2013), **159**, 000–000 DOI 10.1099/mic.0.061952-0

Evolution of optimal codon choices in the family *Enterobacteriaceae*

Andrés Iriarte,[1,2,3] Juan Diego Baraibar,[1] Héctor Romero,[1] Susana Castro-Sowinski[4] and Héctor Musto[1]

Correspondence
Héctor Musto
hmusto@gmail.com

[1]Laboratorio de Organización y Evolución del Genoma, Facultad de Ciencias (UDELAR), Iguá 4225, 11400 Montevideo, Uruguay

[2]Laboratorio de Evolución, Facultad de Ciencias (UDELAR), Iguá 4225, 11400 Montevideo, Uruguay

[3]Área Genética, Depto. de Genética y Mejora Animal, Facultad de Veterinaria (UDELAR), Av. A. Lasplaces 1550, CP 11600, Montevideo, Uruguay

[4]Sección Bioquímica y Biología Molecular, Facultad de Ciencias (UDELAR), Iguá 4225, 11400 Montevideo, Uruguay

Received 5 July 2012
Revised 13 December 2012
Accepted 2 January 2013

The *Enterobacteriaceae* are a large family of Proteobacteria that include many well-known prokaryotic genera, such as *Escherichia*, *Yersinia* and *Salmonella*. The main ideas of synonymous codon usage (CU) evolution and translational selection have been deeply influenced by studies with these bacterial groups. In this work we report the analysis of the CU pattern of completely sequenced bacterial genomes that belong to the *Enterobacteriaceae*. The effect of selection in translation acting at the levels of speed and accuracy, and phylogenetic trends within this group are described. Preferred (optimal) codons were identified. The evolutionary dynamics of these codons were studied and following a Bayesian approach these preferences were traced back to the common ancestor of the family. We found that there is some level of variation in selection among the analysed micro-organisms that is probably associated with lineage-specific trends. The codon bias was largely conserved across the evolutionary time of the family in highly expressed genes and protein conserved regions, suggesting a major role of negative selection. In this sense, the results support the idea that the extant CU bias is finely tuned over the ancestral well-conserved pool of tRNAs.

INTRODUCTION

The *Enterobacteriaceae* are a group of closely related facultative anaerobic bacterial genera, distributed worldwide in soil, water, plants and animals. This group belongs to the Gamma subdivision of the phylum Proteobacteria and contains at least 44 genera and 176 designated species. Phenotypic traits, DNA–DNA hybridization, 16S rDNA and housekeeping gene sequences analyses (Paradis *et al.*, 2005; Pham *et al.*, 2007), among many other markers, have been used in the phylogenetic study and identification of enterobacterial species. Among the well-known members are some pathogenic strains of *Escherichia coli* and *Salmonella enterica*, which can cause severe gastrointestinal disorders (Toth *et al.*, 2006). In addition to animal pathogens, this group includes important plant pathogens, such as *Erwinia carotovora* and *Erwinia amylovora*, and obligately host-associated bacteria, such as the aphid mutualist *Buchnera aphidicola* and the tsetse fly endosymbiont *Wigglesworthia glossinidia.*

Members of the family *Enterobacteriacea* have been used as model micro-organisms to develop ideas on molecular evolution. In particular, the evolution of synonymous codon usage (CU) and the translational selection hypothesis were deeply influenced by studies in *E. coli* [among other classical papers, see Grantham *et al.* (1980) and Ikemura (1985)]. From a historical point of view, *E. coli* is probably the most studied micro-organism, and it has been the model in microbiology and molecular genetics, among other areas, for more than a century. Furthermore, the first study of the estimation of synonymous and non-synonymous distances in relation to CU bias was made using orthologous sequences from *E. coli* and *Salmonella. typhimurium* (Sharp & Li, 1987b).

Abbreviations: BFT, Bayes Factor Test; CAI, Codon Adaptation Index; CR, conserved region; CU, codon usage; dS, synonymous substitution rate; HEG, highly expressed genes; LRCA, last recent common ancestor; MELP, MILC-based expression level predictor; NCR, non-conserved regions; WCA, within-group correspondence analysis.

Four supplementary tables and five supplementary figures are available with the online version of this paper.

Synoymous CU is known to vary between and within genomes. This characteristic has been explained in terms of natural selection and mutational bias. Natural selection has an impact in translation, acting at the levels of speed and accuracy, while mutational bias is associated with the substitution pattern characteristic of each species (for a recent review, see Sharp *et al.*, 2010). As a consequence of the action of selection, the frequency of 'optimal' codons is increased due to the need to maximize the speed of elongation (selection for translation speed) or to minimize the incorporation of wrong amino acids into conserved positions (selection for translation accuracy) (Hershberg & Petrov, 2008).

The increasing number of complete sequenced bacterial genomes allows us to develop studies designed to understand the evolutionary dynamics of optimal codon choices. In particular, there are dozens of members of the family *Enterobacteriaceae* that have been completely sequenced, including strains of the same species, and their physiological and ecological traits have been reasonably well characterized.

Taking advantage of this knowledge, we aimed to analyse the CU biases in members of *Enterobacteriaceae* from an evolutionary perspective.

METHODS

Sequence data. Genome coding sequences of *Enterobacteriaceae* and *Pasteurella multocida* (as outgroup) were obtained from ftp.ncbi.nih.gov. In order to prevent over-representation of some genera, such as *Escherichia*, *Shigella*, *Salmonella* and *Yersinia*, we used a preliminary analysis of molecular distance, discarding indistinguishable species and strains. To do so, the mean nucleotide sequence divergence across the 16S rDNA gene was estimated. A criterion of at least 2 % sequence divergence was used for inclusion of species. In addition, in the case of highly divergent species, for which multiple strains have been sequenced, only one representative was selected. The final dataset comprised 28 genomes, distributed in 19 genera as shown in Table 1. Two enterobacterial parasitic species were included in the analyses. However, note that these species present high evolutionary rates and that the effect of selection for translation is still controversial (Herbeck *et al.*, 2003; Rispe *et al.*, 2004; Wernegreen & Funk, 2004).

Identification of orthologues and phylogenetic reconstruction. Putatively orthologous sequences among all species analysed were identified running a BLAST query of the whole set of proteins, following the best-reciprocal-hit (BRH) criteria with a minimal identity value of 40 % and a minimal e-value of 1×10^{-2}. Groups of orthologues were assembled by searching for triangles of reciprocality. According to this method a group of sequences were defined as orthologues when the BRH results were exclusively limited to sequences of their own set. As a result, 216 groups of orthologous genes were identified among the 28 complete sequenced genomes. Coding sequences were translated into proteins by means of the transeq program implemented in the EMBOSS package (Rice *et al.*, 2000). The protein sequences were aligned using MUSCLE (Edgar, 2004), and subsequently poorly aligned regions and gaps were removed using Gblock with default parameters (Talavera & Castresana, 2007). Phylogenetic trees were inferred using the maximum-likelihood method with an amino acid LG + G model by means of Phyml version 3.0 (Guindon & Gascuel, 2003). The default SH-like test was used to evaluate branch supports. Finally, a consensus tree was inferred from the 216 phylograms using the sumtrees.py program (Sukumaran & Holder, 2010).

Identification of orthologous genes and alignment for molecular distances estimation, conserved region (CR) classification and CU analyses was performed as described above, but considering separately the highly divergent *Buchnera aphidicola* and *Wigglesworthia glossinidia*. As a result, 322 groups of orthologous sequences were assembled for these two organisms and 727 for the remaining species.

Estimation of molecular distances. The retrieved aligned proteins were back-translated to the known DNA sequences by means of the tranalign program implemented in the EMBOSS package (Rice *et al.*, 2000). Pairwise synonymous distances were estimated using the codeml program, included in the PAML 4.4 package (Yang, 2007). This analysis was performed with pairs of sequences displaying synonymous substitution rate (dS) values $\leqslant 2.0$ (Table S1).

Selection effect, CU analyses and t-RNA prediction. Conditions required to ascertain whether there is a significant effect of selection at the level of translation were tested as described in Iriarte *et al.* (2011). These results were confirmed by means of the estimation of the strength of selection on CU as developed by Sharp *et al.* (2005). The orthologous sequences were used to analyse CU (322 and 727 groups as indicated above). Subsequently, analyses were repeated with all genes in order to confirm orthologue-based results. Codon count, base composition and relative synonymous codon usage (RSCU) (Sharp & Li, 1986) were calculated using the program CodonW 1.4.2 (sourceforge.net/projects/codonw/). Within-group correspondence analysis (WCA) (Charif *et al.*, 2005; Suzuki *et al.*, 2008) was calculated as implemented in the ADE4 package of R (Thioulouse *et al.*, 1997). For each gene, the effective number of codons (ENC′) (Novembre, 2002), Codon Adaptation Index (CAI) (Sharp & Li, 1987a) and the MILC-based expression level predictor (MELP) values (Supek & Vlahovicek, 2005) were computed using INCA2.1 (Supek & Vlahovicek, 2004). In all cases, 44 conserved ribosomal proteins genes were used as the reference set of highly expressed genes (HEG). The CU skew (Δ) for each triplet (i) in sequences of the group *A* compared with sequences of the group *B* was defined as:

$$\Delta_{(i)}=\mathbf{Fr}\ A_{(\mathbf{i})}-\mathbf{Fr}\ B_{(\mathbf{i})}/\mathbf{Fr}\ A_{(\mathbf{i})}+\mathbf{Fr}\ B_{(\mathbf{i})}$$

where **Fr** is the relative frequency of the codon **i** among its synonymous codon family. Note that group *A* was composed of the reference set and group *B* was composed of the rest of the orthologous genes when estimating Δ in HEG. Similarly, group *A* was composed of CRs and group B was composed of non-conserved regions (NCRs) when estimating Δ in a CR.

The number of tRNA genes was predicted using the tRNAscan-SEM program, version 1.3 (Lowe & Eddy, 1997) in mixed (general tRNA model) mode, using the default parameters.

Contingency $\chi2$ tests were used to identify triplets that are significantly differentially used ($P<0.01$) between the reference set and the rest of the orthologous genes. Triplets significantly more frequent in the reference set were considered as translationally optimal.

Analysis of CU in conserved protein regions. CR and NCR among orthologues were identified and split using Gblock with the following parameters: b1=X, b2=X, b3=1 and b4=2, where X represents the number of sequences in the alignment (for details, see Talavera & Castresana, 2007). Blocks with a length of at least two amino acids, unchanged in all species and with a maximum of one contiguous non-conserved position, were pooled and used for subsequent studies. The percentage of CR-aligned positions was

Table 1. Micro-organisms studied in this work and principal results

Non-significant correlations are shown in bold type ($P>0.001$).

Micro-organism	GC%	Var*	MELP†	GC3s‡	dS§	*S*‖	Accession no.
Cronobacter turicensis	58.6	24.5	0.93	0.85	0.59	1.02	NC_013282
Dickeya dadantii	56.3	14.4	0.93	0.55	0.62	0.63	NC_012880
Edwardsiella ictaluri	58.9	19.0	0.93	0.80	0.54	1.16	NC_012779
Enterobacter aerogenes	56.3	23.7	0.89	0.74	0.57	1.18	NC_015663
Enterobacter cloacae	58.5	25.2	0.91	0.87	0.56	0.95	NC_014618
Erwinia amylovora	54.7	15.6	0.93	0.72	0.60	0.93	NC_013971
Escherichia blattae	57.7	26.2	0.92	0.90	0.61	1.14	NC_017910
Escherichia coli	52.0	25.7	0.89	0.55	0.63	1.38	NC_017633
Pantoea ananatis	54.8	19.1	0.93	0.77	0.67	1.19	NC_013956
Pantoea sp.	55.9	22.0	0.91	0.75	0.61	1.24	NC_014837
Pectobacterium atrosepticum	52.3	18.1	0.91	0.55	0.59	0.93	NC_004547
Pectobacterium carotovorum	53.3	19.5	0.92	0.64	0.59	0.95	NC_012917
Photorhabdus asymbiotica	43.6	15.1	0.88	0.12	0.52	0.97	NC_012962
Photorhabdus luminescens	44.2	15.1	0.67	0.20	0.51	0.95	NC_005126
Proteus mirabilis	40.2	21.7	0.93	0.60	0.61	1.56	NC_010554
Providencia stuartii	42.8	20.9	0.91	0.71	0.58	1.42	NC_017731
Rahnella aquatilis	53.5	20.2	0.91	0.68	0.62	0.97	NC_017047
Salmonella bongori	52.5	22.6	0.89	0.58	0.43	1.44	NC_015761
Salmonella enterica	53.6	24.2	0.89	0.64	0.43	1.49	NC_011274
Serratia proteamaculans	56.5	19.7	0.90	0.77	0.56	0.85	NC_009832
Sodalis glossinidius	56.3	12.9	0.82	0.78	0.52	0.72	NC_007712
Xenorhabdus bovienii	46.4	17.9	0.75	0.37	0.57	1.18	NC_013892
Xenorhabdus nematophila	45.5	20.7	0.91	0.32	0.66	1.12	NC_014228
Yersinia enterocolitica	48.2	20.0	0.93	0.57	0.67	1.23	NC_017564
Yersinia pestis	49.0	17.9	0.93	0.53	0.69	1.07	NC_004088
Buchnera aphidicola	25.3	7.6	0.68	**0.02**	**0.05**	−0.13	NC_017259
Wigglesworthia glossinidia	23.7	4.7	0.35	0.36	**0.12**	−0.06	NC_004344
Pasteurella multocida (OG)	41.0	16.7	0.88	0.73	Not estimated	1.27	NC_017027

*The of variability explained by the first axis generated by WCA is shown. This axis was always related to the expression level with the exception of *D. dadantii* (second axis) and *B. aphidicola* and *W. glossinidia*, where no axis was found to be related to expression.

†Pearson's correlation coefficients (*r*) of the positions of the genes on the 'expression' axis of WCA against the respective MELP values.

‡Pearson's correlation coefficient (*r*) between GC3s and the positions of the genes on the axis related to expression.

§Minimum absolute Pearson's correlation coefficients (*r*) of the dS estimation for orthologues with their respective MELP values (Tables S1 and S2). Note that all these coefficients are negative.‖Estimation of the strength of selected CU bias (*S*) as developed by Sharp *et al.* (2005).

45.9 %. The robustness of the method to define CR was tested using pairwise p-distances estimation with the default protein model, using the distmat program implemented in the EMBOSS package. The p-distance is the proportion (p) of amino acid sites at which the two sequences compared are different. Global CU in CR and NCR was also compared for each species using contingency $\chi 2$ tests and CU skew.

Ancestral inference and phylogenetic inertia. Under our model, each synonymous codon adopts only two mutually exclusive states: 'significantly preferred' (1) or 'non-significantly preferred' (0), as described above. Thus, the evolution of preferences was mapped across the phylogeny (fully resolved phylogram) defined in advance. Reconstructions were performed following the Bayesian inference method by means of BayesTraits v1.0 (Pagel *et al.*, 2004). States of internal nodes were reconstructed by local approximation, using a reversible-jump MCMC model with 50 million iterations (Pagel & Meade, 2006; Pagel *et al.*, 2004). In this case the Markov chain searches the posterior distribution of the different models of evolution as well as the posterior distributions of the parameters of these models. The Bayes Factor Test (BFT) was used to compare alternative hypotheses (presence/absence of optimal codons in each node) as suggested by the BayesTraits manual (www.evolution.rdg.ac.uk). A BFT value >2 was taken as 'positive' evidence, >5 as 'solid' evidence, and >7 as 'strong' evidence for support of one model over the other. Codons infered to have changes are marked with '*', '**' and '***' in Fig. 1, respectively (adapted from McGaughran & Holland, 2010).

A Bayesian approach for continuous characters was also used to reconstruct ancestral selection coefficients by means of BayesTraits. Alternative models were compared using BFT as suggested by the BayesTraits manual. Posterior distributions of model and scaling parameters were created. The MCMC analyses were run for 100 million iterations. The inferences were based on samples taken after 25 % of the MCMC cycle. In all cases, convergence statistics were checked by means of Tracer 1.5 (Rambaut & Drummond 2007).

Following the method developed by Vieira-Silva & Rocha (2008), the phylogenetic inertia associated with optimal codon choices and

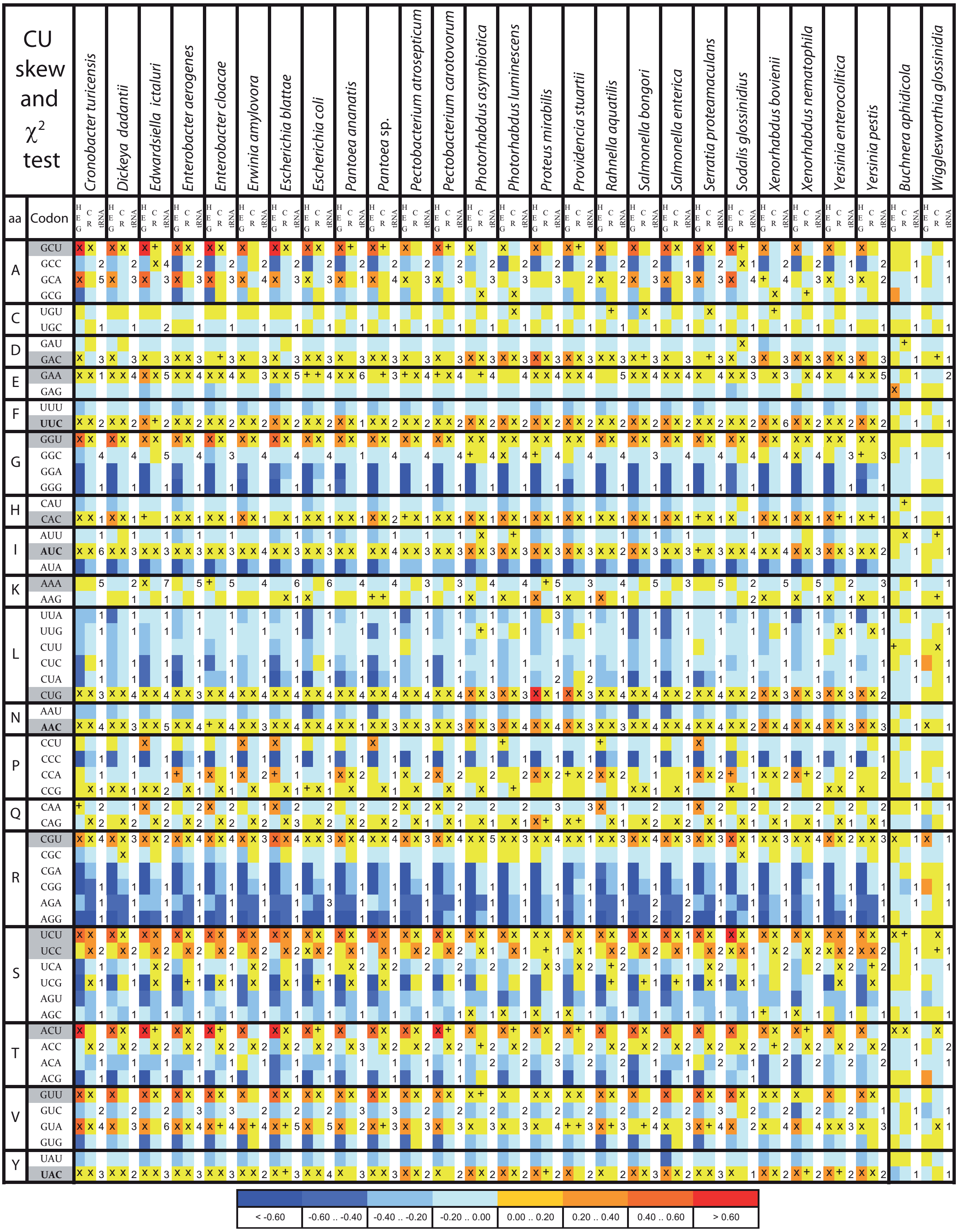


Fig. 1. CU skew (Δ) of HEG and CR (protein CRs) estimated for each organism. Contingency $\chi2$ tests were used to find which

triplets are significantly preferred by the HEG, $P<0.05$ and $P<0.01$, indicated by '+' and '×', respectively. Cognate tRNA gene copy numbers are indicated for each organism. Grey-shaded cells indicate inferred ancestral optimal codons; universal optimal codons as defined by Sharp *et al.* (2005) are shown in bold type.

selection coefficient in *Enterobacteriaceae* was assessed. For each pair of species, the associations between the cophenetic distances and (i) the frequency of identical optimal codon choices and (ii) the absolute selection coefficient difference were studied. The cophenetic distance was estimated based on the branch length in the phylogenetic tree by means of the cophenetic function in the APE package from R (Paradis *et al.*, 2004).

RESULTS AND DISCUSSION

Phylogenetic reconstruction and genomic GC content distribution

A consensus tree was inferred from 216 phylograms (see Fig. S1 in the online version of this paper). This tree was in agreement with previous 16S rDNA-related reconstructions for this bacterial group (Naum *et al.*, 2008; Wertz *et al.*, 2003).

When *B. aphidicola* (GC=30.1 %) and *W. glossinidia* (GC=29.1 %) were not considered, the genomic GC content ranged from 40.2 to 58.9 % (mean=52.1 %, SD=5.5 %), showing an important variation when considering the phylogenetic distance. It has been shown that genomic GC content is a major determinant in global CU among bacteria (Chen *et al.*, 2004; Singer & Hickey, 2000); therefore we studied the GC distribution of coding sequences, discriminated by codon position (Fig. S2). The GC distribution of each codon position, for both complete genomes and orthologous genes, was compared. The difference in GC content between both set of genes was negligible (mean difference=−0.7 %; SD=0.7), although it was significant in some organisms (U-test, $P<0.001$). Therefore, observed minor differences indicate that the orthologous genes were representative of the compositional trends at the genomic scale (Fig. S2).

Natural selection acting at the level of translation in orthologous genes

CU analyses were performed on orthologous genes, 322 for *B. aphidicola* and *W. glossinidia*, and 727 for the remaining organisms. Thus, we focused on vertical inheritance, reducing the effect of horizontal gene transfer and lineage-specific duplications and deletions, comparing the intragenomic CU bias among similar genes in different genomes. Results based on orthologues were confirmed when the whole set of genes of each genome was studied, as shown by the highly significant correlation observed between the axes generated by WCA in the two groups of sequences ($0.72<r<0.99$, $P<0.001$).

In the case of *B. aphidicola* and *W. glossinidia*, the variance explained by the first axes is rather low (Table 1). This suggests that synonymous CU varies little across the genome, and in both cases our analyses confirmed earlier results (Herbeck *et al.*, 2003; Rispe *et al.*, 2004). In the remaining organisms, except for *Dickeya dadantii*, WCA clustered known HEG toward one extreme of the first axis (Fig. S3). Interestingly, for *D. dadantii* not the first but the second axis was related to expression level. These clusters included ribosomal protein coding genes, several translation elongation factors, enolase, glyceraldehyde-3-phosphate dehydrogenase, among other HEG. Besides, the axes associated with expression showed a significant correlation between the expression index values for each gene and their position along that axis (see Table 1). Therefore, expression is a main factor explaining intragenomic CU variation.

There was a large range in dS values among genes. It has been clearly stated that the standard set of HEG in *E. coli* cells (Ishihama *et al.*, 2008) tend to display the lowest values of synonymous divergence (dS). A significant negative correlation between dS and MELP (or CAI, data not shown) was found for all tested micro-organisms except for *B. aphidicola* and *W. glossinidia*, indicating that as a general rule, orthologous genes with extreme synonymous codon bias display lower dS values (Tables 1 and S2). Since the pioneer work of (Sharp & Li, 1987b), there is evidence that supports purifying selection at synonymous sites in *Enterobacteriaceae.* Highly significant correlation values hold for comparisons among phylogenetic distant organisms. From this, we can conclude that (i) natural selection was operative at the translational level in the last common ancestor of this monophyletic group, (ii) it is still a main force shaping CU, and (iii) the optimal codons are almost the same since the divergence from the last common ancestor of the family. This is striking given the divergence time of the group, the different genomic GC contents, generation times, and the diverse eco-physiological features that characterize the family (see below).

We compared the effect of selection acting at the level of speed in translation using the selection coefficient. Results showed that selection was higher in two monophyletic groups: one composed of *Proteus mirabilis* and *Providencia stuartii*, and the other composed of species from the genus *Salmonella* and *E. coli* (Table 1). In contrast, *B. aphidicola*, *W. glossinidia*, *D. dadantii*, *Serratia proteamaculans* and *Sodalis glossinidius* showed the lowest selection coefficients. These results agree with those reported by Supek *et al.* (2010) and Retchless & Lawrence (2011). The estimated correlation coefficient between MELP and ENC′ values also support the observed trends among the analysed organisms (data not shown).

Analysis of CU in relation to accuracy

It is known that the misincorporation of an amino acid can affect fitness, depending on how the substitution impacts on protein structure and function. This problem, in relation to CU, was first studied by Akashi (1994) in *Drosophila* species, and has also been analysed in *E. coli* (Eyre-Walker, 1996; Stoletzki & Eyre-Walker, 2007). In order to understand whether selection acting at the level of accuracy is operative in *Enterobacteriaceae*, the CU biases in conserved versus non-conserved protein regions were compared (see Methods). The estimation of pairwise p-distances showed that the blocks of aligned CR were on average eightfold (±3) less diverged than NCR, supporting the methodology employed to select regions. The estimated CU skew (Δ) in CR was correlated with the Δ in HEG. The latter value is a measure of the bias towards the usage of translational optimal triplets among HEG in relation to the rest. Bias in HEG has been mainly associated with selection for translational speed (see above). The positive and significant correlations suggest that CRs display a bias towards optimal codons (Fig. 1, Table S3), which might be explained as the result of selection for translation acting at the level of accuracy.

The analyses of CU at CRs showed that both components of translational selection (selection for translational speed and accuracy) were operative and favoured almost the same optimal codons across the *Enterobacteriaceae*. Interestingly some conserved deviations from this general rule have also been found (see CAG, ACC, UCC, UCG, UCA and CCG in Fig. 1). In these cases more 'accurate' codons do not match 'faster' translated triplets.

The comparison of the absolute CU skews between HEG (measured as |Δ|HEG) and CR (measured as |Δ|CR) suggests that the effect of selection on speed, when operative in a gene, is stronger than the effect on accuracy (Table S3). This was true for all species analysed except for those where the translational selection effect was negligible. The ΔCR estimated for subgroups of genes, clustered according to expression indexes, showed that the effect of selection on accuracy was widespread in many genes independently of the expression level (Fig. S4). In a way, this is unexpected, since it is not obvious why selection for translational accuracy would not strongly affect HEG in relation to lowly expressed genes (LEG; see, for example, Drummond & Wilke, 2008). In general, these results were in agreement with an earlier paper on three *E. coli* strains (Stoletzki & Eyre-Walker, 2007). Our results also suggest that the methods frequently used to compare CU between HEG and LEG not only detect triplets adapted for speed, but usually also identify 'accuracy codons'. This might be explained at least in part by the fact that the majority of HEG are also strongly conserved at the amino acid level, as is the case for ribosomal proteins. Indeed, selection for speed is mainly restricted to some genes which display a highly pronounced bias in CU, while selection for accuracy is present in CRs of almost all genes, and seems to be more or less independent of expression level.

Optimal codon choices and cognate tRNA gene copy number

Nine conserved optimal triplets were detected in the analysed micro-organisms ($\chi2$, $P<0.01$). This number increased to 11 when the significance was reduced to 0.05. At least one optimal triplet was identified for 13 amino acids, the exceptions being Cys, Glu, Lys, Pro and Gln. Thirty-two triplets never did emerge as significantly preferred in HEG, while 17 showed variable trends among the different species. Recently Wang *et al.* (2011) reported a similar set of optimal codons for some of the organisms analysed here. These results were obtained, as mentioned above, for orthologous genes. Interestingly, the pattern of optimal triplet choices was mainly conserved when the whole genomes were analysed (see Table S3).

The total number of identified optimal codons ranged from 16 to 19, and on average 69 % of them (SD=0.04) matched a cognate tRNA gene in each genome (Fig. 1). Among them, 13 highly conserved major codons matched a cognate tRNA in almost all genomes. No tRNA genes were identified for GCU (Ala), GUU (Val), GGU (Gly) or ACU (Thr), and only one was found for UCU (Ser) in *S. enterica*. Interestingly, these are among the optimal codons with higher conservation. However, as a general rule for the family, tRNAs that recognize the aformentioned codons by wobbling, were identified: GGC (codon GCC), GAC (codon GUC), GCC (codon GGC), GGA (codon UCC) and GGU (codon ACC), respectively (Fig. 1).

A highly conserved CU bias in *Enterobacteriaceae* was found, even in organisms where the number of genes encoding tRNAs has dramatically changed (Fig. 1). See for instance the case of synonymous families encoding Ala and Val in *Pantoea ananatis*, Ala in *Edwardsiella ictaluri*, Leu in *Proteus mirabilis* and Arg in *Sodalis glossinidius*. Accordingly, the most likely scenario is that the effect of changes in the copy number of tRNA genes in the general translational strategy is marginal, and some kind of ancestral stable state for the majority of codon families may prevail. Since major codons are not always recognized by the tRNA with more gene copies (Kahali *et al.*, 2008), 1 several non-mutually exclusive explanations might be proposed, which can be summarized as follows: (i) post-transcriptional modifications in the first position of the anticodon in some isoacceptor tRNAs; (ii) different expression levels that lead to a similar concentration of tRNAs independently of gene copy numbers; (iii) changes in the copy number of tRNA genes (in relation to the ancestral state) that might be unstable states, which do not alter the selectively determined CU bias (Higgs & Ran, 2008); (iv) finally, this uncoupling phenomenon might be the consequence of a reduced effect of natural selection.

We finally classified the 17 codons that showed variable trends in three groups: highly conserved (CAC, GAC, GUA, AUC and AAC), conserved (GCA, GAA, CCA, AAG and AGC) and non-conserved (AAA, GGC, CAG, UCC, CCU, CCG and CAA).

Note that conservation patterns in the optimal CU bias were also evident even in those genomes with the most extreme mutational bias, as can be derived from the GC profiles (Fig. S2).

Ancestral inference analysis of optimal codon choices and selection coefficients

The study of the evolutionary dynamics of optimal codon choices in the family *Enterobacteriaceae*, from extant species towards the last recent common ancestor (LRCA), was performed using the conserved set of 727 orthologous sequences. For this analysis, *B. aphidicola* and *W. glossinidia* were not considered mainly as a consequence of the high evolutionary rate of these species. In addition to the nine optimal codons present in all tested genomes of the *Enterobacteriacea*, the codons GGC (Gly), GUA (Val), CCU (Pro), AGC (Ser), CAC (His), GAC (Asp), AUC (Ile) and AAC (Asn) were inferred as optimal in the LRCA, with statistically significant support (BFT >2).

The analysis did not detect a preferred codon for the amino acids Cys, Gln and Lys in the LRCA. In the case of Cys, as noted above, there was no significant preferred codon in any species (Fig. 1). This is interesting for two reasons: (i) all species share a common tRNA for UGC; and (ii) this amino acid is encoded by a U/C-ended twofold degenerate family, and for these amino acids, the C- ending triplet is always preferred. The lack of a significant preferred codon might be due to the scarcity of this amino acid. In the case of Gln, we note that either CAA or CAG is optimal for some dispersed species, which suggests that an optimal codon for this amino acid is a relatively recently acquired state for some extant species and therefore is not a fixed state in the group (Figs 1 and 2). In addition, we noted that the number and the distribution of tRNA genes for this residue are highly variable and as a consequence it was not possible to establish a link with the preference for CAG or CAA. We did note, however, that CAG is used significantly more in conserved protein regions in all species (Fig. 1), which shows that this codon is optimal for accuracy. Optimal codon choices for Lys were also highly variable. AAA was optimal only in *Edwardsiella ictaluri* and marginally preferred in *Enterobacter cloacae* (Fig. 1). On the other hand, the preference for AAG was restricted to three lineages, suggesting a clear phylogenetic trend (Fig. 2).

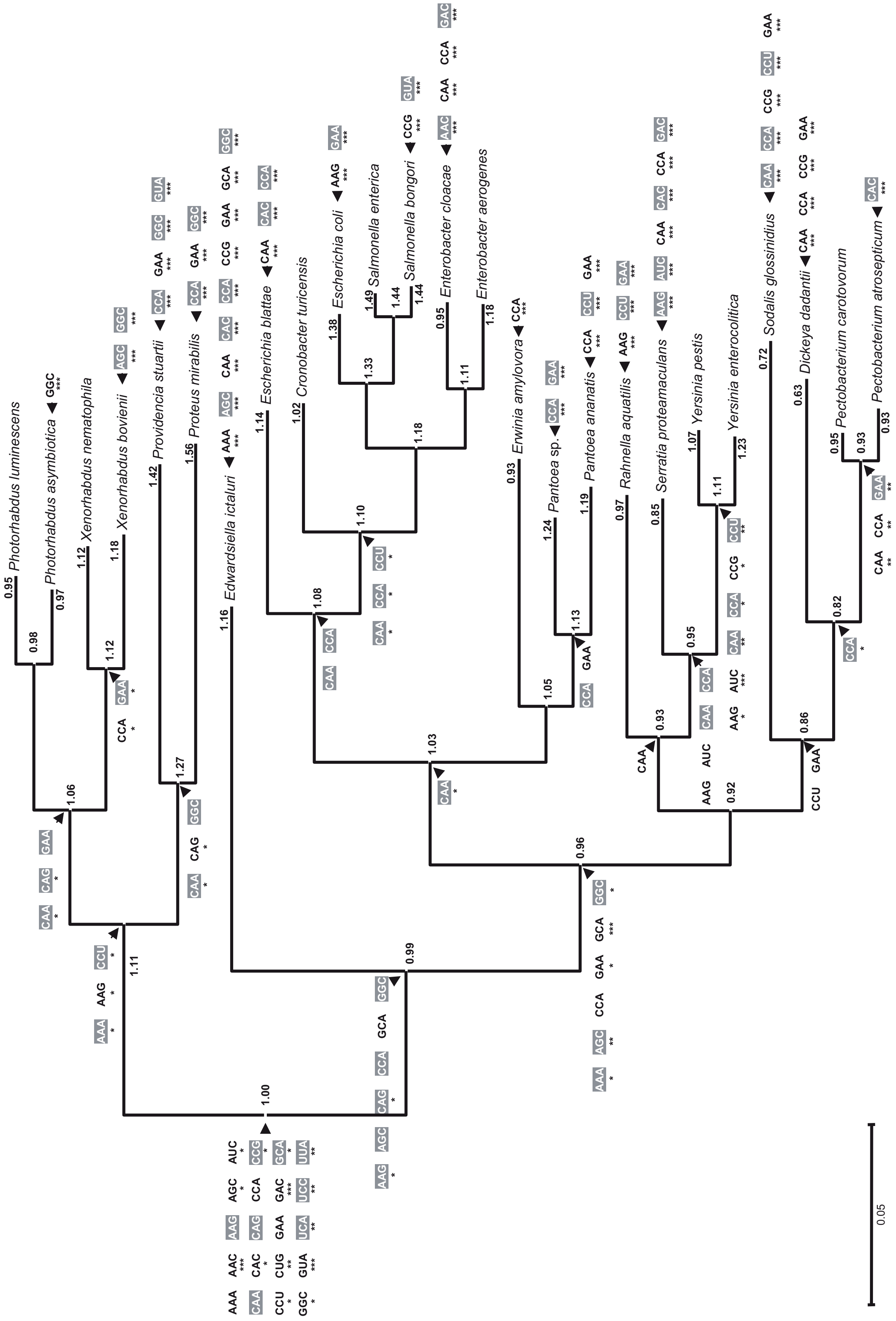
The evolution of selection coefficients (*S*) was reconstructed using a Bayesian approach. In addition, the possible relationship of *S* trends and some key eco-physiological data was assessed. In relation to growth rate, we observed a tendency (although non-signficant) in the sense that slow-growing bacteria tend to display lower selection values, i.e. *D. dadantii*. The lack of significance might be the result of the phylogenetic inertia associated with this character (see below) and the relatively small number of species analysed in each clade. We also observed the associations with biotic relationship reported earlier (Botzman & Margalit, 2011; Sharp *et al.*, 2005). The case of symbiotic species, which showed significantly lower values of *S* (U-test, $P<0.01$), might be better understood under the assumption that these microbes had undergone a general relaxation of selection pressures due to reduced effective population sizes (Sharp *et al.*, 2010). Note that this is also evident when analysing the reconstructed *S* trends of parasitic ‘lineages’ in relation to free-living ones (see for instance *Sodalis glossinidius* and species of the genus *Photorhabdus* in Fig. 2). Interestingly, the results for these species were strikingly different in relation to other symbiotic species such as *B. aphidicola* and *W. glossinidia*, in which the selection effect was negligible. For these species, which are characterized by an increased rate of molecular evolution, prolonged population bottlenecks may result in a dramatic reduction of the effect of selection, even in HEG (Hershberg & Petrov, 2008). For *B. aphidicola* and *W. glossinidia*, results showed that HEG and/or CR display significant more used (χ^2, $P<0.01$ and 0.05) or similar directional CU skews for a few highly conserved optimal codons (AAC, CGU, UCU, UCC, ACU, CAC and GGU). This could be interpreted as vestigial signs of selection, even though mutational biases cannot be completely excluded.

Other distinctive characteristics observed in certain lineages (for example, decreasing optimum growth temperature, strict aerobiosis or specific habitat changes) may also be related with particular trends in *S*; however, these results are inconclusive because there are not enough species with any particular feature to make significant any observed trend (Table S4).

Phylogenetic inertia

The correlation between cophenetic distances and absolute differences in selection coefficient suggests, as expected, that closely related micro-organisms show rather similar values. Besides, the variation is higher when comparing distantly related micro-organisms (data not shown), suggesting a remarkable effect of phylogenetic inertia of this character.

In the case of variable optimal triplets a clear phylogenetic pattern was found for GUA (Val) or AAG (Lys). On the other hand, there are some cases in which the optimal codon choice seems to be independent of the evolutionary distance [i.e. CCA (Pro) or GAA (Glu)]. For these triplets the results showed that organisms having cophenetic distances higher than 0.1 have a 50 % chance of sharing a certain triplet as a preferred codon, which approximate the expectancy by chance (Fig. S5). For the codons GCA (Ala) and AGC (Ser), only organisms having cophenetic distances higher than 0.3 showed different choices. This



Photorhabdus luminescens
Photorhabdus asymbiotica
Xenorhabdus nematophila
Xenorhabdus bovienii
Providencia stuartii
Proteus mirabilis
Edwardsiella ictaluri
Escherichia blattae
Cronobacter turicensis
Escherichia coli
Salmonella enterica
Salmonella bongori
Enterobacter cloacae
Enterobacter aerogenes
Erwinia amylovora
Pantoea sp.
Pantoea ananatis
Rahnella aquatilis
Serratia proteamaculans
Yersinia pestis
Yersinia enterocolitica
Sodalis glossinidius
Dickeya dadantii
Pectobacterium carotovorum
Pectobacterium atrosepticum
0.05

Fig. 2. Ancestral reconstruction of optimal codon choices and selection coefficients. Codons in black type represent ancestral inferred optimal triplets, while non-optimal codons are shown in white type within grey boxes. Only codons that have changed (significant to non-significant or optimal to non-optimal) in a node respective to the ancestral one are indicated. Thus, to get the list of optimal and non-optimal codons in a node, the analysis should be started from the root. Ancestral reconstructions were performed following the Bayesian inference method. A BFT value >2 was taken as 'positive' evidence, >5 as 'solid' evidence, and >7 as 'strong' evidence ('*', '**' and '***', respectively). A Bayesian approach was also used to reconstruct ancestral selection coefficients, which are indicated with black arrows next to nodes and tips.

was in agreement with the Bayesian reconstruction shown in Fig. 2, which suggests that the preference for these two codons represents opposite ancestral choices of the two main monophyletic groups, one comprising the genus *Photorhabdus*, *Xenorhabdus*, *Providencia* and *Proteus*, and the other comprising the rest of the organisms included in the reconstruction (Figs 2 and S5).

Concluding remarks

Our results suggest that natural selection was operative for CU in the last common ancestor of *Enterobacteriaceae* at two levels: speed and accuracy. The extant species share both features, although some minor differences were also found among them. This is striking, given the differences in lifestyles, effective population sizes, generation times, ecological niches, metabolic routes, genomic GC, etc., that characterize the different micro-organisms. Withers *et al.* (2006) found that for most of the tested genomes, CU frequencies matched more closely to the putative ancestral set of tRNA genes than to extant ones. Our results support this statement, which has important implications for understanding CU and tRNA coevolution. We suggest that extant CU bias is finely tuned over the ancestral well-conserved tRNA backbone.

ACKNOWLEDGEMENTS

We thank Hugo Naya and Tamara Fernández-Calero for providing scripting assistance. This work was partially supported by Fondo Clemente Estable 2007, FCE_722 (ANII, Uruguay) and PEDECIBA (Uruguay).

REFERENCES

Akashi, H. (1994). Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* **136**, 927–935.

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

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

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

Drummond, D. A. & Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341–352.

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

Eyre-Walker, A. (1996). Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? *Mol Biol Evol* **13**, 864–872.

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

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

Herbeck, J. T., Wall, D. P. & Wernegreen, J. J. (2003). Gene expression level influences amino acid usage, but not codon usage, in the tsetse fly endosymbiont *Wigglesworthia*. *Microbiology* **149**, 2585–2596.

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

Higgs, P. G. & Ran, W. (2008). Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. *Mol Biol Evol* **25**, 2279–2291.

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

Iriarte, A., Baraibar, J. D., Romero, H. & Musto, H. (2011). Selected codon usage bias in members of the class Mollicutes. *Gene* **473**, 110–118.

Ishihama, Y., Schmidt, T., Rappsilber, J., Mann, M., Hartl, F. U., Kerner, M. J. & Frishman, D. (2008). Protein abundance profiling of the *Escherichia coli* cytosol. *BMC Genomics* **9**, 102.

Kahali, B., Basak, S. & Ghosh, T. C. (2008). Delving deeper into the unexpected correlation between gene expressivity and codon usage bias of *Escherichia coli* genome. *J Biomol Struct Dyn* **25**, 655–661.

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

McGaughran, A. & Holland, B. R. (2010). Testing the effect of metabolic rate on DNA variability at the intra-specific level. *PLoS ONE* **5**, e9686.

Naum, M., Brown, E. W. & Mason-Gamer, R. J. (2008). Is 16S rDNA a reliable phylogenetic marker to characterize relationships below the family level in the Enterobacteriaceae? *J Mol Evol* **66**, 630–642.

Novembre, J. A. (2002). Accounting for background nucleotide composition when measuring codon usage bias. *Mol Biol Evol* **19**, 1390–1394.

Pagel, M. & Meade, A. (2006). Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *Am Nat* **167**, 808–825.

Pagel, M., Meade, A. & Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. *Syst Biol* **53**, 673–684.

Paradis, E., Claude, J. & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**, 289–290.

Paradis, S., Boissinot, M., Paquette, N., Bélanger, S. D., Martel, E. A., Boudreau, D. K., Picard, F. J., Ouellette, M., Roy, P. H. & Bergeron, M. G. (2005). Phylogeny of the *Enterobacteriaceae* based on genes encoding elongation factor Tu and F-ATPase *β*-subunit. *Int J Syst Evol Microbiol* **55**, 2013–2025.

Pham, H. N., Ohkusu, K., Mishima, N., Noda, M., Monir Shah, M., Sun, X., Hayashi, M. & Ezaki, T. (2007). Phylogeny and species identification of the family *Enterobacteriaceae* based on *dnaJ* sequences. *Diagn Microbiol Infect Dis* **58**, 153–161.

Rambaut, A. & Drummond, A. J. (2007). TRACER v. 1.5. See http://beast.bio.ed.ac.uk/Tracer.

Retchless, A. C. & Lawrence, J. G. (2011). Quantification of codon selection for comparative bacterial genomics. *BMC Genomics* **12**, 374.

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

Rispe, C., Delmotte, F., van Ham, R. C. & Moya, A. (2004). Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids. *Genome Res* **14**, 44–53.

Sharp, P. M. & Li, W. H. (1986). Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. *Nucleic Acids Res* **14**, 7737–7749.

Sharp, P. M. & Li, W. H. (1987a). The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* **15**, 1281–1295.

Sharp, P. M. & Li, W. H. (1987b). The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol Biol Evol* **4**, 222–230.

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 Res* **33**, 1141–1153.

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

Singer, G. A. & Hickey, D. A. (2000). Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol Biol Evol* **17**, 1581–1588.

Stoletzki, N. & Eyre-Walker, A. (2007). Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol Biol Evol* **24**, 374–381.

Sukumaran, J. & Holder, M. T. (2010). DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26**, 1569–1571.

Supek, F. & Vlahovicek, K. (2004). INCA: synonymous codon usage analysis and clustering by means of self-organizing map. *Bioinformatics* **20**, 2329–2330.

Supek, F. & Vlahovicek, K. (2005). Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC Bioinformatics* **6**, 182.

Supek, F., Skunca, N., Repar, J., Vlahovicek, K. & Smuc, T. (2010). Translational selection is ubiquitous in prokaryotes. *PLoS Genet* **6**, e1001004.

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

Talavera, G. & Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol* **56**, 564–577.

Thioulouse, J., Chessel, D., Dole'dec, S. & Olivier, J.-M. (1997). ADE-4: a multivariate analysis and graphical display software. *Stat Comput* **7**, 75–83.

Toth, I. K., Pritchard, L. & Birch, P. R. (2006). Comparative genomics reveals what makes an enterobacterial plant pathogen. *Annu Rev Phytopathol* **44**, 305–336.

Vieira-Silva, S. & Rocha, E. P. (2008). An assessment of the impacts of molecular oxygen on the evolution of proteomes. *Mol Biol Evol* **25**, 1931–1942.

Wang, B., Shao, Z. Q., Xu, Y., Liu, J., Liu, Y., Hang, Y. Y. & Chen, J. Q. (2011). Optimal codon identities in bacteria: implications from the conflicting results of two different methods. *PLoS ONE* **6**, e22714.

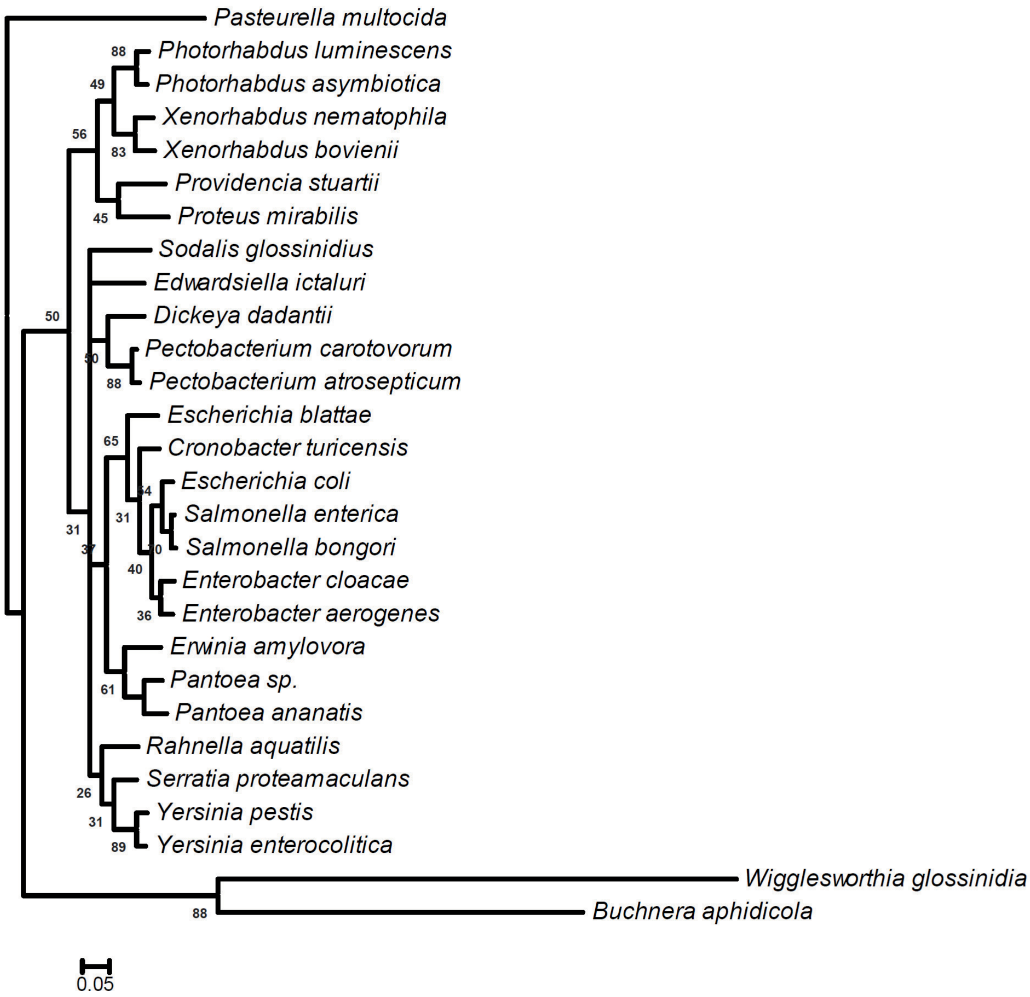
Wernegreen, J. J. & Funk, D. J. (2004). Mutation exposed: a neutral explanation for extreme base composition of an endosymbiont genome. *J Mol Evol* **59**, 849–858.

Wertz, J. E., Goldstone, C., Gordon, D. M. & Riley, M. A. (2003). A molecular phylogeny of enteric bacteria and implications for a bacterial species concept. *J Evol Biol* **16**, 1236–1248.

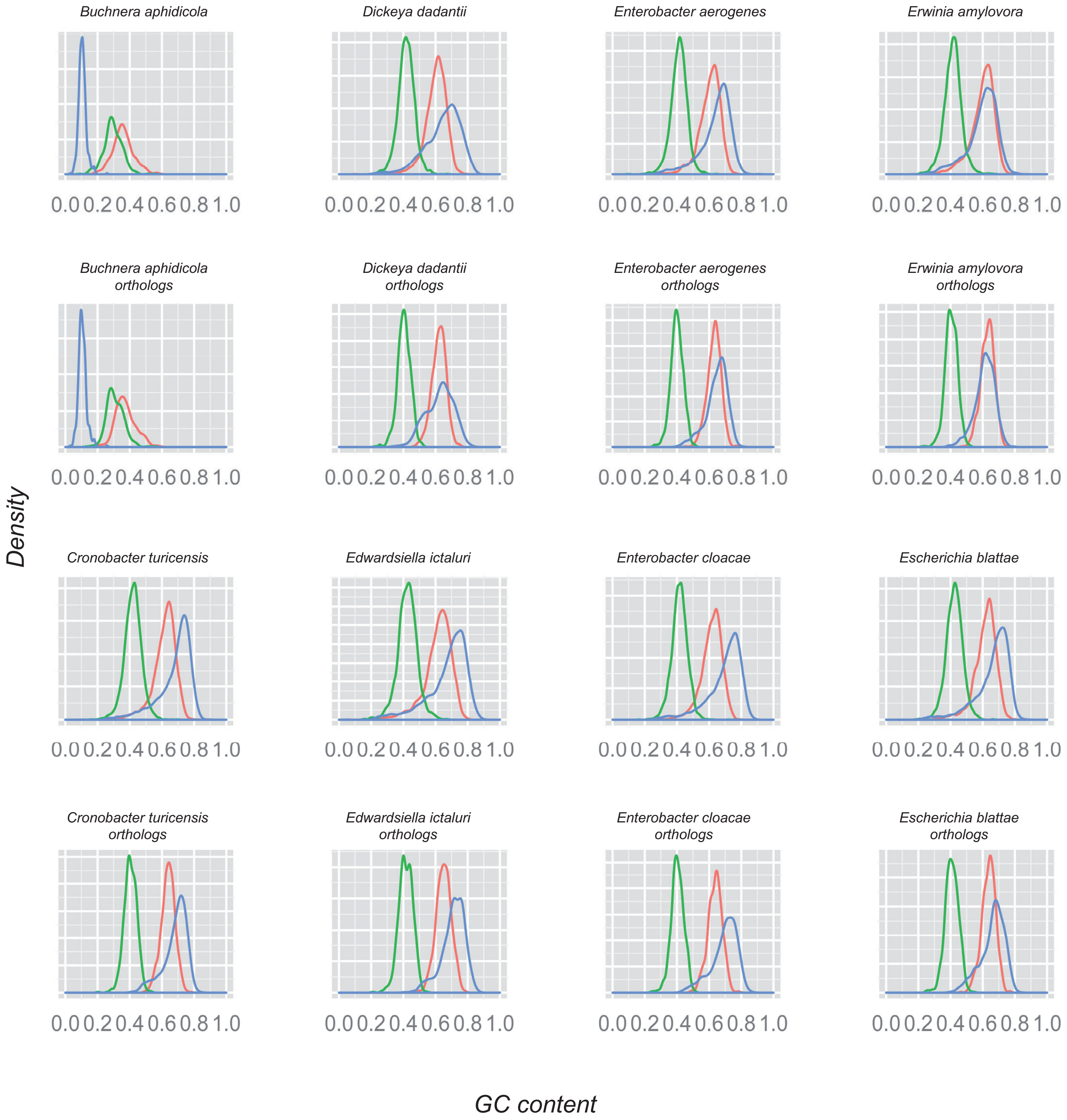
Withers, M., Wernisch, L. & dos Reis, M. (2006). Archaeology and evolution of transfer RNA genes in the *Escherichia coli* genome. *RNA* **12**, 933–942.

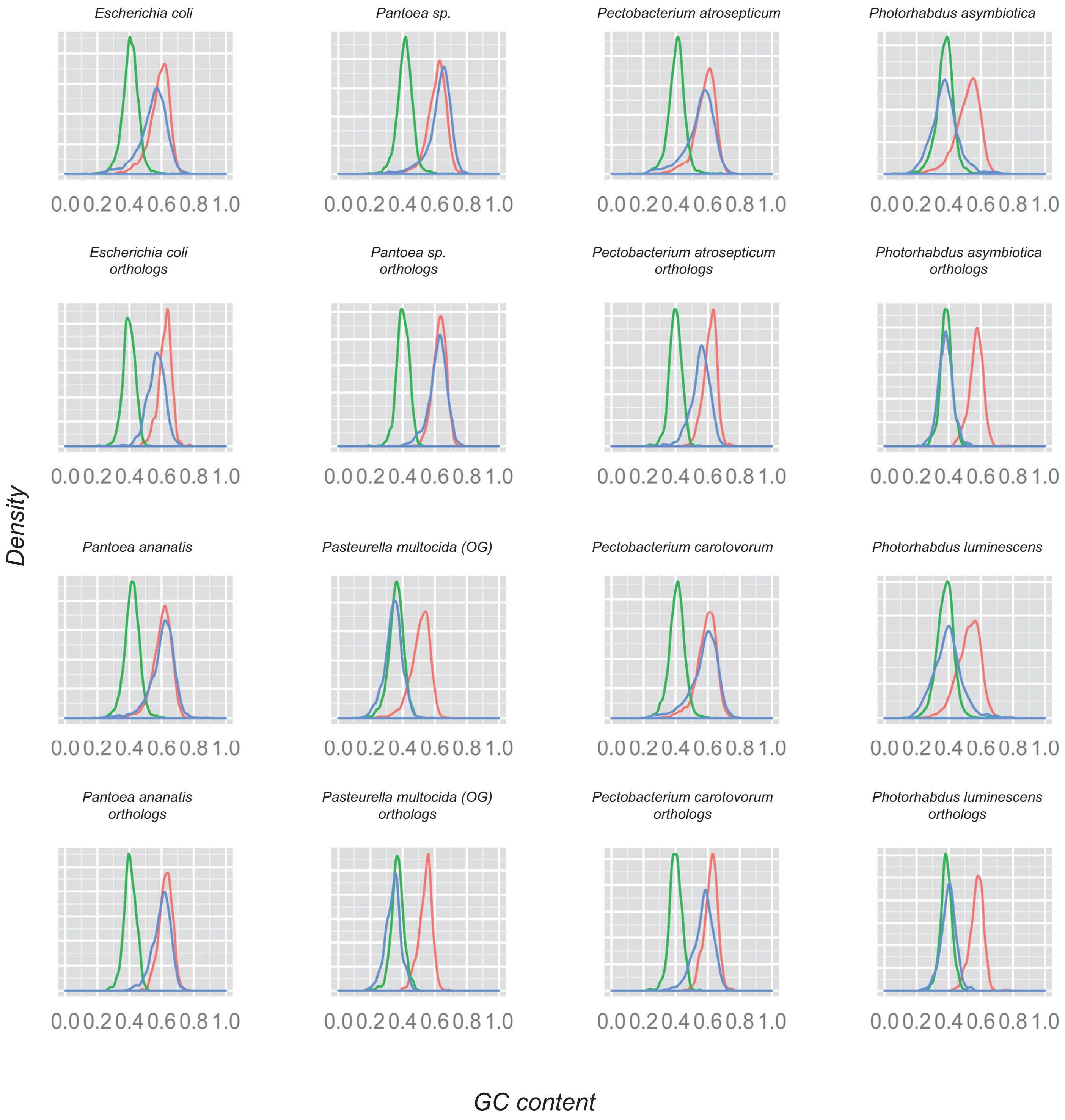
Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586–1591.

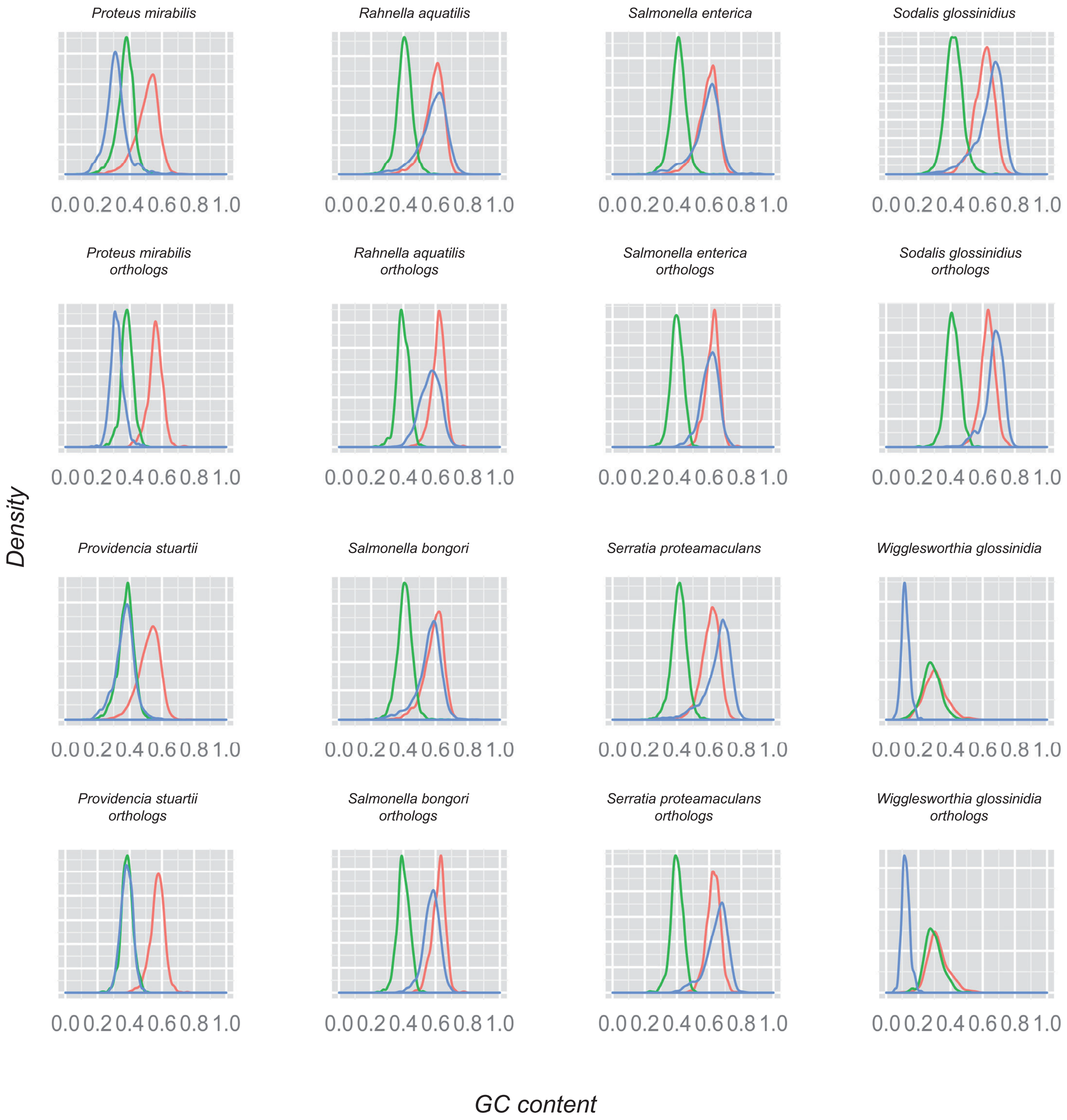
Edited by: R. Lan

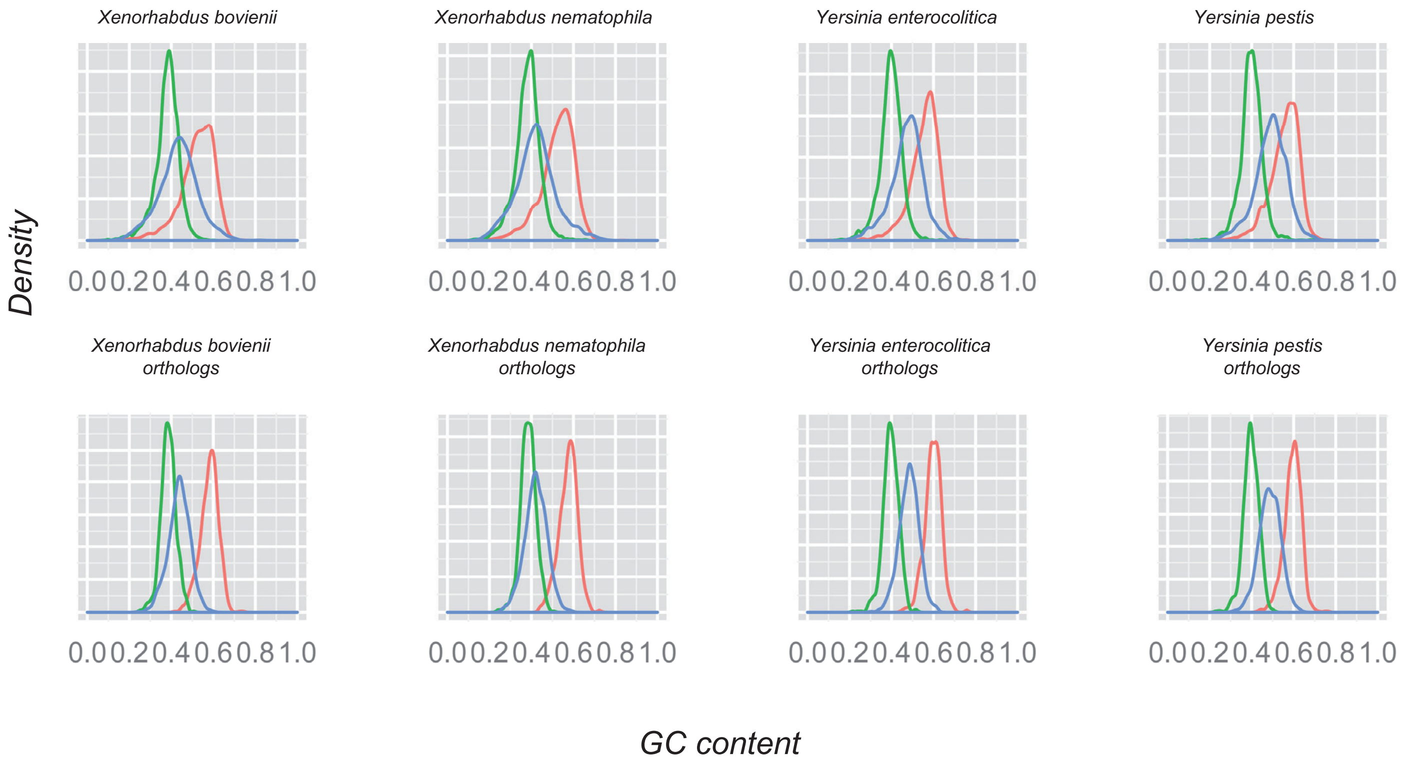


Supplementary Figure 1. Phylogenetic relationships among microorganisms analyzed. A consensus tree was inferred from 227 phylograms based on orthologous proteins. Reconstruction was made using the Maximum likelihood method with the amino acid LG+G model.

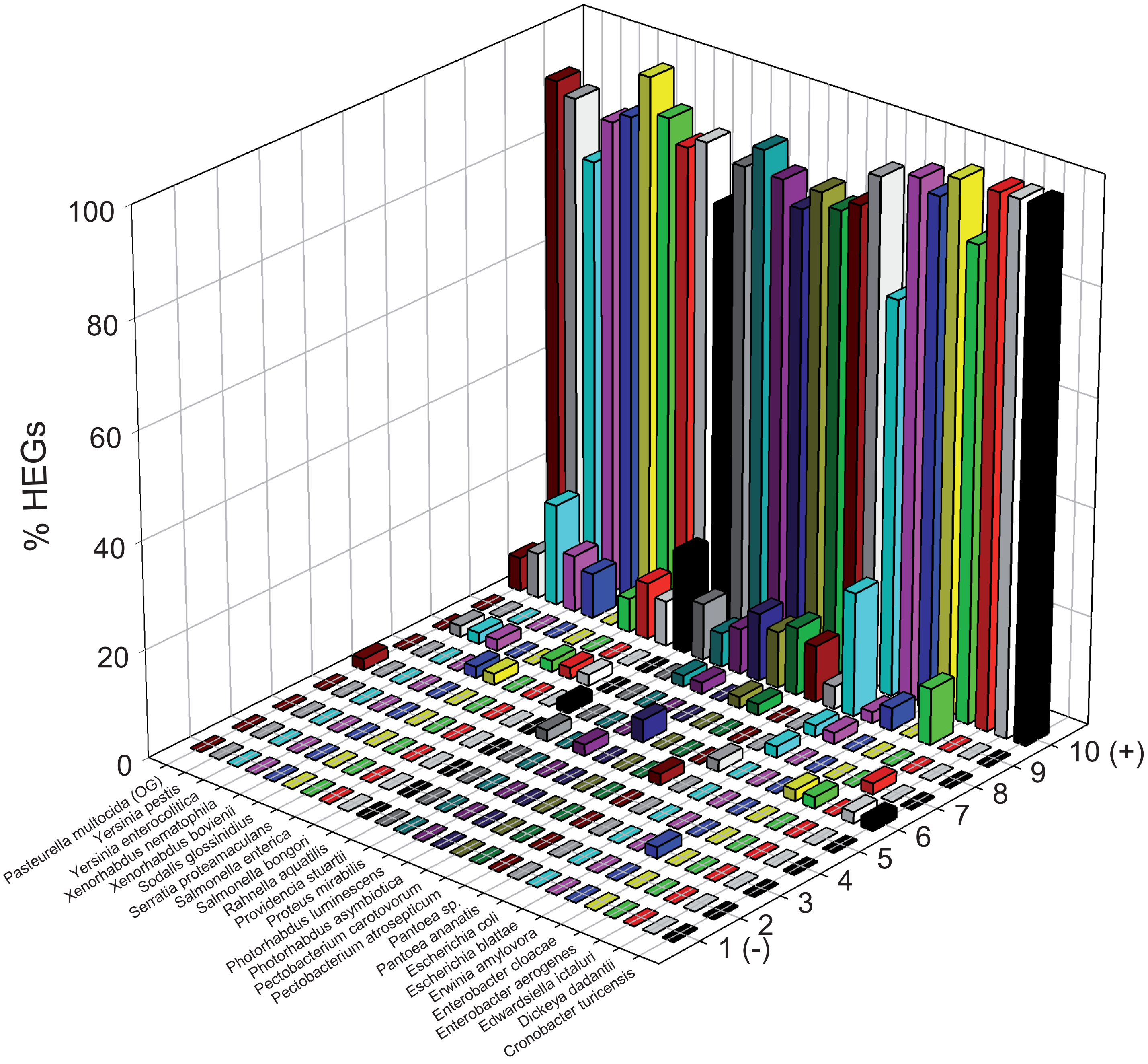

Buchnera aphidicola
0.0 0.2 0.4 0.6 0.8 1.0
Dickeya dadantii
0.0 0.2 0.4 0.6 0.8 1.0
Enterobacter aerogenes
0.0 0.2 0.4 0.6 0.8 1.0
Erwinia amylovora
0.0 0.2 0.4 0.6 0.8 1.0
Buchnera aphidicola orthologs
0.0 0.2 0.4 0.6 0.8 1.0
Dickeya dadantii orthologs
0.0 0.2 0.4 0.6 0.8 1.0
Enterobacter aerogenes orthologs
0.0 0.2 0.4 0.6 0.8 1.0
Erwinia amylovora orthologs
0.0 0.2 0.4 0.6 0.8 1.0
Density
Cronobacter turicensis
0.0 0.2 0.4 0.6 0.8 1.0
Edwardsiella ictaluri
0.0 0.2 0.4 0.6 0.8 1.0
Enterobacter cloacae
0.0 0.2 0.4 0.6 0.8 1.0
Escherichia blattae
0.0 0.2 0.4 0.6 0.8 1.0
Cronobacter turicensis orthologs
0.0 0.2 0.4 0.6 0.8 1.0
Edwardsiella ictaluri orthologs
0.0 0.2 0.4 0.6 0.8 1.0
Enterobacter cloacae orthologs
0.0 0.2 0.4 0.6 0.8 1.0
Escherichia blattae orthologs
0.0 0.2 0.4 0.6 0.8 1.0
GC content


Escherichia coli
Pantoea sp.
Pectobacterium atrosepticum
Photorhabdus asymbiotica
Escherichia coli orthologs
Pantoea sp. orthologs
Pectobacterium atrosepticum orthologs
Photorhabdus asymbiotica orthologs
Pantoea ananatis
Pasteurella multocida (OG)
Pectobacterium carotovorum
Photorhabdus luminescens
Pantoea ananatis orthologs
Pasteurella multocida (OG) orthologs
Pectobacterium carotovorum orthologs
Photorhabdus luminescens orthologs
0.0 0.2 0.4 0.6 0.8 1.0
Density
GC content

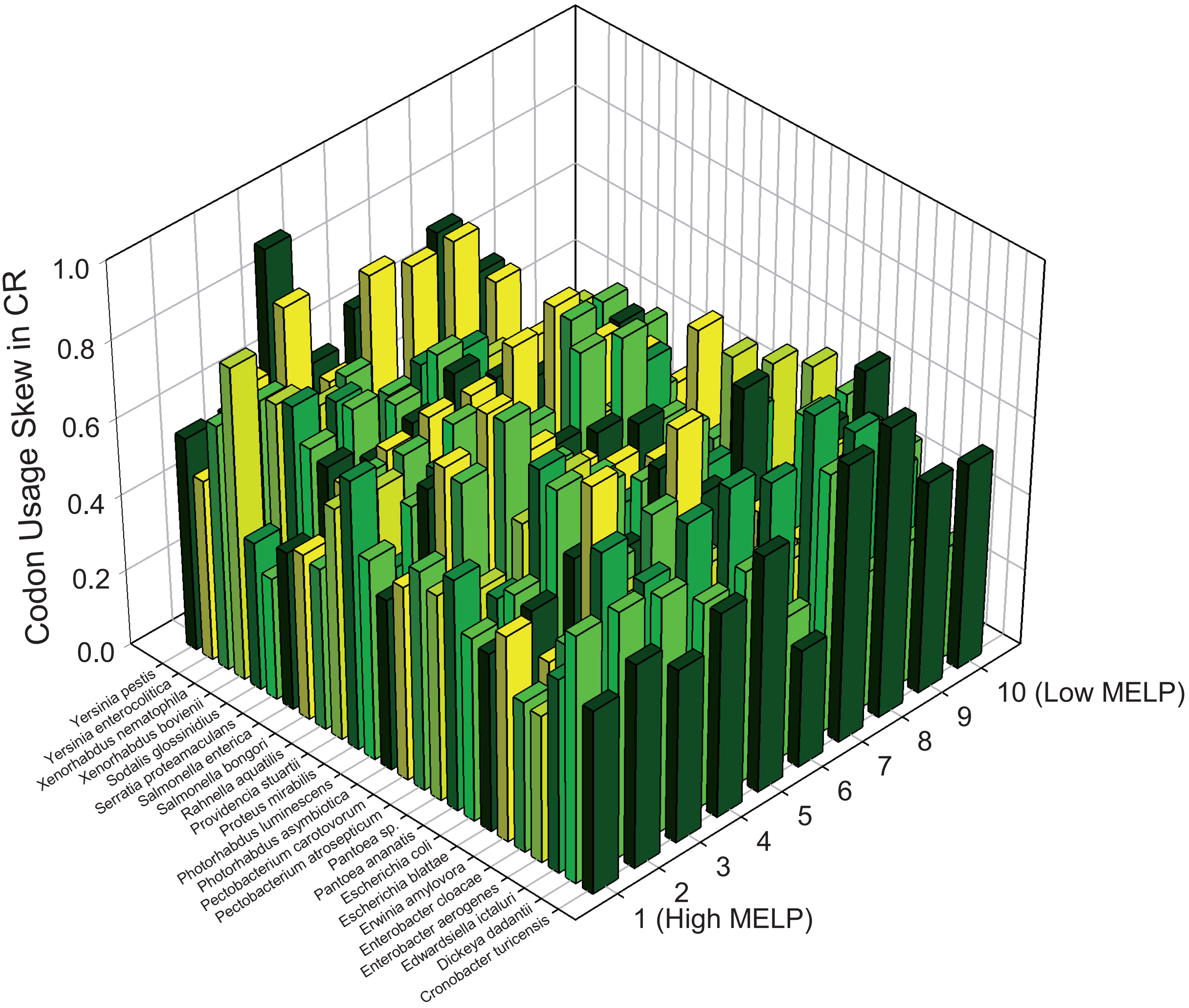

Proteus mirabilis
Rahnella aquatilis
Salmonella enterica
Sodalis glossinidius
Proteus mirabilis orthologs
Rahnella aquatilis orthologs
Salmonella enterica orthologs
Sodalis glossinidius orthologs
Providencia stuartii
Salmonella bongori
Serratia proteamaculans
Wigglesworthia glossinidia
Providencia stuartii orthologs
Salmonella bongori orthologs
Serratia proteamaculans orthologs
Wigglesworthia glossinidia orthologs
0.0 0.2 0.4 0.6 0.8 1.0
Density
GC content



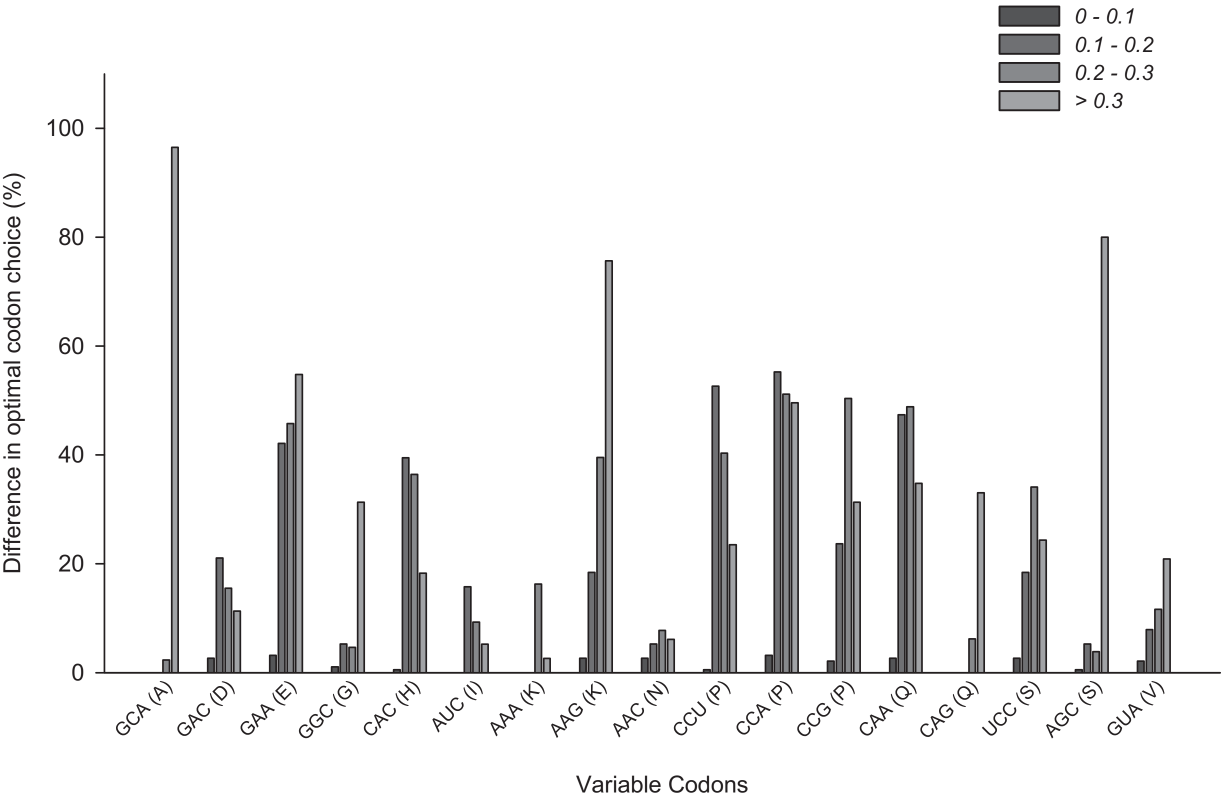
Supplementary Figure 2. Density plot of the GC content at the three codon positions for all (up) and orthologous genes (down) for each species analyzed in the present study. Red = GC1, Green = GC2, Blue = GC3.



Supplementary Figure 3. Histogram of the distribution of genes encoding ribosomal proteins (44), elongation factors (fusA and tsf), enolase (eno) and glyceraldehyde-3-phosphate dehydrogenase (gapA), along the WCA axis related to expression levels (see Table 1) for each species. Note that no axis was related to expression in the case of *B. aphidicola* and *W. glossinidia*. The axis was divided into 10 parts, each of them containing an equal number of genes.



Supplementary Figure 4. Histogram of the distribution of CU skews in conserved regions (ΔCR) of genes clustered according to MELP values. The sum of ΔCR was estimated using UUC, UCC, UCU, GGU, AUC, CUG, AAC and CGU, which were highly conserved among all considered species (Fig. 1). All gene clusters showed positive ΔCR toward these triplets, although to differing extents. Data was not shown for *B. aphidicola* and *W. glossinidia*, since for both organisms selection for translation was negligible for most clusters.



Supplementary Figure 5. Difference in optimal codon choices between species in relation to their phylogenetic distances. The histogram illustrates the frequency of genomes that show different preferences for each variable synonymous codon (absence –presence) in relation to their cophenetic distance.

Supplementary Table 1.

Number of putative orthologous sequences included in the synonymous pairwise distance analysis.

Only sequences displaying dS values ≤ 2 and a minimal of 40% identity value were considered for analysis. By means of BRH-based method a maximum of 727 orthologous genes were found.

	Cronobacter turicensis	*Dickeya dadantii*	*Edwardsiella ictaluri*	*Enterobacter aerogenes*	*Enterobacter cloacae*	*Erwinia amylovora*	*Escherichia blattae*	*Escherichia coli*	*Pantoea ananatis*	*Pantoea sp.*	*Pectobacterium atrosepticum*	*Pectobacterium carotovorum*	*Photorhabdus asymbiotica*	*Photorhabdus luminescens*	*Proteus mirabilis*	*Providencia stuartii*	*Rahnella aquatilis*	*Salmonella bongori*	*Salmonella enterica*	*Serratia proteamaculans*	*Sodalis glossinidius*	*Xenorhabdus bovienii*	*Xenorhabdus nematophila*	*Yersinia enterocolitica*	*Yersinia pestis*
Cronobacter turicensis		641	687	722	725	695	718	709	681	706	632	643	242	265	126	177	644	710	716	701	679	327	276	459	438
Dickeya dadantii			641	639	652	611	630	597	587	625	671	668	325	355	144	174	616	591	616	675	631	390	344	524	495
Edwardsiella ictaluri				682	703	665	692	595	630	675	593	627	235	257	109	136	619	595	639	703	679	294	268	465	442
Enterobacter aerogenes					724	694	711	713	675	688	616	638	300	330	154	203	657	718	716	693	658	340	286	519	491
Enterobacter cloacae						693	718	714	680	699	633	651	262	285	137	186	655	715	724	699	680	320	279	480	453
Erwinia amylovora							677	651	684	714	616	627	302	337	144	198	652	640	665	694	641	350	299	518	477
Escherichia blattae								683	661	691	578	613	295	301	152	182	636	678	686	690	679	323	286	486	474
Escherichia coli									631	665	601	614	372	389	196	245	638	717	722	657	565	398	360	546	503
Pantoea ananatis										719	585	609	298	310	166	205	626	626	650	661	582	357	320	503	480
Pantoea sp.											613	640	324	343	163	215	662	655	670	694	643	377	345	561	524
Pectobacterium atrosepticum												725	393	416	222	260	642	575	595	668	571	438	398	565	557
Pectobacterium carotovorum													396	414	213	252	640	606	617	676	594	439	411	570	558
Photorhabdus asymbiotica														726	454	429	399	345	325	379	198	683	689	519	508
Photorhabdus luminescens															455	427	418	363	353	401	230	679	677	530	514
Proteus mirabilis																471	197	180	166	176	92	394	426	323	307
Providencia stuartii																	236	213	211	212	125	407	392	331	325
Rahnella aquatilis																		606	628	691	604	442	417	609	582
Salmonella bongori																			722	645	571	370	341	491	485
Salmonella enterica																				676	619	366	336	489	472
Serratia proteamaculans																					689	449	366	672	653
Sodalis glossinidius																						257	238	440	423
Xenorhabdus bovienii																							725	486	472
Xenorhabdus nematophila																								484	468
Yersinia enterocolitica																									725
Yersinia pestis																									

Buchnera aphidicola		
Wigglesworthia glossinidia	287	

Supplementary Table 2

Correlation coefficients between the MELP index (Organisms MELP column) and the respective pairwise synonymous divergence (dS).

The results showed that genes with high MELP values are also characterized by small levels of synonymous divergence.
Grey cells correspond to non-significant correlations coefficients (p > 0.001).

Organism (MELP)	*Cronobacter turicensis*	*Dickeya dadantii*	*Edwardsiella ictaluri*	*Enterobacter aerogenes*	*Enterobacter cloacae*	*Erwinia amylovora*	*Escherichia blattae*	*Escherichia coli*	*Pantoea ananatis*	*Pantoea sp.*	*Pectobacterium atrosepticum*	*Pectobacterium carotovorum*	*Photorhabdus asymbiotica*	*Photorhabdus luminescens*	*Proteus mirabilis*	*Providencia stuartii*	*Rahnella aquatilis*	*Salmonella bongori*	*Salmonella enterica*	*Serratia proteamaculans*	*Sodalis glossinidius*	*Xenorhabdus bovienii*	*Xenorhabdus nematophila*	*Yersinia enterocolitica*	*Yersinia pestis*
Cronobacter turicensis	■	0.59	0.63	0.64	0.62	0.62	0.69	0.67	0.67	0.63	0.67	0.65	0.78	0.78	0.71	0.80	0.64	0.67	0.65	0.60	0.60	0.78	0.78	0.75	0.75
Dickeya dadantii	0.64	■	0.62	0.67	0.63	0.65	0.70	0.72	0.71	0.67	0.66	0.66	0.80	0.81	0.79	0.82	0.68	0.73	0.71	0.62	0.63	0.78	0.81	0.77	0.77
Edwardsiella ictaluri	0.60	0.58	■	0.58	0.54	0.58	0.59	0.69	0.64	0.60	0.67	0.63	0.77	0.76	0.67	0.76	0.67	0.69	0.65	0.54	0.55	0.77	0.78	0.75	0.76
Enterobacter aerogenes	0.59	0.58	0.57	■	0.60	0.59	0.64	0.63	0.63	0.61	0.65	0.65	0.76	0.75	0.66	0.75	0.61	0.64	0.63	0.57	0.57	0.77	0.77	0.72	0.74
Enterobacter cloacae	0.61	0.57	0.56	0.64	■	0.60	0.66	0.66	0.65	0.63	0.65	0.63	0.77	0.77	0.68	0.78	0.66	0.66	0.63	0.60	0.57	0.78	0.78	0.77	0.77
Erwinia amylovora	0.62	0.61	0.60	0.64	0.61	■	0.65	0.68	0.67	0.61	0.70	0.67	0.79	0.77	0.74	0.77	0.67	0.69	0.67	0.61	0.61	0.77	0.80	0.74	0.74
Escherichia blattae	0.68	0.64	0.61	0.67	0.67	0.64	■	0.69	0.69	0.66	0.73	0.71	0.77	0.78	0.73	0.77	0.69	0.71	0.69	0.62	0.61	0.77	0.77	0.75	0.75
Escherichia coli	0.69	0.68	0.71	0.69	0.67	0.66	0.71	■	0.70	0.69	0.72	0.71	0.73	0.73	0.63	0.70	0.70	0.66	0.66	0.71	0.65	0.73	0.76	0.75	0.74
Pantoea ananatis	0.69	0.68	0.68	0.71	0.69	0.68	0.72	0.73	■	0.69	0.73	0.71	0.81	0.80	0.78	0.82	0.71	0.74	0.72	0.67	0.68	0.80	0.81	0.76	0.77
Pantoea sp.	0.64	0.63	0.63	0.67	0.65	0.63	0.68	0.68	0.68	■	0.68	0.66	0.76	0.73	0.71	0.77	0.63	0.70	0.69	0.61	0.61	0.76	0.76	0.70	0.72
Pectobacterium atrosepticum	0.72	0.64	0.66	0.74	0.71	0.69	0.75	0.74	0.72	0.71	■	0.59	0.74	0.74	0.73	0.74	0.71	0.73	0.72	0.70	0.65	0.74	0.76	0.75	0.74
Pectobacterium carotovorum	0.69	0.65	0.67	0.72	0.67	0.66	0.74	0.73	0.72	0.69	0.59	■	0.75	0.75	0.73	0.73	0.69	0.73	0.72	0.69	0.63	0.74	0.76	0.74	0.74
Photorhabdus asymbiotica	0.78	0.78	0.73	0.80	0.78	0.79	0.78	0.81	0.79	0.77	0.76	0.77	■	0.52	0.75	0.77	0.77	0.79	0.80	0.78	0.69	0.71	0.72	0.75	0.75
Photorhabdus luminescens	0.77	0.80	0.73	0.78	0.78	0.78	0.79	0.80	0.80	0.79	0.77	0.79	0.54	■	0.75	0.77	0.77	0.80	0.80	0.78	0.72	0.71	0.72	0.74	0.74
Proteus mirabilis	0.74	0.74	0.68	0.79	0.76	0.76	0.78	0.81	0.79	0.78	0.80	0.79	0.79	0.78	■	0.81	0.81	0.80	0.79	0.77	0.61	0.82	0.81	0.80	0.80
Providencia stuartii	0.74	0.73	0.67	0.76	0.75	0.72	0.75	0.78	0.77	0.76	0.76	0.76	0.77	0.77	0.78	■	0.77	0.77	0.78	0.76	0.58	0.78	0.78	0.79	0.78
Rahnella aquatilis	0.66	0.63	0.67	0.68	0.68	0.65	0.72	0.70	0.69	0.64	0.69	0.68	0.72	0.72	0.71	0.73	■	0.71	0.70	0.68	0.62	0.71	0.73	0.72	0.73
Salmonella bongori	0.68	0.70	0.73	0.68	0.66	0.70	0.72	0.68	0.72	0.71	0.72	0.71	0.78	0.77	0.73	0.77	0.72	■	0.43	0.71	0.69	0.78	0.79	0.76	0.78
Salmonella enterica	0.66	0.67	0.70	0.67	0.64	0.67	0.71	0.67	0.71	0.70	0.71	0.71	0.77	0.78	0.72	0.76	0.71	0.43	■	0.70	0.66	0.77	0.78	0.77	0.78
Serratia proteamaculans	0.61	0.56	0.56	0.63	0.61	0.60	0.61	0.67	0.64	0.59	0.67	0.66	0.76	0.75	0.70	0.78	0.66	0.69	0.67	■	0.56	0.76	0.77	0.69	0.70
Sodalis glossinidius	0.64	0.61	0.61	0.65	0.60	0.64	0.64	0.73	0.71	0.65	0.71	0.69	0.81	0.81	0.73	0.79	0.69	0.71	0.69	0.61	■	0.83	0.83	0.78	0.79
Xenorhabdus bovienii	0.79	0.78	0.78	0.82	0.82	0.79	0.82	0.82	0.80	0.80	0.80	0.79	0.71	0.72	0.77	0.78	0.78	0.82	0.81	0.81	0.74	■	0.66	0.78	0.78
Xenorhabdus nematophila	0.80	0.79	0.77	0.81	0.81	0.79	0.81	0.83	0.80	0.80	0.78	0.79	0.72	0.73	0.77	0.78	0.78	0.83	0.81	0.81	0.68	0.66	■	0.78	0.78
Yersinia enterocolitica	0.75	0.76	0.74	0.78	0.79	0.74	0.78	0.79	0.77	0.74	0.77	0.78	0.75	0.73	0.74	0.78	0.75	0.77	0.77	0.73	0.70	0.75	0.77	■	0.67
Yersinia pestis	0.77	0.77	0.77	0.78	0.79	0.76	0.79	0.80	0.79	0.75	0.77	0.78	0.76	0.76	0.78	0.78	0.76	0.80	0.80	0.74	0.75	0.78	0.79	0.69	■

Organism (MELP)	*Buchnera aphidicola*	*Wigglesworthia glossinidia*
Buchnera aphidicola	■	0.05
Wigglesworthia glossinidia	0.12	■

Supplementary Table 3

Correlation coefficients between the Codon Usage Skews (Δ) estimated for highly expressed genes (HEG) and the respective skews in conserved regions (CR).

Average (AVG) and standard deviation (SD) of the absolute value of the estimated Δ (|Δ|) of codons is shown for each organisms. |Δ| was derived from the estimated Δ for each codon in each group of genes according to the equation presented in Materials and Methods.

Correlations between |Δ| HEG(AG) and |Δ| HEG(OR) show that the bias in highly expressed genes estimated using orthologous genes (OR) is highly similar to the bias estimated using all genes (AG).

Observed correlations between ΔHEG and ΔCR in each organism suggest that highly expressed genes and conserved regions show similar codon usage biases. Average skews among codons and standard deviation show that the observed bias is more pronounced in highly expressed genes than in conserved regions.

$\Delta HEG_{(AG)}$: Δ of highly expressed genes set (reference set = 44 ribosomal proteins) respective to all genes in each genome.

$\Delta HEG_{(OR)}$: Δ of highly expressed genes set (reference set = 44 ribosomal proteins) respective to set of orthologs.

ΔCR: Δ of conserved regions respective to non conserved regions

Grey boxes indicate non significant correlation values ($p > 0.001$).

Organism	$\|\Delta\|$ $HEG_{(AG)}$		$\|\Delta\|$ $HEG_{(OR)}$		$\|\Delta\|$ CR		Pearson's *r*		
							$\Delta HEG_{(AG)}$ *vs.* $\Delta HEG_{(OR)}$	$\Delta HEG_{(AG)}$ *vs.* ΔCR	$\Delta HEG_{(OR)}$ *vs.* ΔCR
	AVG	SD	AVG	SD	AVG	SD			
Cronobacter turicensis	0.379	0.270	0.329	0.242	0.119	0.131	0.990	0.712	0.630
Dickeya dadantii	0.325	0.255	0.295	0.241	0.088	0.096	0.996	0.729	0.693
Edwardsiella ictaluri	0.378	0.272	0.349	0.266	0.098	0.108	0.983	0.670	0.607
Enterobacter aerogenes	0.386	0.271	0.324	0.255	0.106	0.107	0.982	0.767	0.686
Enterobacter cloacae	0.371	0.276	0.317	0.270	0.111	0.121	0.979	0.716	0.614
Erwinia amylovora	0.316	0.249	0.285	0.223	0.092	0.113	0.984	0.670	0.581
Escherichia blattae	0.379	0.277	0.341	0.264	0.108	0.114	0.990	0.692	0.630
Escherichia coli	0.389	0.282	0.331	0.266	0.113	0.122	0.986	0.759	0.704
Pantoea ananatis	0.325	0.249	0.284	0.220	0.097	0.123	0.992	0.670	0.595
Pantoea sp.	0.359	0.277	0.311	0.265	0.107	0.113	0.987	0.707	0.646
Pectobacterium atrosepticum	0.312	0.248	0.268	0.226	0.085	0.100	0.987	0.735	0.670
Pectobacterium carotovorum	0.340	0.256	0.298	0.238	0.090	0.100	0.990	0.729	0.668
Photorhabdus asymbiotica	0.284	0.240	0.244	0.217	0.075	0.083	0.991	0.744	0.716
Photorhabdus luminescens	0.297	0.238	0.258	0.214	0.072	0.084	0.992	0.752	0.721
Proteus mirabilis	0.338	0.254	0.315	0.238	0.067	0.064	0.997	0.787	0.768
Providencia stuartii	0.330	0.251	0.301	0.234	0.071	0.068	0.995	0.804	0.783
Rahnella aquatilis	0.338	0.254	0.278	0.220	0.096	0.106	0.989	0.735	0.659
Salmonella bongori	0.388	0.264	0.342	0.251	0.100	0.109	0.992	0.728	0.671
Salmonella enterica	0.392	0.275	0.344	0.269	0.104	0.113	0.989	0.732	0.676
Serratia proteamaculans	0.335	0.274	0.283	0.253	0.104	0.113	0.983	0.682	0.605
Sodalis glossinidius	0.280	0.230	0.256	0.223	0.079	0.096	0.970	0.638	0.516
Xenorhabdus bovienii	0.314	0.249	0.281	0.231	0.076	0.080	0.992	0.723	0.677
Xenorhabdus nematophila	0.327	0.254	0.298	0.234	0.075	0.082	0.993	0.745	0.710
Yersinia enterocolitica	0.315	0.253	0.271	0.235	0.086	0.090	0.993	0.778	0.744
Yersinia pestis	0.314	0.244	0.278	0.223	0.081	0.101	0.992	0.751	0.717
Buchnera aphidicola	0.067	0.063	0.070	0.065	0.049	0.052	0.985	*0.171*	*0.149*
Wigglesworthia glossinidia	0.074	0.088	0.074	0.087	0.051	0.051	0.982	*0.328*	*0.286*
Pasteurella multocida	0.328	0.261	0.323	0.252	0.083	0.100	0.993	0.601	0.549

Supplementary Table 4.

List of the 28 genomes composing our dataset and their characteristics.

Characteristics were were retrieved from the literature and online databases. We defined the minimum generation time as the smallest value reported for one species. The optimum growth temperature of the species was indicated under Column “OGT”.

Microorganisms	Motility [2]	OxygenReq [2]	Habitat [2,7]	Biotic relationship [2]	OGT (ºC) [1,2]	Generation time (hr) [3,4,5,6,8,9,10,11,12,13, 14,15,16,17,18,19]
Cronobacter turicensis	yes	Facultative	Host-Associated	NA	30	0.37
Dickeya dadantii	no	Facultative	Multiple	Free living	30	1.50
Edwardsiella ictaluri	no	Facultative	Aquatic	Free living	28	0.57
Enterobacter aerogenes	yes	Facultative	Host-Associated	Free living	30	0.38
Enterobacter cloacae	NA	Facultative	Soil	Free living	30	0.55
Erwinia amylovora	no	Facultative	Host-Associated	Free living	28	1.10
Escherichia blattae	yes	Facultative	NA	NA	37	NA
Escherichia coli	yes	Facultative	Host-Associated	Free living	37	0.35
Pantoea ananatis	yes	Facultative	Multiple	Free living	30	1.20
Pantoea sp.	no	Facultative	Multiple	Free living	Mesophilic	0.92
Pectobacterium atrosepticum	yes	Facultative	Multiple	Free living	30	1.20
Pectobacterium carotovorum	yes	Facultative	Host-Associated	Free living	26	NA
Photorhabdus asymbiotica	no	Facultative	Host-Associated	Symbiotic	28	NA
Photorhabdus luminescens	yes	Facultative	Host-Associated	Symbiotic	28	0.30
Proteus mirabilis	no	Aerobic	Host-Associated	Free living	37	0.42
Providencia stuartii	yes	Facultative	Multiple	NA	37	1.29
Rahnella aquatilis	yes	Facultative	Aquatic	Free living	30	NA
Salmonella bongori	yes	Facultative	Host-Associated	Free living	30	NA
Salmonella enterica	no	Facultative	Multiple	Free living	37	0.33
Serratia proteamaculans	yes	Facultative	Multiple	Free living	30	0.59
Sodalis glossinidius	no	Microaerophilic	Host-Associated	Symbiotic	25	0.43
Xenorhabdus bovienii	no	Facultative	Host-Associated	Free living	28	0.31
Xenorhabdus nematophila	no	Facultative	Host-Associated	Free living	26	0.56
Yersinia enterocolitica	yes	Facultative	Multiple	Free living	30	0.50
Yersinia pestis	yes	Facultative	Multiple	Free living	37	1.25
Buchnera aphidicola	NA	NA	Host-Associated	Symbiotic	22	36.00
Wigglesworthia glossinidia	NA	NA	Host-Associated	Symbiotic	Mesophilic	36.00
Pasteurella multocida	no	Facultative	Host-Associated	Free living	37	1.00

Supplementary Table 4 References:

1) DSMZ Bacteria Collection (http://www.cabri.org/CABRI/srs-doc/index.html)

2) BacMap Database (http://bacmap.wishartlab.com, Stothard et al. 2005 "BacMap: an interactive picture atlas of annotated bacterial genomes")

3) Freilich et al. 2009 ("Metabolic-network-driven analysis of bacterial ecological strategies")

4) Vieira-Silva and Rocha 2010 ("The Systemic Imprint of Growth and Its Uses in Ecological (Meta)Genomics")

5) Cooney et al. 2009. "Growth of Enterobacteriaceae in milk" 1st International Conference on Cronobacter, Poster Abstract 22.

6) Hugouvieux-Cotte-Pattat and Charaoui-Boukerzaza 2009 ("Catabolism of raffinose, sucrose, and melibiose in Erwinia chrysanthemi 3937")

7) NCBI BioProject Database (http://www.ncbi.nlm.nih.gov/bioproject)

8) Ingham et al. 2006 ("Rapid antibiotic sensitivity testing and trimethoprim-mediated filamentation of clinical isolates of the Enterobacteriaceae assayed on a novel porous culture support")

9) Obayori et al. 2012. ("Degradation of weathered crude oil (Escravos Light) by bacterial strains from hydrocarbons-polluted site")

10) Biosecurity Australia (2011) Draft report for the non-regulated analysis of existing policy for apples from New Zealand. Department of Agriculture, Fisheries and Forestry, Canberra. (taken from cites Hildebrand 1938; Billing 1974b; Shrestha et al. 200

11) Adams et al. 2009 ("Effects of symbiotic bacteria and tree chemistry on the growth and reproduction of bark beetle fungal symbionts")

12) Hasegawa et al. 2005 ("Elevated Temperature Enhances Virulence of Erwinia carotovora subsp. carotovora Strain EC153 to Plants and Stimulates Production of the Quorum Sensing Signal, N-Acyl Homoserine Lactone, and Extracellular Proteins")

13) Sosa et al. 2006 ("Proteus mirabilis isolates of different origins do not show correlation with virulence attributes and can colonize the urinary tract of mice")

14) Surekha Rani et al. 2008 ("Isolation and characterization of a chlorpyrifosdegrading bacterium from agricultural soil and its growth response")

15) Knodlera et al. 2010 ("Dissemination of invasive Salmonella via bacterial-induced extrusion of mucosal epithelia")

16) Irwin et al . 2008. ("Binding of nontarget microorganisms from food washes to anti-Salmonella and anti-E. coli O157 immunomagnetic beads: most probable composition of background Eubacteria")

17) Nyamboli, MA. 2008 ("Mass production of entomopathogenic nematodes for plant protection")

18) Herbert E. 2008 (Chapter 2, pp 66. in "Regulation of Xenorhabdus Nematophila Mutualism and Pathogenicity by the CPXRA Two-component System")

19) Neuhaus et al. 2000 ("Restart of Exponential Growth of Cold-Shocked Yersinia enterocolitica Occurs after Down-Regulation of cspA1/A2 mRNA")

5 Capítulo II: Estudio comparativo del UCS en el género *Aspergillus*

Translational selection on codon usage in the genus Aspergillus

Iriarte A, Sanguinetti M, Fernández-Calero T, Naya H, Ramón A, Musto H.

Gene 2012, 506(1):98-105.

Resumen:

Aspergillus es un género de hongos que incluye más de 200 especies descriptas. Muchos miembros del grupo son patógenos relevantes o especies con importancia económica. Al momento de escribir este trabajo el uso de codones sinónimos ha sido analizado en una única especie del grupo y con un pequeño número de genes. En este artículo analizamos los patrones de uso de codones en ocho genomas completamente secuenciados que pertenecen a este género. Los resultados sugieren que la selección operando sobre la velocidad y la precisión en la traducción son los principales factores que determinan la variabilidad en el uso de codones sinónimos en todas las especies estudiadas, y por lo tanto sugieren que la selección era activa en el último ancestro común del grupo. El análisis de distancias sinónimas y la composición demuestran que los genes altamente expresados evolucionan más lentamente en sitios sinónimos independientemente de los sesgos composicionales. Se identificó un núcleo conservado de codones traduccionalmente óptimos y se estudió el conjunto de tRNAs en cada genoma. Se encontró que la gran mayoría de los tripletes preferidos concuerdan con los tRNA isoaceptores con más copias en el genoma. Se discuten los posibles escenarios que pueden explicar las diferencias observadas entre las especies analizadas. Por último destacamos la aplicación biotecnológica de esta investigación en relación con la expresión heteróloga de proteínas.

Gene 506 (2012) 98–105

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Translational selection on codon usage in the genus *Aspergillus*

Andrés Iriarte [a,b,c], Manuel Sanguinetti [d], Tamara Fernández-Calero [e], Hugo Naya [e,f], Ana Ramón [d], Héctor Musto [a,*]

[a] *Laboratorio de Organización y Evolución del Genoma, Depto. de Ecología y Evolución, Facultad de Ciencias (UDELAR), Iguá 4225, 11400 Montevideo, Uruguay*
[b] *Laboratorio de Evolución, Depto. de Ecología y Evolución, Facultad de Ciencias (UDELAR), Iguá 4225, 11400 Montevideo, Uruguay*
[c] *Área Genética, Depto. de Genética y Mejora Animal, Facultad de Veterinaria (UDELAR), Av. A. Lasplaces 1550, CP 11600, Montevideo, Uruguay*
[d] *Sección Bioquímica, Depto. de Biología Celular y Molecular, Facultad de Ciencias (UDELAR), Iguá 4225, 11400 Montevideo, Uruguay*
[e] *Unidad de Bioinformática, Institut Pasteur de Montevideo, Mataojo 2020, CP 11400 Montevideo, Uruguay*
[f] *Grupo de Mejoramiento Genético Animal, Depto. de Producción Animal y Pasturas, Facultad de Agronomía (UDELAR), Av. Garzón 780, CP 12900, Montevideo, Uruguay*

ARTICLE INFO

Article history:
Accepted 15 June 2012
Available online 26 June 2012

Keywords:
Aspergilli
Neosartorya
Translational selection
Optimal codons
tRNA content

ABSTRACT

Aspergillus is a genus of mold fungi that includes more than 200 described species. Many members of the group are relevant pathogens and other species are economically important. Only one species has been analyzed for codon usage, and this was performed with a small number of genes. In this paper, we report the codon usage patterns of eight completely sequenced genomes which belong to this genus. The results suggest that selection for translational efficiency and accuracy are the major factors shaping codon usage in all of the species studied so far, and therefore they were active in the last common ancestor of the group. Composition and molecular distances analyses show that highly expressed genes evolve slower at synonymous sites. We identified a conserved core of translationally optimal codons and study the tRNA gene pool in each genome. We found that the great majority of preferred triplets match the respective cognate tRNA with more copies in the respective genome. We discuss the possible scenarios that can explain the observed differences among the species analyzed. Finally we highlight the biotechnological application of this research regarding heterologous protein expression.

1. Introduction

The genetic code is “redundant”, since usually each amino acid is coded by more than one codon. However, it is widely known that triplets coding for the same amino acid are not equally frequent, both among organisms and among genes within a single species (Grantham et al., 1981). Global codon usage depends mainly on the GC content of each genome (Grantham, 1980; Sueoka, 1962). On the contrary, variable synonymous codon usage among genes mostly reflects a balance between biases generated by mutation, natural selection and random genetic drift (Bulmer, 1991; Sharp and Li, 1986). In several unicellular species (both prokaryotes and eukaryotes) it has been demonstrated that highly expressed genes (HEGs) tend to use a subset of codons, which usually match the most abundant tRNAs (Ikemura, 1985; Yamao et al., 1991). These are known as “major codons”, and its usage among these sequences is driven by natural selection acting at the level of speed (and accuracy) during translation. On the other hand, lowly expressed genes (LEGs) tend to display a codon usage pattern that mostly depends on the mutational bias (Ermolaeva, 2001). Recently, a link was observed between optimal codon choice and genomic GC content (Hershberg and Petrov, 2009).

For model unicellular eukaryotic species, such as *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, it has been known for a long time that synonymous codon usage bias in a gene is correlated with its own transcription level (Ikemura, 1985; Sharp and Cowe, 1991; Sharp et al., 1988). This phenomenon is now recognized as crucial in shaping gene expression and cellular function through its effects at different levels, ranging from RNA processing to protein translation and folding (Plotkin and Kudla, 2011).

The Aspergilli constitute a genus with more than 200 members, classified within the Pezizomycotina, the largest subphylum of the Ascomycota. This taxonomic group comprises species whose characteristics are of high pathological, agricultural, industrial, pharmaceutical and scientific importance (Kale et al., 2003; Scazzocchio, 2006). Despite belonging to the same genus, *Aspergillus* species have diverged significantly (Galagan et al., 2005), though they are sufficiently related such that orthologs can be easily identified for the majority of genes.

Only two papers have analyzed codon usage in this genus, and both studied the biases in *A. nidulans* (Gurr et al., 1987; Lloyd and Sharp, 1991). Although both papers used different approaches, they

Abbreviations: GC, molar content of Guanine+Cytosine; GC3, GC content at third-codon positions; HEGs, Highly Expressed Genes; LEGs, Lowly Expressed Genes; COA, Correspondence Analysis; WCA, Within group Correspondence Analysis; RSCU, Relative Synonymous Codon Usage; MELP, MILC-based Expression Level Predictor; dS, Synonymous substitution rate; CR, Conserved Regions; NCR, Non-Conserved Regions.

* Corresponding author. Tel.: +598 2 5258618.
E-mail address: hmusto@fcien.edu.uy (H. Musto).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2012.06.027

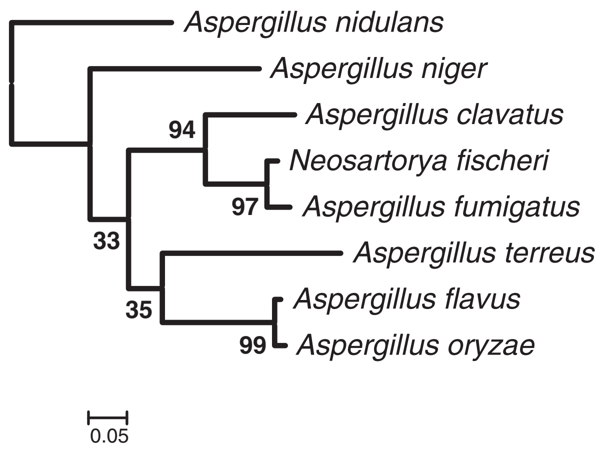


Fig. 1. Phylogenetic reconstruction of the eight sequenced *Aspergillus* species. A consensus tree was inferred from 3191 phylograms based on orthologous genes.

both concluded that in *A. nidulans* HEGs present a biased codon usage towards a group of approximately 19–20 optimal codons. They also stated that most of these triplets are G- or C-ending, making highly expressed genes more GC-rich at silent sites than LEGs.

Taking advantage of the available information concerning complete genome sequences and microarray data in the Aspergilli, we aim to understand the particular codon usage patterns observed in all available completely sequenced members of this group. We take a comparative perspective attempting to analyze the relative effect of selection on the observed codon usage patterns, since its significance is not well established for many members of this group. We conclude that natural selection has been shaping codon usage in these species, both at the level of speed and accuracy of translation since the split of these species from last common ancestor.

2. Materials and methods

2.1. Sequences and phylogenetic reconstruction

Complete genome and coding sequences were obtained from the *Aspergillus* Comparative Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org/), which comprised eight species: *Aspergillus flavus* (Payne et al., 2008), *A. oryzae* (Machida et al., 2005), *A. terreus* (www.broadinstitute.org), *A. niger* (www.jgi.doe.gov), *Neosartorya fischeri*, teleomorph of *A. fischerianus* (Fedorova et al., 2008), *A. fumigatus* (Nierman et al., 2005), *A. clavatus* (Fedorova et al., 2008), *A. nidulans* (Galagan et al., 2005).

Putatively orthologous sequences among all species analyzed were identified running a BLAST query of the whole set of proteins, following the best-reciprocal-hit (BRH) criteria. A group of sequences were defined as orthologs, when the BRH results were exclusively limited to sequences of their own set. As a result 3,191 groups of orthologs were assembled. Proteins were aligned using MUSCLE (Edgar, 2004) and subsequently poorly aligned regions and gaps were removed from the alignment using Gblock (Talavera and Castresana, 2007). The phylogenetic tree for each set of orthologous proteins was inferred using a Maximum Likelihood method with an amino acid LG+G model by means of Phyml version 3.0 (Guindon and Gascuel, 2003). The default SH-like test was used to evaluate branch supports. Finally, a consensus tree was inferred from the 3,191 phylograms using sumtrees.py (Sukumaran and Holder, 2010).

2.2. Codon usage analysis and t-RNA prediction

Codon usage, correspondence analysis (COA) (Greenacre, 1984), base composition and the relative synonymous codon usage (RSCU) (Sharp and Li, 1986) were calculated using the CodonW 1.4.2 program (written by John Peden and available at http://sourceforge.net/projects/codonw/). To confirm the results obtained by COA, within-group correspondence analysis (WCA) (Charif et al., 2005; Suzuki et al., 2008) was calculated as implemented in ADE4 package of R (Thioulouse et al., 1997). Moreover, for each gene, we computed MELP values (Supek and Vlahovicek, 2005) using INCA2.1 (Supek and Vlahovicek, 2004). MELP is an expression level predictor, independent of length and composition of the sequences. This index measures the similarity of codon usage between genes and a group of selected HEGs. It is derived from a method for quantifying the codon usage bias in genes or gene groups, based on a corrected log-ratio χ^2 statistic (Supek and Vlahovicek, 2005).

Contingency χ^2 tests were used to identify which triplets are significantly preferred in putative HEGs ($p<0.01$) and were considered as translationally optimal codons. Results were confirmed with optimal codons detected by COA comparing the codon usage pattern of genes at both extremes of the axis related to expression level (10% of genes at either extreme).

The strength of selection on codon usage was calculated according to Sharp et al. (2005), using the reference set of 160 HEGs (see below).

We predicted the tRNA genes in each genome by using tRNAscan-SE search server, version 1.21 (Lowe and Eddy, 1997) in mixed (general tRNA model) mode with default parameters.

Table 1
Main results for each species.

Organism	Genomic GC%	No. genes	%Var[a]	r (MELP)[b]	r (dS)[c]	r (GC_3)[d]	r ($GC_{introns}$)[e]	S[f] (Sharp)
Aspergillus flavus	48.35	12,604	14.2	0.92	>0.71[g]	0.77	0.07	0.95
Aspergillus oryzae	48.24	12,063	14.9	0.92	>0.73[g]	0.77	0.12	0.97
Aspergillus terreus	52.88	8900	23.4	0.78	>0.62	0.95	0.36	1.09
Aspergillus niger	50.33	10,701	17.0	0.86	>0.60	0.84	0.10	0.95
Neosartorya fischeri	49.43	10,406	15.8	0.87	>0.64[g]	0.81	0.17	0.89
Aspergillus fumigatus	49.80	9887	15.3	0.78	>0.65[g]	0.88	0.22	0.75
Aspergillus clavatus	49.21	9121	19.8	0.71	>0.66	0.95	0.26	0.89
Aspergillus nidulans	50.32	10,701	13.4	0.87	>0.75	0.77	0.12	0.89

[a] Variability explained by the first axis generated by COA (correspondence analysis) on RSCU (relative synonymous codon usage) values. This axis is always related to expression levels.

[b] Correlation coefficients of the positions of the genes on the first axis of COA against the respective MELP (expression level predictor) value.

[c] Minimum correlation coefficient of the dS (synonymous distance) estimation for orthologs with their respective MELP values (see Supplementary Table 1).

[d] Correlation coefficients of the positions in the expression-related axis with GC_3 of the genes.

[e] Correlation coefficients of the GC content at introns with GC content at the third codon position of the respective genes.

[f] Estimation of the strength of selected codon usage bias (S) as developed by Sharp et al. (2005) based on the 160 genes defined as highly expressed in all species.

[g] Estimations between the closely related species *A. flavus* and *A. oryzae* as well as between *N. fischeri* and *A. fumigatus* are not considered (see text for details).

2.3. Identification of conserved HEGs

We defined 160 orthologous genes as highly expressed in all the species based on the microarray data generated in the study of Andersen et al. (2008), that represent 5% of the total orthologous genes. This study predicted expression level for all genes in *A. nidulans*, *A. niger* and *A. oryzae* in two physiological conditions. Correlation analyses show that there is a slight variation between replicates and conditions, especially at the top 10% of expression (r>0.89, p<0.001). Thus, we averaged replicates and conditions for each species and selected orthologous genes that are simultaneously within this range in the three studied organisms.

This comparative transcriptomics study of three relatively distant *Aspergillus* species allowed us to propose a conserved core of HEGs. These genes were subsequently considered as a "reference set" for codon usage analysis (see above). Not surprisingly among these sequences we found many genes known to be highly expressed (i.e. translation elongation factors, ribosomal proteins, HSP 70, enzymes from the Krebs cycle, tubulin, etc.).

Table 2
Putative optimal codons and number of cognate tRNA genes detected for each species analyzed.

Table 2 (*continued*)

		Aspergillus flavus	*Aspergillus oryzae*	*Aspergillus terreus*	*Aspergillus niger*	*Neosartorya fischeri*	*Aspergillus fumigatus*	*Aspergillus clavatus*	*Aspergillus nidulans*
Ala	GCU	xx 10	xx 10	xx 6	xx 12	xx 8	xx 8	xx 8	xx 8
	GCC	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0
	GCA	3	3	2	9	3	3	3	2
	GCG	3	3	3	3	2	2	2	3
Arg	CGU	xx 12	xx 12	xx 8	xx 15	xx 11	xx 9	xx 14	xx 9
	CGC	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0
	CGA	3	2	1	2	2	2	2	2
	CGG	3	3	3	3	2	2	2	2
	AGA	3	2	2	0	8	1	3	2
	AGG	2	2	2	1	2	2	2	2
Asn	AAU	0	0	0	0	0	0	0	0
	AAC	xx 9	xx 9	xx 1	xx 0	xx 14	xx 5	xx 8	xx 7
Asp	GAU	0	0	1	1	0	0	0	0
	GAC	xx 13	xx 13	xx 8	xx 14	xx 10	xx 9	xx 14	xx 9
Cys	UGU	0	0	0	0	0	0	0	0
	UGC	xx 3	xx 3	xx 3	x 4	xx 3	xx 3	xx 3	xx 3
Gln	CAA	3	4	2	4	2	2	2	2
	CAG	xx 7	xx 7	xx 5	xx 8	xx 8	xx 6	xx 6	xx 5
Glu	GAA	7	8	3	5	15	4	8	3
	GAG	xx 1	xx 11	xx 7	xx 12	xx 8	xx 8	xx 8	xx 8
Gly	GGU	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0
	GGC	x 15	x 15	9	17	12	11	x 11	11
	GGA	2	4	3	6	3	3	3	3
	GGG	1	1	1	1	1	1	1	1
His	CAU	0	0	1	0	0	0	0	0
	CAC	xx 7	xx 8	xx 3	xx 6	xx 16	xx 4	xx 7	xx 5
Ile	AUU	10	10	7	11	7	7	7	7
	AUC	xx 0	xx 1	xx 0	xx 0	xx 9	xx 0	xx 7	xx 1
	AUA	2	2	2	1	2	1	1	1
Leu	UUA	2	2	0	2	1	1	0	1
	UUG	0	0	0	3	3	2	2	2
	CUU	xx 0	xx 0	0	x 9	6	x 6	6	xx 6
	CUC	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0
	CUA	1	1	1	2	0	0	0	2
	CUG	xx 1	xx 4	xx 4	xx 5	xx 4	xx 4	xx 4	3
Lys	AAA	4	5	1	2	12	1	9	3
	AAG	xx 1	xx 11	xx 7	xx 15	xx 9	xx 8	xx 9	xx 8
Phe	UUU	0	0	0	0	0	0	0	0
	UUC	xx 7	xx 8	xx 4	xx 10	xx 18	xx 5	xx 9	xx 5
Pro	CCU	xx 9	xx 9	x 2	xx 9	xx 5	xx 5	xx 5	xx 6
	CCC	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0
	CCA	2	2	1	0	0	2	2	2
	CCG	1	0	0	3	2	2	3	2
Ser	UCU	xx 8	xx 7	xx 5	xx 8	xx 5	xx 5	xx 5	xx 5
	UCC	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0
	UCA	1	1	1	2	2	1	1	1
	UCG	2	3	2	4	2	2	2	2
	AGU	0	0	0	0	0	0	0	0
	AGC	xx 6	xx 6	x 4	6	xx 4	xx 4	xx 3	x 4
Thr	ACU	xx 9	xx 10	xx 6	xx 10	xx 6	xx 6	xx 6	xx 6
	ACC	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0
	ACA	2	3	0	2	15	3	7	2
	ACG	2	2	3	3	2	2	2	2
Tyr	UAU	0	0	0	0	1	0	0	0
	UAC	xx 11	xx 12	xx 6	xx 8	xx 5	xx 5	xx 5	xx 5
Val	GUU	xx 11	xx 10	x 6	xx 11	xx 7	xx 7	xx 8	xx 8
	GUC	xx 1	xx 1	xx 0	xx 0	xx 0	xx 0	xx 0	xx 0
	GUA	3	4	1	2	13	1	6	1
	GUG	4	4	3	4	2	2	2	2

Contingency χ^2 tests were used to find which triplets are significantly preferred by the reference set, which comprises 160 orthologous genes defined as highly expressed. p values<0.001 and<0.01 are indicated with "xx" and "x", respectively. The number of copies of cognate tRNA genes are indicated for each species. "Gray" rows indicate universal optimal codons as defined by Sharp et al. (2005).

2.4. Molecular distances

Putatively orthologous sequences among the eight species were identified and aligned as mentioned above. The retrieved aligned proteins were back-translated to the known DNA sequences. Pairwise synonymous distances (Sharp and Li, 1987) were estimated using Codeml program, included in PAML 4.4 package (Yang, 2007). The analyses were performed with pairs of sequences displaying dS values≤1.5 and a minimal identity value of 40% (Supplementary Tables 1 and 2).

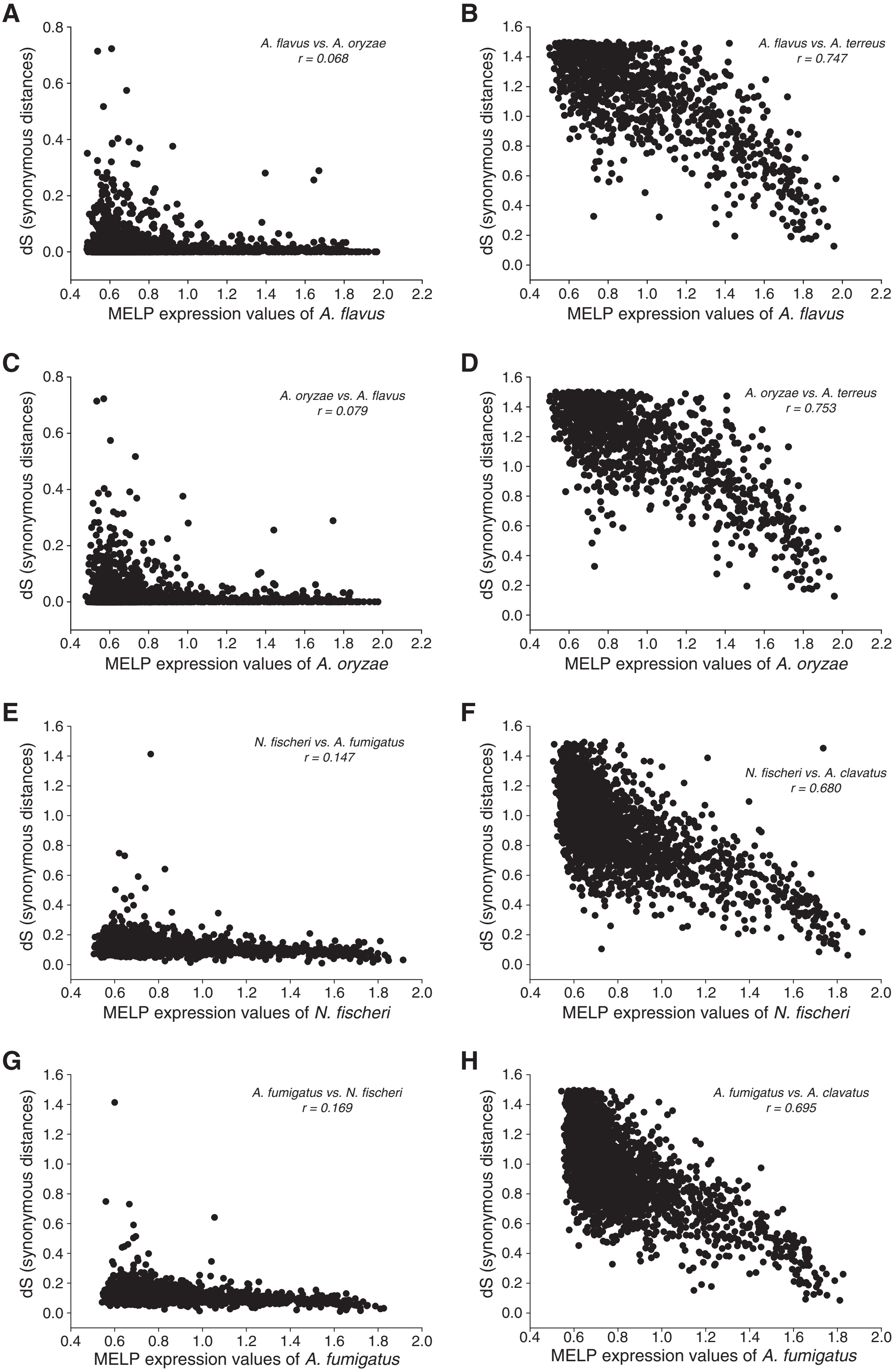


Fig. 2. Regression plots of MELP expression values against estimated dS. Plot of the MELP (expression level predictor, axis x) values against dS (synonymous distance, axis y) between the pair of organisms indicated in each plot.

2.5. Comparison of codon usage between conserved and non conserved gene regions

Conserved (CR) and non-conserved regions (NCR) were identified and splitted using Gblock with the following parameters –b1 = 8, –b2 = 8, –b3 = 1 and –b4 = 2 (for details, see Talavera and Castresana, 2007). This means that blocks with a length of at least two amino acids, unchanged in the eight species, with a maximum of one contiguous non-conserved position are pooled for subsequent studies. The percentage and standard deviation of codons in CR were 53.3 ± 1.1 and in NCR 46.7 ± 1.1. In order to test the robustness of the method, pairwise p-distances were calculated for each region in each gene with the default protein model using the *distmat* program implemented in the EMBOSS package. This distance is the proportion (p) of amino acid sites at which the two sequences compared are different (Supplementary Fig. 1). Global codon usage in both regions were compared for each species as mentioned in 2.2. Following the method described in 2.3, we identified the 30% of the genes showing the highest and lowest levels of expression. Both groups were analyzed independently (Supplementary Table 3).

3. Results and discussion

3.1. Codon usage analyses

A molecular phylogeny of the eight *Aspergillus* species considered in this work was inferred from the composite data (Fig. 1). The phylogenetic tree was basically in agreement with previous reconstructions for this group (Wang et al., 2009). For each species, a codon usage table is reported as Supplementary Table 4. Correspondence analysis (COA) on the Relative Synonymous Codon Usage (RSCU) values suggests that synonymous codon usage bias is associated with expression levels in the analyzed species, since in all cases we found that genes coding for ribosomal proteins and other Highly Expressed Genes (HEGs) are clustered at one side of the first axis (see Supplementary Fig. 2). We stress that the same result was obtained when COA was done on codon counts and with within-group COA (WCA, Charif et al., 2005). This is deduced from the high correlation coefficients obtained among the main axes generated by COA on RSCU and the ones generated by the other two methods ($r > 0.88$). The first axes accounted for between 13.4% and 23.4% of the total variability, depending on the organism. On the other hand, Lowly Expressed Genes (LEGs) and genes encoding hypothetical proteins are distributed almost randomly along those axes. Besides, for each species we found a strong and significant correlation between the MELP (expression level predictor) value for each gene and the respective position along the “expression” axis (Table 1). We remark that the reference set, which was based on microarray data, is not only composed by genes encoding ribosomal proteins but also by other HEGs.

It is interesting to note the relatively high differences found in each species in the variability explained by the first axis (see Table 1). We point out that as this factor increases, the correlation of the MELP index with the expression axis decreases ($r = -0.68$). These observations can be explained by considering that this axis is not only related to expression levels but also to GC_3 content, and that translational optimal codons tend to be GC-rich at third codon positions (see Table 2). Therefore, we suggest that mutational bias and translational selection (fixing G/C– ended codons) are superimposed in the detected trend, and therefore drive the genome in the same direction. This is confirmed by the correlation found in each species between GC_3 and the GC content of the corresponding introns (Table 1), which reflects strongly the most neutral changes (mutational bias).

When the reference set was compared to the rest of the genes, we found a conserved core of 21 optimal codons (χ^2, $p < 0.01$) for all species and at least 23 optimal triplets in each member of this genus (Table 2). Among these codons, at least 67% perfectly match the respective cognate tRNAs present in each genome (Table 2). These results not only suggest that the triplets that we detected are effectively translationally optimal but also that they are highly conserved across the studied species. We should note, however, that some optimal codons (for instance, CCC, GCC and AUC) do not display a cognate tRNA, in some or all species. This is most probably due to modifications in the first position of the anticodon in some isoacceptor tRNA, which permits these tRNAs to recognize the optimal triplet, among other synonymous codons. In spite of the high conservation of optimal codons, there are some interesting differences concerning tRNAs. For example, there are nine tRNA genes complementary to AUC in *N. fischeri*, but zero in *A. fumigatus*. On the other hand, the ratio of AAA:AAG is 12:9 in *N. fischeri*, but 1:8 in *A. fumigatus*. These differences are striking, given the very close relationship of these species. At least three non-mutually exclusive scenarios could explain these differences. First, the variation in tRNA gene number in each genome could be compensated by different expression levels, which can lead to similar intracellular concentrations. Second, different modifications in the sequence of the tRNA molecules can exist (not only at the first anticodon positions), which could “hide” the anticodon prediction. Third, the observed variations in tRNA genes could be due to differences in the evolutionary plasticity of each genome, i.e., unequal rate of duplications, deletions, rearrangements, etc.

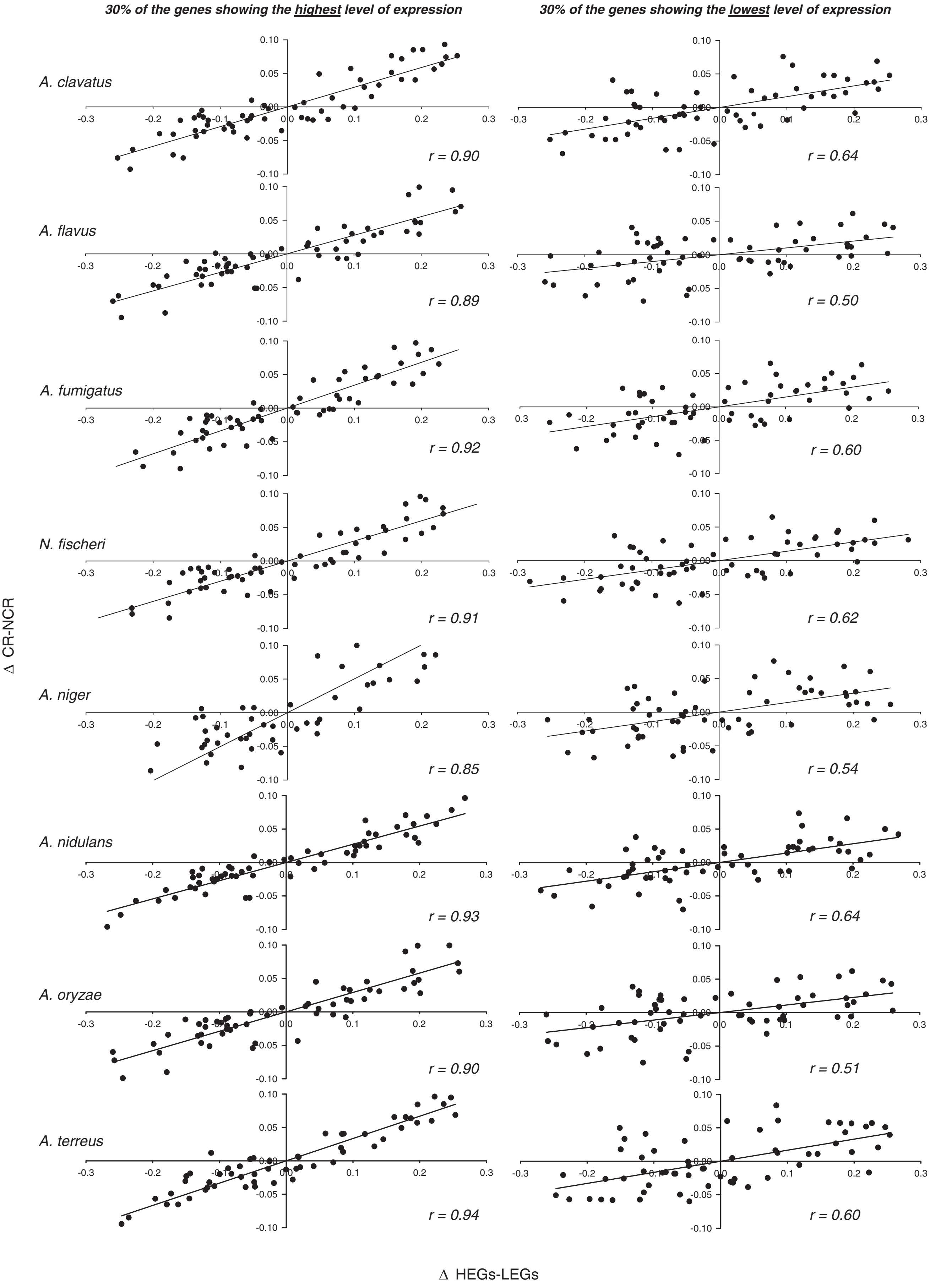
As mentioned above, we used orthologous sequences as the reference set, so in order to avoid misleading results from partial data we repeated all the analyses for *A. nidulans*, *A. oryzae* and *A. niger* using the whole data set from the microarray experiment reported by Andersen et al. (2008). The new defined reference sets comprised the top 5% HEGs of the total data set of each species ($n > 530$ genes). As expected, we found no significant differences with the reported results (data not shown).

We found that the percentages of variation explained by the “expression” axes were relatively high, but variable, which indicates that some variation exists among the species. Since neutral evolutionary forces, like mutational bias or genetic drift, might have affected differentially the codon usage in these genomes, it is not surprising to find some level of variation across the genus as shown by the estimated selection coefficient (Table 1). Codon usage analyses were also performed only with the orthologous sequences, which comprised 3,191 genes from each genome. Theses analyses confirmed the results obtained with all genes (data not shown). The two main axes generated with the orthologous sequences show a high correlation with the ones generated using the total set of genes from each species ($r \geq 0.98$, $p < 0.001$ and $r \geq 0.91$, $p < 0.001$; respectively), which shows that “species-specific” genes (acquired by duplication, horizontal gene transfer, etc.) have a very minor effect on the intragenomic codon usage patterns.

3.2. Estimation of molecular distances

In a pioneering work Sharp and Li (1987) showed that when natural selection is operative at the level of translation, there is a negative correlation between synonymous divergence and the expression levels. Following this idea, we found that each species displays a high

Fig. 3. Conserved regions in HEGs (highly expressed genes) and LEGs (lowly expressed genes) prefer translationally optimal codons.The differences in the relative frequencies of codon usage between conserved and non-conserved regions (Δ CR-NCR) are plotted against the difference of relative frequencies of codon usage between HEGs and LEGs (Δ HEGs-LEGs). Each point is a codon, and the x value of that triplet is the Δ between HEGs and LEGs, and is the same in the two columns for each species. The y axis value is for conserved regions for HEGs on the left column, while it is for LEGs on the right column. Note that for each species, the r value is always higher for conserved regions in HEGs.


30% of the genes showing the highest level of expression
30% of the genes showing the lowest level of expression
A. clavatus
r = 0.90
r = 0.64
A. flavus
r = 0.89
r = 0.50
A. fumigatus
r = 0.92
r = 0.60
N. fischeri
r = 0.91
r = 0.62
A. niger
r = 0.85
r = 0.54
A. nidulans
r = 0.93
r = 0.64
A. oryzae
r = 0.90
r = 0.51
A. terreus
r = 0.94
r = 0.60
0.10
0.05
0.00
-0.05
-0.10
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
Δ CR-NCR
Δ HEGs-LEGs

correlation coefficient value ($-0.75 \leq r \leq -0.60$; $p<0.001$) between synonymous distances (dS) and MELP values (Table 1). The correlation values found for every pair of species are shown in Supplementary Table 1. That table shows, as expected, that HEGs display less synonymous changes than LEGs. The only exception is found in the comparison between extremely closely related species, which can be easily explained when considering the lack of time for accumulating substitutions.

These results are important for at least three complementary reasons: first, from a parsimonious point of view, natural selection for codon usage was operative in the last common ancestor of all these species. Second, purifying selection is still operative in all of the analyzed species (although probably with different strength). Third, as a consequence of the latter point and as above mentioned, synonymous sites had diverged less in the HEGs than in the rest of the sequences. Not surprisingly, there is no correlation in the comparisons between the closely related pair of species *A. flavus* with *A. oryzae* and *A. fumigatus* with *N. fischeri* (see Fig. 2 and Supplementary Table 1). Indeed, the scatter of the points in the plane shows that HEGs do not reach high dS values, while the rest of the genes can "freely" diverge and consequently they reach high dS figures. This interpretation is supported by the high r values found in the comparisons with more distant species (Fig. 2).

3.3. Mutational bias does not explain the evolution of third codon positions

A strictly neutral model states that codon usage mainly results from biases in the mutational effect. Therefore mutational pressure should influence in a similar way the regions and codon positions less affected by natural selection, namely third silent sites and introns (Francino and Ochman, 2001; Musto et al., 1997). In order to test if this is the case among the Aspergilli, we studied the correlations that hold between these positions and sites.

As mentioned before, in these species the majority of optimal codons contain a G or a C in the third position and therefore HEGs show a bias towards high GC content. Contrary to the neutral prediction stated above, we found that the GC content of introns is correlated neither with gene expression level nor with GC content at synonymous sites. These observations suggest that the effect of mutational bias on the observed codon usage is negligible or absent in the majority of the species analyzed, with the exception of *A. terreus* and *A. clavatus*, since the latter organisms show a marginal correlation (see Table 1). Needless to say, these observations support the idea that translational selection for codon usage is operative among these species. The minor differences that we found between species are probably due to genetic drift and/or a small effect of mutational bias acting idiosyncratically across species-specific lineages.

3.4. Analysis of codon usage in relation to accuracy

It is known that the misincorporation of an amino acid can affect fitness, depending on how this substitution impact on protein functions. This was first studied by Akashi (1994) in *Drosophila* species. In order to understand if this happens too among the species studied here, we made the following analysis. Every orthologous protein was splitted in two regions: conserved and non-conserved, which were then pooled. By "conserved" we understand blocks with a length of at least two amino acids unchanged in the eight species, with a maximum of one contiguous non-conserved position. Then, codon usage in each "pool" for each species was calculated. The difference (Δ) between the relative frequency of codon usage in conserved *vs*. non conserved regions was plotted against the Δ of the relative frequency of codon usage between HEGs and LEGs. The latter value is a measure of the bias towards the usage of translational optimal triplets among HEGs in relation to the rest (mainly associated with the speed of translation, see above). Fig. 3 shows that the frequency of optimal codons is higher at conserved amino acid regions than in the rest of the protein, given that the preferred codons at conserved regions (positive values at the x axis) are also preferred by HEGs (positive values at y axis). This is true not only for the 30% of the genes showing the highest levels of expression (left column in Fig. 3), but also for the 30% with the lower expression levels (right column in Fig. 3), although the correlation is higher in the first case. We note that the slopes are also higher for CR in HEGs, which implies that the selective differences between conserved and non-conserved regions are smaller in LEGs than in HEGs. Furthermore, the relative contribution of selection for accuracy to the differences observed between HEGs and LEGs is marginal as is suggested by both the slopes and the ranges of the values on each axis. Taken together, these results indicate that there is a strong link between synonymous codon usage and amino acid constraint, which in turn implies that natural selection is acting at the level of accuracy. In other words, translational optimal codons are more frequent and significantly more used (see Supplementary Table 3) not only at heavily expressed genes but also at highly conserved protein regions.

4. Conclusions

Here we report that *i*) natural selection acting at the level of translation (simultaneously at two levels: speed and accuracy), is the main force shaping codon usage variation among genes in each of the species analyzed; *ii*) the role of mutational bias in this variation is rather marginal, since all species display more or less the same genomic GC content and the base composition of introns evolve independently of exons (and mainly of their third codon positions); *iii*) we have defined for each species a set of "optimal codons", and the majority of them are conserved in the whole genus; *iv*) the "universal optimal codons" (Sharp et al., 2005) are among this group; and finally *v*) natural selection is acting on the extant species since their split from the last common ancestor.

We highlight the possible biotechnological application of the data presented here regarding heterologous expression of non-fungal proteins in the Aspergilli. Codon optimization as a strategy to obtain higher yields in the heterologous production of proteins in members of this genus has been proved successful for the expression of a synthetic gene encoding potato alpha-glucan phosphorylase in *A. niger* (Koda et al., 2005). Therefore, we believe that our report on the dilucidation of codon usage in the members of the genus *Aspergillus* can lead to a more rational design of the primary DNA sequence encoding a desired protein towards the optimization of its heterologous production. In addition to growing knowledge of other factors affecting expression, like gene regulation at the level of promoters, positional effects on gene expression, mRNA processing and stability, etc., our work can lead to a strengthening of the value of *Aspergillus* as a host for heterologous protein production.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2012.06.027.

Acknowledgments

This work was supported by Fondo Clemente Estable 2007, FCE_722 (ANII, Uruguay), Fondo Clemente Estable 2009, FCE_2776 (ANII, Uruguay) and PEDECIBA (Uruguay). We also thank the contributions made by two anonymous referees.

References

Akashi, H., 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics 136, 927–935.

Andersen, M.R., Vongsangnak, W., Panagiotou, G., Salazar, M.P., Lehmann, L., Nielsen, J., 2008. A trispecies *Aspergillus* microarray: comparative transcriptomics of three *Aspergillus* species. Proc. Natl. Acad. Sci. U. S. A. 105, 4387–4392.

Bulmer, M., 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129, 897–907.

Charif, D., Thioulouse, J., Lobry, J.R., Perriere, G., 2005. Online synonymous codon usage analyses with the ade4 and seqinR packages. Bioinformatics 21, 545–547.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

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

Fedorova, N.D., et al., 2008. Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus*. PLoS Genet. 4, e1000046.

Francino, M.P., Ochman, H., 2001. Deamination as the basis of strand-asymmetric evolution in transcribed *Escherichia coli* sequences. Mol. Biol. Evol. 18, 1147–1150.

Galagan, J.E., et al., 2005. Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. Nature 438, 1105–1115.

Grantham, R., 1980. Nucleic acid sequence similarities: 'poly(A) tendency'. FEBS Lett. 121, 193–199.

Grantham, R., Gautier, C., Gouy, M., Jacobzone, M., Mercier, R., 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 9, r43–r74.

Greenacre, M.J., 1984. Theory and applications of correspondence analysis. Academic Press, London.

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

Gurr, S.J., Unkles, S.E., Kinghorn, J.R., 1987. The structure and organization of nuclear genes of filamentous fungi. In: Kinghorn, J.R. (Ed.), Gene Structure in Eukaryotic Microbes. IRL Press, Oxford, pp. 93–139.

Hershberg, R., Petrov, D.A., 2009. General rules for optimal codon choice. PLoS Genet. 5, e1000556.

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

Kale, S.P., Cary, J.W., Baker, C., Walker, D., Bhatnagar, D., Bennett, J.W., 2003. Genetic analysis of morphological variants of *Aspergillus parasiticus* deficient in secondary metabolite production. Mycol. Res. 107, 831–840.

Koda, A., Bogaki, T., Minetoki, T., Hirotsune, M., 2005. High expression of a synthetic gene encoding potato alpha-glucan phosphorylase in *Aspergillus niger*. J. Biosci. Bioeng. 100, 531–537.

Lloyd, A.T., Sharp, P.M., 1991. Codon usage in Aspergillus nidulans. Mol. Gen. Genet. 230, 288–294.

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

Machida, M., et al., 2005. Genome sequencing and analysis of *Aspergillus oryzae*. Nature 438, 1157–1161.

Musto, H., Caccio, S., Rodriguez-Maseda, H., Bernardi, G., 1997. Compositional constraints in the extremely GC-poor genome of *Plasmodium falciparum*. Mem. Inst. Oswaldo Cruz 92, 835–841.

Nierman, W.C., et al., 2005. Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. Nature 438, 1151–1156.

Payne, G.A., Yu, J., Nierman, W.C., Machida, M., Bhatnagar, D., Cleveland, T.E., Dean, R.A., 2008. A first glance into the genome sequence of Aspergillus flavus. In: Goldman, G.H., Osmani, S.A. (Eds.), The Aspergilli: Genomics, Medical Aspects, Biotechnology, and Research MMethods. CRC Press, Boca Raton, pp. 15–23.

Plotkin, J.B., Kudla, G., 2011. Synonymous but not the same: the causes and consequences of codon bias. Nat. Rev. Genet. 12, 32–42.

Scazzocchio, C., 2006. *Aspergillus* genomes: secret sex and the secrets of sex. Trends Genet. 22, 521–525.

Sharp, P.M., Cowe, E., 1991. Synonymous codon usage in *Saccharomyces cerevisiae*. Yeast 7, 657–678.

Sharp, P.M., Li, W.H., 1986. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. Nucleic Acids Res. 14, 7737–7749.

Sharp, P.M., Li, W.H., 1987. The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol. Biol. Evol. 4, 222–230.

Sharp, P.M., Cowe, E., Higgins, D.G., Shields, D.C., Wolfe, K.H., Wright, F., 1988. Codon usage patterns in *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. Nucleic Acids Res. 16, 8207–8211.

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 Res. 33, 1141–1153.

Sueoka, N., 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. U. S. A. 48, 582–592.

Sukumaran, J., Holder, M.T., 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26, 1569–1571.

Supek, F., Vlahovicek, K., 2004. INCA: synonymous codon usage analysis and clustering by means of self-organizing map. Bioinformatics 20, 2329–2330.

Supek, F., Vlahovicek, K., 2005. Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. BMC Bioinformatics 6, 182.

Suzuki, H., Brown, C.J., Forney, L.J., Top, E.M., 2008. Comparison of correspondence analysis methods for synonymous codon usage in bacteria. DNA Res. 15, 357–365.

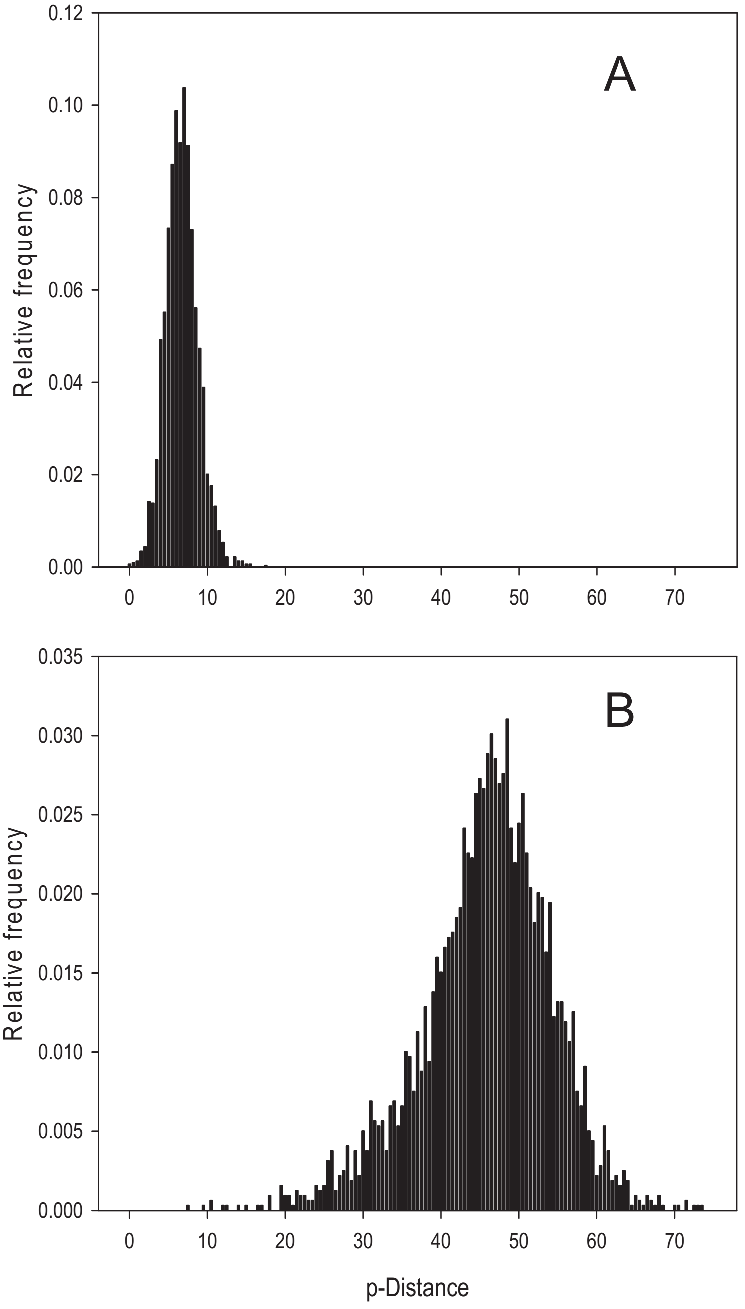
Talavera, G., Castresana, J., 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56, 564–577.

Thioulouse, J., Chessel, D., Dolédec, S., Olivier, J.M., 1997. ADE-4: a multivariate analysis and graphical display software. Stat. Comput. 7, 75–83.

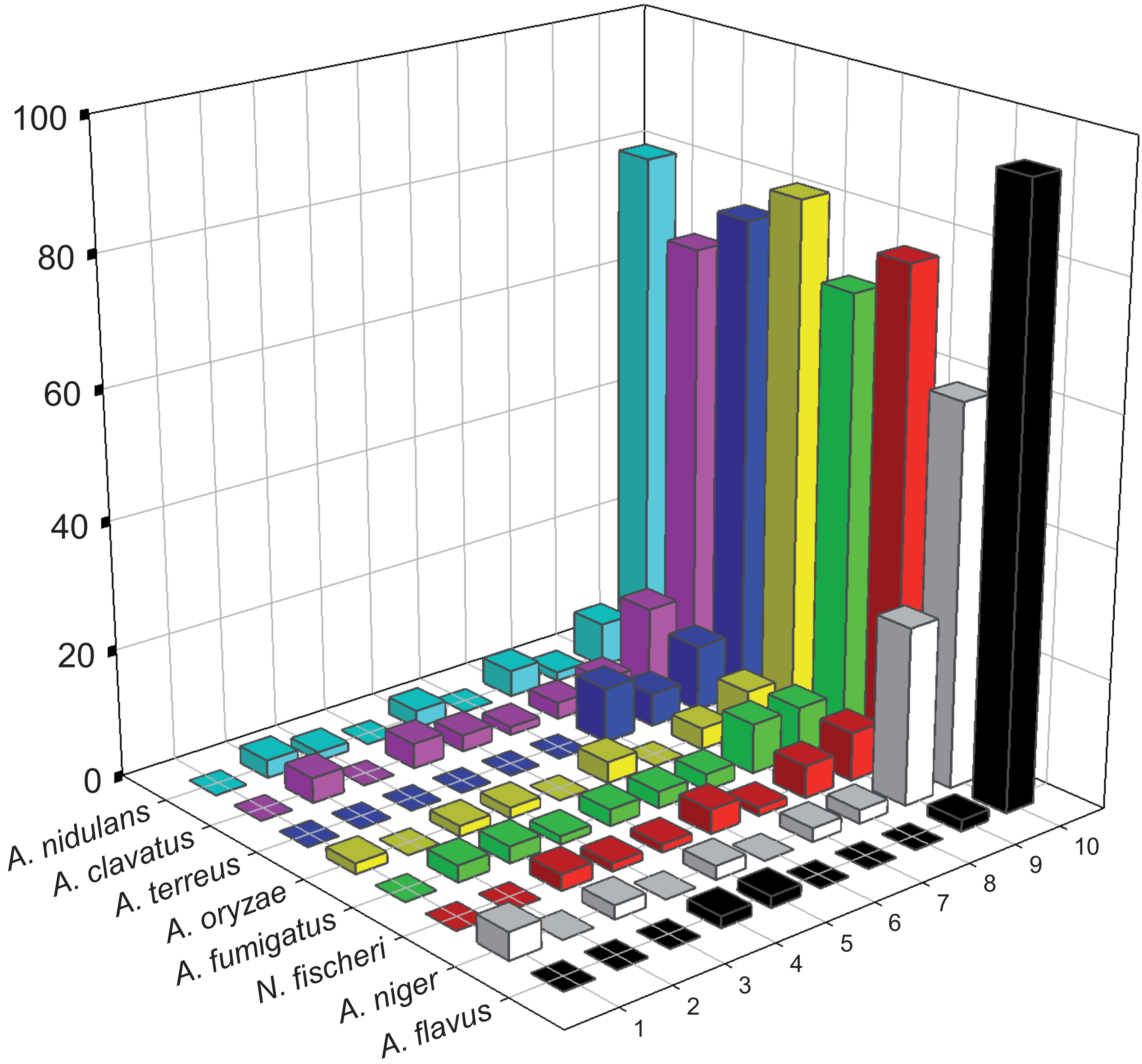
Wang, H., Xu, Z., Gao, L., Hao, B., 2009. A fungal phylogeny based on 82 complete genomes using the composition vector method. BMC Evol. Biol. 9, 195.

Yamao, F., Andachi, Y., Muto, A., Ikemura, T., Osawa, S., 1991. Levels of tRNAs in bacterial cells as affected by amino acid usage in proteins. Nucleic Acids Res. 19, 6119–6122.

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



Supplementary Fig. 1. Estimated p-distance of conserved and non-conserved regions. Histogram of the frequency of conserved (A) and non-conserved (B) regions in the 3191 orthologous proteins distributed according to the mean p-distance [this distance is the proportion (p) of amino acid sites at which the two sequences compared are different]. Conserved regions are aligned blocks with a length of at least two amino acids, which are unchanged in the eight species and with a maximum of one contiguous non-conserved position. The graph shows that the protein regions selected as conserved using the software Gblock are also characterized by less diverged sequences.



Supplementary Fig. 2. Distribution of genes encoding ribosomal proteins. Histogram of the distribution of genes encoding ribosomal proteins along the axis related to expression levels (see Table 1) for each species. The axis was divided into 10 parts, each of them containing an equal number of genes.

Supplementary Table 1.

Correlation coefficients between the expression level predictor MELP, calculated for each gene in each organism (Organisms MELP) and the respective pairwise synonymous divergence (dS).

Genes with high MELP values are also characterized by small levels of dS. The particular case of *N. fischeri vs. A. fumigatus* and *A. flavus vs. A. oryzae* with a significant but rather small correlation values may be explained by a low time of divergence. All values correspond to significant correlations coefficients ($p < 0.001$).

Organism (MELP)	*A. flavus*	*A. oryzae*	*A. terreus*	*A. niger*	*N. fischeri*	*A. fumigatus*	*A. clavatus*	*A. nidulans*
Aspergillus flavus		-0.068	-0.747	-0.769	-0.752	-0.752	-0.771	-0.715
Aspergillus oryzae	-0.079		-0.753	-0.759	-0.749	-0.750	-0.771	-0.734
Aspergillus terreus	-0.718	-0.725		-0.724	-0.629	-0.630	-0.642	-0.706
Aspergillus niger	-0.684	-0.671	-0.692		-0.655	-0.630	-0.668	-0.606
Neosartorya fischeri	-0.733	-0.734	-0.647	-0.753		-0.147	-0.680	-0.738
Aspergillus fumigatus	-0.736	-0.741	-0.653	-0.750	-0.169		-0.695	-0.746
Aspergillus clavatus	-0.758	-0.761	-0.666	-0.762	-0.709	-0.713		-0.749
Aspergillus nidulans	-0.754	-0.768	-0.763	-0.779	-0.778	-0.793	-0.785	

Supplementary Table 2.

Number of putative orthologous sequences. Only sequences displaying *dS* values ≤ 1.5 and a minimal identity of 40% are shown.

Organism (MELP)	*A. flavus*	*A. oryzae*	*A. terreus*	*A. niger*	*N. fischeri*	*A. fumigatus*	*A. clavatus*	*A. nidulans*
Aspergillus flavus								
Aspergillus oryzae	3187							
Aspergillus terreus	1042	1048						
Aspergillus niger	1088	1089	1190					
Neosartorya fischeri	1101	1118	1075	1029				
Aspergillus fumigatus	1008	1001	981	912	3190			
Aspergillus clavatus	1101	1092	1275	1114	3104	3070		
Aspergillus nidulans	525	520	618	671	587	539	631	

Supplementary Table 3.

Contingency $\chi 2$ tests were used to find which codons are significantly preferred among HEGs, and conserved regions (CR). p values < 0.01.

Optimal codons tend to be significantly more used at conserved amino acid regions. This is true not only for the 30% of the genes showing the highest level of expression, but also for the 30% with the lower expression levels.

		Aspergillus flavus			*Aspergillus oryzae*			*Aspergillus terreus*			*Aspergillus niger*			*Neosartorya fischeri*			*Aspergillus fumigatus*			*Aspergillus clavatus*			*Aspergillus nidulans*		
		HEGs	**CR** *(30% highest expression)*	**CR** *(30% lowest expression)*	**HEGs**	**CR** *(30% highest expression)*	**CR** *(30% lowest expression)*	**HEGs**	**CR** *(30% highest expression)*	**CR** *(30% lowest expression)*	**HEGs**	**CR** *(30% highest expression)*	**CR** *(30% lowest expression)*	**HEGs**	**CR** *(30% highest expression)*	**CR** *(30% lowest expression)*	**HEGs**	**CR** *(30% highest expression)*	**CR** *(30% lowest expression)*	**HEGs**	**CR** *(30% highest expression)*	**CR** *(30% lowest expression)*	**HEGs**	**CR** *(30% highest expression)*	**CR** *(30% lowest expression)*
Ala	GCA																								
	GCC	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	GCG									x						x			x			x			
	GCU	x	x		x	x		x			x			x			x			x			x	x	
Arg	AGA																								
	AGG																								
	CGA			x			x																		
	CGC	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	CGG			x			x						x			x			x			x			x
	CGU	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Asn	AAC	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	AAU																								
Asp	GAC	x	x		x	x		x			x	x	x	x			x			x			x	x	x
	GAU																								
Cys	UGC	x			x			x			x	x	x	x			x	x		x			x		
	UGU																								
Gln	CAA																								
	CAG	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
Glu	GAA																								
	GAG	x	x		x	x		x	x	x	x	x		x	x		x	x		x	x		x	x	
Gly	GGA						x																		
	GGC	x		x	x		x			x			x			x			x	x		x			x
	GGG																								
	GGU	x	x		x	x			x		x	x		x	x		x	x		x	x		x	x	
His	CAC	x	x		x	x		x	x	x	x	x		x	x		x	x		x	x	x	x	x	x
	CAU																								
Ile	AUA																								
	AUC	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	AUU																								
Leu	CUA																								
	CUC	x			x			x			x	x	x	x			x		x	x		x	x		
	CUG	x	x		x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x			
	CUU	x			x						x						x						x		
	UUA																								
	UUG																								

Lys	AAA																								
	AAG	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x		x	x	x	x	x	
Phe	UUC	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x	x
	UUU																								
Pro	CCA																								
	CCC	x	x		x	x		x	x	x	x	x		x	x	x	x	x	x	x	x	x	x	x	
	CCG			x			x			x			x			x			x			x			
	CCU	x			x			x			x			x			x			x			x		
Ser	AGC	x	x		x			x						x			x	x		x			x		
	AGU																								
	UCA																								
	UCC	x	x	x	x	x	x	x	x		x	x	x	x	x	x	x		x	x	x	x	x	x	x
	UCG			x			x			x			x			x			x			x			x
	UCU	x			x			x			x			x			x			x			x		
Thr	ACA																								
	ACC	x	x		x	x		x	x	x	x	x		x	x	x	x	x	x	x	x	x	x	x	
	ACG			x			x			x			x			x			x			x			x
	ACU	x			x			x			x			x			x			x			x		
Tyr	UAC	x	x		x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
	UAU																								
Val	GUA																								
	GUC	x	x		x	x		x	x		x	x	x	x	x	x	x	x	x	x	x		x	x	x
	GUG									x									x						
	GUU	x			x			x			x			x			x			x			x		

Supplementary Table 4.

Codon usage table for each species. Codon count and the RSCU values are indicated for orthologs and whole genomes.

Aspergillus clavatus

```
Codon usage of orthologous genes:
Phe UUU   22505 0.6 Ser UCU   26647   1 Tyr UAU   21298 0.8 Cys UGU    7549 0.7
    UUC   48621 1.4     UCC   37698 1.3     UAC   30396 1.2     UGC   12734 1.3
Leu UUA    7637 0.3     UCA   19334 0.7 TER UAA     898 0.8 TER UGA    1402 1.3
    UUG   29216   1     UCG   34685 1.2     UAG     890 0.8 Trp UGG   25218   1

    CUU   29007   1 Pro CCU   30903   1 His CAU   22450   1 Arg CGU   19821 0.9
    CUC   47305 1.6     CCC   37881 1.2     CAC   24208   1     CGC   35499 1.7
    CUA   12507 0.4     CCA   24562 0.8 Gln CAA   29260 0.7     CGA   20464   1
    CUG   51458 1.7     CCG   30430   1     CAG   53050 1.3     CGG   25232 1.2

Ile AUU   28903 0.9 Thr ACU   23138 0.8 Asn AAU   28864 0.8 Ser AGU   18253 0.7
    AUC   55829 1.8     ACC   40939 1.4     AAC   41914 1.2     AGC   32540 1.2
    AUA    8150 0.3     ACA   23176 0.8 Lys AAA   29982 0.6 Arg AGA   14917 0.7
Met AUG   41301   1     ACG   26201 0.9     AAG   66456 1.4     AGG   11171 0.5

Val GUU   26276 0.9 Ala GCU   41017   1 Asp GAU   57632   1 Gly GGU   32067   1
    GUC   47335 1.6     GCC   57914 1.4     GAC   57137   1     GGC   48836 1.5
    GUA    9928 0.3     GCA   31201 0.7 Glu GAA   50612 0.8     GGA   27264 0.8
    GUG   35775 1.2     GCG   41547   1     GAG   78708 1.2     GGG   21277 0.7

Total: 1977025 codons
```

```
Codon usage of all genes in the genome:
Phe UUU   53906 0.7 Ser UCU   57647 0.9 Tyr UAU   51367 0.8 Cys UGU   20144 0.8
    UUC  112217 1.4     UCC   84865 1.4     UAC   72480 1.2     UGC   33447 1.3
Leu UUA   17437 0.3     UCA   42787 0.7 TER UAA    2361 0.8 TER UGA    4003 1.3
    UUG   65756   1     UCG   75862 1.2     UAG    2750 0.9 Trp UGG   63617   1

    CUU   64291 0.9 Pro CCU   66220   1 His CAU   52444   1 Arg CGU   42452 0.9
    CUC  111243 1.6     CCC   85088 1.2     CAC   56489   1     CGC   79656 1.7
    CUA   29775 0.4     CCA   55102 0.8 Gln CAA   64667 0.7     CGA   44735   1
    CUG  122419 1.8     CCG   68958   1     CAG  119816 1.3     CGG   56847 1.2

Ile AUU   67041 0.9 Thr ACU   52384 0.8 Asn AAU   64600 0.8 Ser AGU   41196 0.7
    AUC  131281 1.8     ACC   96332 1.5     AAC   94093 1.2     AGC   73384 1.2
    AUA   20152 0.3     ACA   52865 0.8 Lys AAA   65200 0.6 Arg AGA   32950 0.7
Met AUG   96439   1     ACG   61925 0.9     AAG  142909 1.4     AGG   26101 0.6

Val GUU   59079 0.9 Ala GCU   90712 0.9 Asp GAU  125920   1 Gly GGU   71301 0.9
    GUC  110410 1.6     GCC  134594 1.4     GAC  128171   1     GGC  113512 1.5
    GUA   23373 0.3     GCA   72115 0.7 Glu GAA  107616 0.8     GGA   63465 0.8
    GUG   85832 1.2     GCG   96823   1     GAG  173382 1.2     GGG   54476 0.7

Total: 4510481 codons
```

Aspergillus flavus

Codon usage of orthologous genes:

Phe	UUU	24635	0.7	Ser	UCU	31952	1.2	Tyr	UAU	23758	0.9	Cys	UGU	8735	0.9
	UUC	45029	1.3		UCC	33729	1.2		UAC	28321	1.1		UGC	11052	1.1
Leu	UUA	12279	0.4		UCA	23342	0.8	TER	UAA	1122	1.1	TER	UGA	1224	1.2
	UUG	32193	1.1		UCG	28071	1		UAG	826	0.8	Trp	UGG	24753	1
	CUU	34313	1.2	Pro	CCU	34630	1.2	His	CAU	23515	1.1	Arg	CGU	23461	1.2
	CUC	38546	1.4		CCC	31295	1		CAC	21347	1		CGC	28620	1.4
	CUA	17480	0.6		CCA	28444	0.9	Gln	CAA	34528	0.9		CGA	20277	1
	CUG	36587	1.3		CCG	26195	0.9		CAG	46340	1.2		CGG	21454	1.1
Ile	AUU	33187	1.1	Thr	ACU	28123	1	Asn	AAU	31915	0.9	Ser	AGU	20961	0.8
	AUC	46154	1.5		ACC	35938	1.3		AAC	40911	1.1		AGC	28845	1
	AUA	11823	0.4		ACA	26408	0.9	Lys	AAA	34290	0.7	Arg	AGA	15006	0.7
Met	AUG	40301	1		ACG	22709	0.8		AAG	63662	1.3		AGG	13061	0.6
Val	GUU	32645	1.1	Ala	GCU	46277	1.2	Asp	GAU	60349	1.1	Gly	GGU	36774	1.2
	GUC	39430	1.4		GCC	48287	1.2		GAC	52953	0.9		GGC	40613	1.3
	GUA	14105	0.5		GCA	35258	0.9	Glu	GAA	58400	0.9		GGA	30045	1
	GUG	30900	1.1		GCG	31080	0.8		GAG	69849	1.1		GGG	19571	0.6

Total: 1937883 codons

Codon usage of all genes in the genome:

Phe	UUU	83254	0.8	Ser	UCU	86984	1.1	Tyr	UAU	80040	0.9	Cys	UGU	33649	0.9
	UUC	138357	1.3		UCC	96695	1.2		UAC	92045	1.1		UGC	40740	1.1
Leu	UUA	38831	0.4		UCA	67299	0.9	TER	UAA	3926	0.9	TER	UGA	4922	1.2
	UUG	95823	1.1		UCG	77557	1		UAG	3639	0.9	Trp	UGG	88031	1
	CUU	101643	1.2	Pro	CCU	92290	1.1	His	CAU	74387	1.1	Arg	CGU	61576	1.1
	CUC	121454	1.4		CCC	88472	1.1		CAC	66280	0.9		CGC	78293	1.4
	CUA	57096	0.7		CCA	83461	1	Gln	CAA	99880	0.9		CGA	58559	1
	CUG	114662	1.3		CCG	74048	0.9		CAG	132398	1.1		CGG	59971	1.1
Ile	AUU	106623	1.1	Thr	ACU	82990	1	Asn	AAU	97924	0.9	Ser	AGU	62310	0.8
	AUC	148606	1.5		ACC	110541	1.3		AAC	119495	1.1		AGC	83927	1.1
	AUA	42570	0.4		ACA	81636	0.9	Lys	AAA	98503	0.7	Arg	AGA	44203	0.8
Met	AUG	126037	1		ACG	70486	0.8		AAG	171243	1.3		AGG	39388	0.7
Val	GUU	97600	1.1	Ala	GCU	130833	1.1	Asp	GAU	172282	1.1	Gly	GGU	108498	1.1
	GUC	121889	1.3		GCC	145819	1.2		GAC	153818	0.9		GGC	124618	1.3
	GUA	44863	0.5		GCA	107965	0.9	Glu	GAA	157737	0.9		GGA	95451	1
	GUG	99052	1.1		GCG	94011	0.8		GAG	196797	1.1		GGG	68464	0.7

Total: 5802441 codons

Aspergillus fumigatus

```
Codon usage of orthologous genes:
Phe UUU   25503 0.7 Ser UCU   29851 1.1 Tyr UAU   21370 0.8 Cys UGU    8144 0.8
    UUC   45228 1.3     UCC   34728 1.2     UAC   30554 1.2     UGC   12951 1.2
Leu UUA    9125 0.3     UCA   22986 0.8 TER UAA     870 0.8 TER UGA    1389 1.3
    UUG   31521 1.1     UCG   31707 1.1     UAG     932 0.9 Trp UGG   25179   1

    CUU   32321 1.1 Pro CCU   34332 1.1 His CAU   23246   1 Arg CGU   21254   1
    CUC   43193 1.5     CCC   32699 1.1     CAC   23061   1     CGC   31723 1.5
    CUA   14311 0.5     CCA   26106 0.9 Gln CAA   31074 0.8     CGA   21358   1
    CUG   46100 1.6     CCG   28602 0.9     CAG   50306 1.2     CGG   23521 1.1

Ile AUU   32925 1.1 Thr ACU   26272 0.9 Asn AAU   30790 0.9 Ser AGU   20030 0.7
    AUC   49561 1.6     ACC   36445 1.3     AAC   40459 1.1     AGC   31256 1.1
    AUA   10660 0.3     ACA   25858 0.9 Lys AAA   32763 0.7 Arg AGA   16302 0.8
Met AUG   40674   1     ACG   24962 0.9     AAG   64593 1.3     AGG   13452 0.6

Val GUU   29197   1 Ala GCU   44671 1.1 Asp GAU   55578   1 Gly GGU   33030   1
    GUC   45207 1.5     GCC   51773 1.2     GAC   57226   1     GGC   44534 1.4
    GUA   12088 0.4     GCA   34237 0.8 Glu GAA   53948 0.8     GGA   28762 0.9
    GUG   32410 1.1     GCG   37068 0.9     GAG   74519 1.2     GGG   20888 0.7

Total: 1967383 codons
```

```
Codon usage of all genes in the genome:
Phe UUU   65058 0.7 Ser UCU   68100   1 Tyr UAU   55369 0.8 Cys UGU   24135 0.8
    UUC  112580 1.3     UCC   84332 1.3     UAC   78098 1.2     UGC   38011 1.2
Leu UUA   22403 0.3     UCA   53595 0.8 TER UAA    2508 0.8 TER UGA    4367 1.3
    UUG   74222   1     UCG   73980 1.1     UAG    3012 0.9 Trp UGG   69741   1

    CUU   77254 1.1 Pro CCU   77998 1.1 His CAU   58671   1 Arg CGU   48091 0.9
    CUC  110312 1.5     CCC   79565 1.1     CAC   58147   1     CGC   75638 1.5
    CUA   36765 0.5     CCA   63095 0.9 Gln CAA   73499 0.8     CGA   50322   1
    CUG  118130 1.6     CCG   68973   1     CAG  121542 1.3     CGG   57680 1.1

Ile AUU   80558   1 Thr ACU   63610 0.9 Asn AAU   74086 0.9 Ser AGU   47682 0.7
    AUC  127413 1.6     ACC   93099 1.3     AAC   97690 1.1     AGC   75653 1.1
    AUA   28285 0.4     ACA   62729 0.9 Lys AAA   75508 0.7 Arg AGA   39288 0.8
Met AUG  101214   1     ACG   62797 0.9     AAG  145877 1.3     AGG   34545 0.7

Val GUU   69435 0.9 Ala GCU  104928   1 Asp GAU  129012   1 Gly GGU   77642   1
    GUC  113360 1.5     GCC  129829 1.3     GAC  136403   1     GGC  112882 1.4
    GUA   30422 0.4     GCA   84720 0.8 Glu GAA  120045 0.8     GGA   72086 0.9
    GUG   83824 1.1     GCG   93127 0.9     GAG  173816 1.2     GGG   58311 0.7

Total: 4805069 codons
```

Neosartorya fischeri

```
Codon usage of orthologous genes:
Phe UUU   24104 0.7 Ser UCU   29166   1 Tyr UAU   20794 0.8 Cys UGU    7575 0.8
    UUC   46418 1.3     UCC   35147 1.2     UAC   31221 1.2     UGC   12675 1.3
Leu UUA    8427 0.3     UCA   21696 0.8 TER UAA     850 0.8 TER UGA    1403 1.3
    UUG   31124 1.1     UCG   32512 1.2     UAG     938 0.9 Trp UGG   25100   1

    CUU   31334 1.1 Pro CCU   34242 1.1 His CAU   22610   1 Arg CGU   20974   1
    CUC   43756 1.5     CCC   33472 1.1     CAC   23250   1     CGC   32253 1.5
    CUA   13707 0.5     CCA   24946 0.8 Gln CAA   30672 0.8     CGA   21061   1
    CUG   47067 1.6     CCG   29486   1     CAG   50953 1.3     CGG   23465 1.1

Ile AUU   31863   1 Thr ACU   25554 0.9 Asn AAU   30133 0.9 Ser AGU   19449 0.7
    AUC   50154 1.6     ACC   36859 1.3     AAC   41152 1.2     AGC   31631 1.1
    AUA    9985 0.3     ACA   24955 0.9 Lys AAA   31927 0.7 Arg AGA   16123 0.8
Met AUG   40730   1     ACG   25570 0.9     AAG   66091 1.4     AGG   13162 0.6

Val GUU   28548   1 Ala GCU   44345 1.1 Asp GAU   55486   1 Gly GGU   33304   1
    GUC   46545 1.6     GCC   53137 1.3     GAC   58632   1     GGC   45144 1.4
    GUA   11556 0.4     GCA   33976 0.8 Glu GAA   53958 0.8     GGA   29074 0.9
    GUG   32442 1.1     GCG   38176 0.9     GAG   76067 1.2     GGG   20437 0.6

Total: 1968563 codons
```

```
Codon usage of all genes in the genome:
Phe UUU   67577 0.7 Ser UCU   69536   1 Tyr UAU   58929 0.8 Cys UGU   23989 0.8
    UUC  121345 1.3     UCC   89011 1.3     UAC   85168 1.2     UGC   39461 1.2
Leu UUA   22704 0.3     UCA   53778 0.8 TER UAA    2642 0.8 TER UGA    4507 1.3
    UUG   77978   1     UCG   78552 1.1     UAG    3320   1 Trp UGG   74622   1

    CUU   79142   1 Pro CCU   81011 1.1 His CAU   60713   1 Arg CGU   49043 0.9
    CUC  118007 1.5     CCC   84007 1.1     CAC   61274   1     CGC   79669 1.5
    CUA   38290 0.5     CCA   64371 0.9 Gln CAA   76297 0.7     CGA   51732   1
    CUG  126155 1.6     CCG   73856   1     CAG  130546 1.3     CGG   60300 1.1

Ile AUU   83595   1 Thr ACU   65616 0.9 Asn AAU   77685 0.9 Ser AGU   49245 0.7
    AUC  136084 1.6     ACC   99424 1.3     AAC  104719 1.2     AGC   80442 1.2
    AUA   28745 0.4     ACA   65227 0.9 Lys AAA   78617 0.7 Arg AGA   40493 0.8
Met AUG  108342   1     ACG   67491 0.9     AAG  156646 1.3     AGG   35641 0.7

Val GUU   72807 0.9 Ala GCU  110629   1 Asp GAU  137633   1 Gly GGU   83156   1
    GUC  122842 1.6     GCC  139151 1.3     GAC  147535   1     GGC  120176 1.4
    GUA   31059 0.4     GCA   90159 0.8 Glu GAA  127421 0.8     GGA   77330 0.9
    GUG   89537 1.1     GCG  100688 0.9     GAG  187838 1.2     GGG   61699 0.7

Total: 5085204 codons
```

Aspergillus niger

Codon usage of orthologous genes:

```
Phe UUU   22664 0.6 Ser UCU   29049   1 Tyr UAU   21040 0.8 Cys UGU    8378 0.8
    UUC   49753 1.4     UCC   40691 1.4     UAC   33220 1.2     UGC   13504 1.2
Leu UUA    7832 0.3     UCA   19319 0.7 TER UAA    1257 0.9 TER UGA    1923 1.3
    UUG   32573 1.1     UCG   30968 1.1     UAG    1195 0.8 Trp UGG   26069   1

    CUU   30641   1 Pro CCU   32618   1 His CAU   21915 0.9 Arg CGU   22381 1.1
    CUC   45980 1.5     CCC   38825 1.2     CAC   25381 1.1     CGC   36508 1.7
    CUA   14823 0.5     CCA   23490 0.7 Gln CAA   30674 0.8     CGA   18325 0.9
    CUG   47569 1.6     CCG   31286   1     CAG   51128 1.3     CGG   24158 1.1

Ile AUU   31018   1 Thr ACU   25507 0.9 Asn AAU   27189 0.8 Ser AGU   20115 0.7
    AUC   52335 1.7     ACC   42810 1.5     AAC   44532 1.2     AGC   31752 1.1
    AUA   10091 0.3     ACA   22149 0.8 Lys AAA   26974 0.6 Arg AGA   13503 0.6
Met AUG   41919   1     ACG   26843 0.9     AAG   69662 1.4     AGG   13095 0.6

Val GUU   28701   1 Ala GCU   43779   1 Asp GAU   56538   1 Gly GGU   35401 1.1
    GUC   44966 1.5     GCC   56828 1.4     GAC   58359   1     GGC   45664 1.4
    GUA   11741 0.4     GCA   32214 0.8 Glu GAA   52589 0.8     GGA   29017 0.9
    GUG   35879 1.2     GCG   35889 0.9     GAG   77259 1.2     GGG   20507 0.6
```

Total: 1999962 codons

Codon usage of all genes in the genome:

```
Phe UUU   74306 0.7 Ser UCU   81764   1 Tyr UAU   68586 0.9 Cys UGU   37726 0.8
    UUC  131937 1.3     UCC  108276 1.3     UAC   90674 1.1     UGC   53541 1.2
Leu UUA   29544 0.3     UCA   65229 0.8 TER UAA    7903 0.7 TER UGA   20368 1.7
    UUG   93213 1.1     UCG   83166   1     UAG    7503 0.6 Trp UGG   91729   1

    CUU   89422   1 Pro CCU   87258   1 His CAU   72492   1 Arg CGU   57803 0.9
    CUC  128157 1.5     CCC  101751 1.2     CAC   72859   1     CGC   93135 1.5
    CUA   51873 0.6     CCA   79122 0.9 Gln CAA   92445 0.8     CGA   62247   1
    CUG  128424 1.5     CCG   85853   1     CAG  128780 1.2     CGG   68579 1.1

Ile AUU   90383   1 Thr ACU   72911 0.9 Asn AAU   79159 0.8 Ser AGU   58315 0.7
    AUC  141080 1.6     ACC  115924 1.4     AAC  109924 1.2     AGC   86918 1.1
    AUA   38625 0.4     ACA   71554 0.9 Lys AAA   75697 0.7 Arg AGA   48565 0.8
Met AUG  116988   1     ACG   74745 0.9     AAG  153913 1.3     AGG   47223 0.8

Val GUU   79639 0.9 Ala GCU  114884   1 Asp GAU  146695   1 Gly GGU   91467   1
    GUC  118057 1.4     GCC  146050 1.3     GAC  143477   1     GGC  122301 1.3
    GUA   39363 0.5     GCA   97773 0.9 Glu GAA  134156 0.9     GGA   91186   1
    GUG  102817 1.2     GCG   97190 0.9     GAG  180057 1.2     GGG   71202 0.8
```

Total: 5603903 codons

Aspergillus nidulans

Codon usage of orthologous genes:

```
Phe UUU   26307 0.8 Ser UCU   31904 1.1 Tyr UAU   21578 0.8 Cys UGU    7430 0.8
    UUC   43831 1.3     UCC   32586 1.2     UAC   31054 1.2     UGC   12334 1.3
Leu UUA   11203 0.4     UCA   25542 0.9 TER UAA     945 0.9 TER UGA    1232 1.2
    UUG   26971 0.9     UCG   28408   1     UAG    1012   1 Trp UGG   24826   1

    CUU   36452 1.3 Pro CCU   33905 1.1 His CAU   21974   1 Arg CGU   20415   1
    CUC   44101 1.5     CCC   29567   1     CAC   22885   1     CGC   32961 1.6
    CUA   17875 0.6     CCA   27416 0.9 Gln CAA   32978 0.8     CGA   21251   1
    CUG   37415 1.3     CCG   30337   1     CAG   47619 1.2     CGG   21790   1

Ile AUU   34906 1.1 Thr ACU   27026   1 Asn AAU   30732 0.8 Ser AGU   19313 0.7
    AUC   45272 1.5     ACC   33312 1.2     AAC   42096 1.2     AGC   31316 1.1
    AUA   13127 0.4     ACA   28166   1 Lys AAA   35432 0.7 Arg AGA   14058 0.7
Met AUG   39022   1     ACG   25687 0.9     AAG   61164 1.3     AGG   14741 0.7

Val GUU   33901 1.2 Ala GCU   45423 1.1 Asp GAU   55701   1 Gly GGU   32465   1
    GUC   41622 1.4     GCC   48012 1.2     GAC   57255   1     GGC   43917 1.4
    GUA   12920 0.4     GCA   35081 0.9 Glu GAA   55846 0.9     GGA   28390 0.9
    GUG   27865   1     GCG   36660 0.9     GAG   72068 1.1     GGG   22214 0.7
```

Total: 1950814 codons

Codon usage of all genes in the genome:

```
Phe UUU   73372 0.8 Ser UCU   78224 1.1 Tyr UAU   62148 0.8 Cys UGU   24332 0.8
    UUC  117834 1.2     UCC   83295 1.2     UAC   87199 1.2     UGC   40539 1.3
Leu UUA   29141 0.4     UCA   64002 0.9 TER UAA    3248 0.9 TER UGA    4367 1.2
    UUG   69960 0.9     UCG   72069   1     UAG    3823   1 Trp UGG   75567   1

    CUU   92982 1.2 Pro CCU   82064 1.1 His CAU   60230   1 Arg CGU   48763 0.9
    CUC  122200 1.6     CCC   76403   1     CAC   61947   1     CGC   83276 1.6
    CUA   49144 0.6     CCA   72483 0.9 Gln CAA   81326 0.8     CGA   52492   1
    CUG  108805 1.4     CCG   77689   1     CAG  126077 1.2     CGG   56827 1.1

Ile AUU   91630 1.1 Thr ACU   69798 0.9 Asn AAU   80416 0.9 Ser AGU   50737 0.7
    AUC  128099 1.5     ACC   91478 1.2     AAC  108875 1.2     AGC   83271 1.2
    AUA   38416 0.5     ACA   75154   1 Lys AAA   86263 0.7 Arg AGA   39447 0.7
Met AUG  106205   1     ACG   70672 0.9     AAG  148806 1.3     AGG   41730 0.8

Val GUU   86292 1.1 Ala GCU  113983   1 Asp GAU  138140   1 Gly GGU   82029 0.9
    GUC  114661 1.5     GCC  130207 1.2     GAC  148018   1     GGC  121893 1.4
    GUA   35832 0.5     GCA   96885 0.9 Glu GAA  133416 0.8     GGA   77526 0.9
    GUG   78328   1     GCG   99016 0.9     GAG  184868 1.2     GGG   67590 0.8
```

Total: 5161509 codons

Aspergillus oryzae

Codon usage of orthologous genes:

```
Phe UUU   24001 0.7 Ser UCU   30471 1.1 Tyr UAU   23016 0.9 Cys UGU    8696 0.9
    UUC   43766 1.3     UCC   32298 1.2     UAC   27457 1.1     UGC   10941 1.1
Leu UUA   11987 0.4     UCA   22320 0.8 TER UAA    1141 1.1 TER UGA    1187 1.1
    UUG   31391 1.1     UCG   26787   1     UAG     863 0.8 Trp UGG   24168   1

    CUU   33364 1.2 Pro CCU   32928 1.2 His CAU   22703 1.1 Arg CGU   22580 1.2
    CUC   37555 1.4     CCC   29935   1     CAC   20616   1     CGC   27521 1.4
    CUA   17037 0.6     CCA   27017 0.9 Gln CAA   33007 0.9     CGA   19497   1
    CUG   35739 1.3     CCG   24977 0.9     CAG   44522 1.2     CGG   20727 1.1

Ile AUU   32282 1.1 Thr ACU   27090   1 Asn AAU   30576 0.9 Ser AGU   20221 0.8
    AUC   44895 1.5     ACC   34614 1.3     AAC   39601 1.1     AGC   27649   1
    AUA   11613 0.4     ACA   25431 0.9 Lys AAA   33133 0.7 Arg AGA   14428 0.7
Met AUG   39109   1     ACG   21928 0.8     AAG   61597 1.3     AGG   12768 0.7

Val GUU   31856 1.1 Ala GCU   44585 1.2 Asp GAU   57967 1.1 Gly GGU   35644 1.2
    GUC   38327 1.3     GCC   46510 1.2     GAC   50829 0.9     GGC   39247 1.3
    GUA   13899 0.5     GCA   33988 0.9 Glu GAA   56287 0.9     GGA   28996 0.9
    GUG   30045 1.1     GCG   29814 0.8     GAG   67610 1.1     GGG   18980 0.6
```

Total: 1871734 codons

Codon usage of all genes in the genome:

```
Phe UUU   79031 0.8 Ser UCU   81562 1.1 Tyr UAU   74889 0.9 Cys UGU   32421 0.9
    UUC  129459 1.2     UCC   90185 1.2     UAC   84744 1.1     UGC   38440 1.1
Leu UUA   36915 0.4     UCA   63457 0.9 TER UAA    3929   1 TER UGA    4593 1.1
    UUG   91166 1.1     UCG   71949   1     UAG    3541 0.9 Trp UGG   82965   1

    CUU   95940 1.2 Pro CCU   86021 1.1 His CAU   69438 1.1 Arg CGU   57893 1.1
    CUC  113335 1.4     CCC   82172   1     CAC   61479 0.9     CGC   72840 1.4
    CUA   54018 0.7     CCA   77685   1 Gln CAA   92400 0.9     CGA   54891   1
    CUG  107879 1.3     CCG   68748 0.9     CAG  123503 1.1     CGG   56567 1.1

Ile AUU  100198 1.1 Thr ACU   77689   1 Asn AAU   91125 0.9 Ser AGU   58579 0.8
    AUC  138491 1.5     ACC  102361 1.3     AAC  110708 1.1     AGC   77619 1.1
    AUA   40956 0.4     ACA   76309   1 Lys AAA   91553 0.7 Arg AGA   41986 0.8
Met AUG  118373   1     ACG   65803 0.8     AAG  158973 1.3     AGG   37831 0.7

Val GUU   91762 1.1 Ala GCU  122537 1.1 Asp GAU  159928 1.1 Gly GGU  102099 1.1
    GUC  113390 1.3     GCC  135033 1.2     GAC  141610 0.9     GGC  116124 1.3
    GUA   43228 0.5     GCA  101108 0.9 Glu GAA  146068 0.9     GGA   89324   1
    GUG   93403 1.1     GCG   87745 0.8     GAG  182276 1.1     GGG   64463 0.7
```

Total: 5422707 codons

Aspergillus terreus

Codon usage of orthologous genes:

```
Phe UUU   20375  0.6 Ser UCU   23892  0.9 Tyr UAU   18004  0.7 Cys UGU    7580  0.8
    UUC   49000  1.4     UCC   42222  1.5     UAC   33708  1.3     UGC   12625  1.3
Leu UUA    5437  0.2     UCA   13710  0.5 TER UAA     796  0.8 TER UGA    1368  1.3
    UUG   26141  0.9     UCG   36914  1.3     UAG    1027    1 Trp UGG   25163    1

    CUU   23344  0.8 Pro CCU   26878  0.9 His CAU   20290  0.9 Arg CGU   20072    1
    CUC   50527  1.8     CCC   42239  1.3     CAC   26139  1.1     CGC   42004    2
    CUA   10208  0.4     CCA   19190  0.6 Gln CAA   25518  0.6     CGA   18705  0.9
    CUG   56986    2     CCG   37950  1.2     CAG   54250  1.4     CGG   27636  1.3

Ile AUU   25108  0.9 Thr ACU   19424  0.7 Asn AAU   22582  0.7 Ser AGU   16700  0.6
    AUC   56683  1.9     ACC   45738  1.6     AAC   46486  1.4     AGC   31838  1.2
    AUA    5981  0.2     ACA   16015  0.6 Lys AAA   27193  0.6 Arg AGA   10401  0.5
Met AUG   41502    1     ACG   31861  1.1     AAG   67058  1.4     AGG    8110  0.4

Val GUU   20840  0.7 Ala GCU   33582  0.8 Asp GAU   49398  0.9 Gly GGU   29548  0.9
    GUC   50308  1.7     GCC   66100  1.6     GAC   65903  1.1     GGC   51924  1.6
    GUA    7525  0.3     GCA   23924  0.6 Glu GAA   47313  0.8     GGA   25522  0.8
    GUG   40247  1.4     GCG   46263  1.1     GAG   78708  1.3     GGG   21398  0.7
```

Total: 1951081 codons

Codon usage of all genes in the genome:

```
Phe UUU   53176  0.6 Ser UCU   52720  0.9 Tyr UAU   47288  0.7 Cys UGU   22185  0.8
    UUC  112389  1.4     UCC   89725  1.5     UAC   80104  1.3     UGC   34907  1.2
Leu UUA   14636  0.2     UCA   33294  0.6 TER UAA    2150  0.7 TER UGA    3819  1.3
    UUG   60106  0.9     UCG   76845  1.3     UAG    2928    1 Trp UGG   65967    1

    CUU   56284  0.8 Pro CCU   57987  0.8 His CAU   49704  0.9 Arg CGU   42349  0.9
    CUC  116174  1.7     CCC   88039  1.3     CAC   60209  1.1     CGC   87895  1.9
    CUA   27945  0.4     CCA   47007  0.7 Gln CAA   57360  0.7     CGA   41799  0.9
    CUG  129053  1.9     CCG   81523  1.2     CAG  119161  1.4     CGG   59424  1.3

Ile AUU   61427  0.9 Thr ACU   45799  0.7 Asn AAU   53825  0.7 Ser AGU   37642  0.6
    AUC  132194  1.9     ACC  101658  1.6     AAC  101271  1.3     AGC   71193  1.2
    AUA   19315  0.3     ACA   41888  0.6 Lys AAA   60181  0.6 Arg AGA   26131  0.6
Met AUG   96901    1     ACG   71901  1.1     AAG  136233  1.4     AGG   21742  0.5

Val GUU   50917  0.7 Ala GCU   76803  0.8 Asp GAU  109754  0.9 Gly GGU   65255  0.9
    GUC  115699  1.7     GCC  145746  1.5     GAC  145043  1.1     GGC  119886  1.6
    GUA   20438  0.3     GCA   62166  0.6 Glu GAA  100756  0.8     GGA   62568  0.8
    GUG   93234  1.3     GCG  103966  1.1     GAG  166991  1.3     GGG   56192  0.7
```

Total: 4448867 codons

6 Capítulo III: Análisis del UCS en *Virus del Nilo Occidental* y *Virus Influenza A*

6.1

Pandemic influenza A virus codon usage revisited: biases, adaptation and implications for vaccine strain development

Goñi N[#], Iriarte A[#], Comas V, Soñora M, Moreno P, Moratorio G, Musto H, Cristina J.

Virology Journal 2012, 9:263

[#]Equal contributors

Resumen:

El Virus de Influenza A (IAV) es un miembro de la familia Orthomyxoviridae y contiene ocho segmentos de un genoma de ARN monocatenario con polaridad negativa. Comprender el alcance y las causas del sesgo en el uso de codones sinónimos es esencial para la comprensión cabal de la evolución viral. En este estudio se realizó un análisis exhaustivo del uso de codones de 310 cepas de IAV de la pandemia ocurrida en 2009. Se encontró un uso sesgado en codones codificantes para Ala, Arg, Pro, Thr y Ser. Los sesgos observados parecen estar fuertemente influenciado por los sesgos subyacentes en la composición de bases. Los análisis multivariados muestran que existe cierto grado de variabilidad entre las cepas que pudría asociarse a un proceso evolutivo de divergencia. Los resultados muestran una asociación general entre el sesgo de uso de codones y la frecuencia de dinucleótidos al mismo tiempo que muestran una pobre adaptación del virus al conjunto de tRNAs del hospedero. En suma, los resultados sugieren que la presión mutacional es una fuerza principal para explicar el uso de codones en las cepas analizadas H1N1. Muy probablemente existe un proceso dinámico que se refleja en la variación observada entre las cepas incluidas en estos análisis. Algunos resultados sugieren que puede existir un equilibrio del sesgo mutacional y la selección natural, que permitiría al virus explorar y re-adaptar su uso de codones a distintos entornos. La recodificación de IAV teniendo en cuenta estos resultados puede proporcionar pistas importantes para el desarrollo de vacunas nuevas y apropiadas.

6.2

A detailed comparative analysis on the overall codon usage patterns in West Nile virus

Moratorio G[#], Iriarte A[#], Moreno P, Musto H, Cristina J.

Infection, Genetics and Evolution 2013, Jan 16.

[#]Equal contributors

Resumen:

El Virus del Nilo Occidental (WNV) es un miembro de la familia Flaviviridae y su genoma consta de un RNA de cadena simple de 11kb con sentido positivo. El WNV se mantiene en un ciclo enzoótico entre mosquitos y aves, pero también puede infectar y causar enfermedad en caballos y seres humanos, que sirven como hospederos finales incidentales. Comprender el alcance y las causas del sesgo en el uso de codones sinónimos es esencial para la comprensión de la evolución viral. En este estudio se realizó un análisis exhaustivo de 449 cepas de WNV. El número efectivo de los codones (ENC) indica que el uso de codones está poco sesgado. El análisis del uso relativo de codones (RSCU) sugiere que la variabilidad observada entre las cepas de WNV (aisladas de aves, equinos, seres humanos y mosquitos) es menor y está influenciada principalmente por las frecuencias relativas de dinucleótidos. Tomando en conjunto, los resultados de este trabajo sugieren que el uso de codones en WNV es el resultado de los sesgos composicionales y la necesidad de escapar de las respuestas de las células antivirales, constituyendo un proceso dinámico de mutación y selección para volver a adaptarlo a diferentes entornos.

Goñi et al. Virology Journal 2012, 9:263
http://www.virologyj.com/content/9/1/263

VIROLOGY JOURNAL

RESEARCH Open Access

Pandemic influenza A virus codon usage revisited: biases, adaptation and implications for vaccine strain development

Natalia Goñi[1†], Andrés Iriarte[2,3†], Victoria Comas[1], Martín Soñora[1], Pilar Moreno[1,4], Gonzalo Moratorio[1,5], Héctor Musto[2] and Juan Cristina[1*]

Abstract

Background: Influenza A virus (IAV) is a member of the family *Orthomyxoviridae* and contains eight segments of a single-stranded RNA genome with negative polarity. The first influenza pandemic of this century was declared in April of 2009, with the emergence of a novel H1N1 IAV strain (H1N1pdm) in Mexico and USA. Understanding the extent and causes of biases in codon usage is essential to the understanding of viral evolution. A comprehensive study to investigate the effect of selection pressure imposed by the human host on the codon usage of an emerging, pandemic IAV strain and the trends in viral codon usage involved over the pandemic time period is much needed.

Results: We performed a comprehensive codon usage analysis of 310 IAV strains from the pandemic of 2009. Highly biased codon usage for Ala, Arg, Pro, Thr and Ser were found. Codon usage is strongly influenced by underlying biases in base composition. When correspondence analysis (COA) on relative synonymous codon usage (RSCU) is applied, the distribution of IAV ORFs in the plane defined by the first two major dimensional factors showed that different strains are located at different places, suggesting that IAV codon usage also reflects an evolutionary process.

Conclusions: A general association between codon usage bias, base composition and poor adaptation of the virus to the respective host tRNA pool, suggests that mutational pressure is the main force shaping H1N1 pdm IAV codon usage. A dynamic process is observed in the variation of codon usage of the strains enrolled in these studies. These results suggest a balance of mutational bias and natural selection, which allow the virus to explore and re-adapt its codon usage to different environments. Recoding of IAV taking into account codon bias, base composition and adaptation to host tRNA may provide important clues to develop new and appropriate vaccines.

Keywords: Influenza A virus, Codon usage, Evolution

Background

Influenza A virus (IAV) is a member of the family *Orthomyxoviridae* and contains eight segments of a single-stranded RNA genome with negative polarity [1]. IAV is one of the most important infectious diseases in humans [2]. Unlike most pathogens where exposure leads to lasting immunity in the host, IAV presents a moving antigenic target [3], evading specific immunity triggered by previous infections. This process, called antigenic drift, is the result of the selective fixation of mutations in the gene encoding the hemagglutinin (HA) protein and to a lesser extent in the neuraminidase (NA) protein [4]. Variants that best escape the host immune response are thought to have a significant reproductive advantage [5]. Another process, called reassortment, is also considered a major force in the evolution of IAV [4]. It occurs when the virus acquires an HA and/or NA of a different IAV subtype (via reassortation) of one or more gene segments. This process has

* Correspondence: cristina@cin.edu.uy
†Equal contributors
[1]Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares, Facultad de Ciencias, Universidad de la República, Iguá 4225, Montevideo 11400, Uruguay
Full list of author information is available at the end of the article

been in the basis of the devastating influenza pandemics that occurred several times in the last century [6].

The first influenza pandemic of this century was declared in April of 2009, with the emergence of a novel H1N1 IAV strain (H1N1pdm) in Mexico and USA [7,8]. By November of 2009, the virus was detected in about 207 countries, infecting more than 620,000 individuals worldwide and accounting for more than 7,800 deaths [7]. This strain was a multiple reassortant with genes derived from viruses that originally circulated in the swine, avian and human populations [9].

It has been observed that IAV is subjected to host immune selection pressure and undergoes rapid evolution, especially when the virus crosses the host species barrier [10]. The replication cycle of IAV depends on host machinery and the virus utilizes host cellular components for its protein synthesis. Therefore, the interplay of codon usage of virus and host could affect viral replication. For these reasons, a detailed understanding of IAV evolution and host adaptation is crucial.

Due to the degeneracy of the genetic code, most amino acids are coded by more than one codon. Synonymous triplets are not used randomly. In several organisms, natural selection and mutational input seem to bias codon use toward a certain subset of codons [11]. Two major models have been proposed to explain codon usage: the translational selection and the mutational models [12]. Codon usage bias related to translation efficiency (at two different levels: speed and accuracy) seems to be linked to local cognate isoacceptors tRNAs abundances, which in turn determine the major codon preferences [13]. On the other hand, discrepancies on codon usage could be due to genome compositional constraints and mutational biases [14]. Nevertheless, these two models cannot be considered as mutually exclusive.

Although previous studies have been performed on the general codon usage of IAV [2,12,15,16], a deep and comprehensive study to investigate the effect of selection pressure imposed by the human host on the codon usage of an emerging, pandemic IAV strain and the trends in viral codon usage involved over the pandemic time period is much needed.

In order to gain insight into these matters, we performed a comprehensive codon usage analysis of 310 H1N1pdm IAV strains, isolated from April to September of 2009, for which the complete genome sequences are available.

Results

In order to study the extent of codon usage bias in H1N1pdm IAV strains in relation to seasonal H1N1 and H3N2 as well as human and swine host cells, the relative synonymous codon usage (RSCU) [14] values for each codon were calculated for the 310 H1N1pdm strains enrolled in these studies and compared with seasonal IAV strains and host organisms. The results of these studies are shown in Table 1.

All codons containing the dinucleotide CpG were underrepresented in all IAV viruses. Important differences were found between human and swine hosts and IAV strains. Particularly, high biased frequencies (Δ RSCU ≥ 0.30) were found for Leu, Ile, Val, Ser, Pro, Thr, Ala, His, Gln, Glu, Arg and Gly. Interestingly, the huge majority of preferred codons in the viruses are A-ended. In the case of Arg, there is a strong bias towards an increase in AGA and AGG, while the CGN codons are depleted (see Table 1).

To observe if H1N1pdm IAV strain sequences display similar codon usage biases, the effective number of codons (ENC) [17] values were calculated for the 310 strains enrolled in this study (mean of 52.51 ± 0.05). ENC varies from 20 to 61, where the larger the extent of codon bias in a gene, the smaller the ENC value. Thus, a value of 52.5 strongly suggests that the overall codon usage among these strains is only slightly biased.

Since codon usage by its very nature is multivariate, it is necessary to analyze the data using multivariate statistical techniques, like correspondence analysis (COA) [18]. The correlation between the position on the first dimensional factor generated by this analysis on RSCU (20.7% of the total variability) for each strain and the respective G + C content at synonymous variable third position (GC_3s) values was significant ($r = -0.47$, $p < 0.0001$). Interestingly, this dimensional factor also significantly correlated with A content at synonymous variable third position (A_3s, $r = 0.68$, $p < 0.0001$) and G content at the same position (G_3s, $r = -0.71$, $p < 0.0001$) (Figure 1). This means that the major factor shaping codon usage among these strains is an opposite trend between purines at third codon positions. Furthermore, this result is mainly due to the frequencies of the codons CGA (Arg) on one side of the distribution and GCG (Ala) and CGG (Arg) at the other side (see Additional file 1: Table S1). In other words, the differential usage of three low frequent codons (RSCU ≤ 0.63) is among the major factor shaping codon usage among these strains.

It has been suggested that dinucleotide biases can affect codon bias [19]. To study the possible effect of dinucleotide composition on codon usage of the H1N1pdm IAV strains, the relative abundances of the 16 dinucleotides in the ORFs of the 310 strains enrolled in these studies were established. The results of these analyses are shown in Table 2.

As it can be seen in the table, the occurrences of dinucleotides are not randomly distributed and no dinucleotides were present at the expected frequencies (Table 2). The relative abundance of CpG showed a strong deviation from the “normal range” (mean ± S.D. = 0.319 ± 0.0020) and were markedly underrepresented. Interestingly, when the second

Table 1 Codon usage in 2009 H1N1 pdm Influenza A Virus, displayed as RSCU[a] values

AA	Cod	HC	Swine	H1N1pdm	H1N1[b]	H3N2	AA	Cod	HC	Swine	H1N1pdm	H1N1	H3N2
Phe	UUU	0.92	0.79	0.85	0.98	0.96	Ser	UCU	1.14	0.99	1.08	1.12	0.91
	UUC	1.08	1.21	1.15	1.02	1.04		UCC	1.32	1.50	0.74	0.87	0.97
Leu	UUA	0.48	0.32	0.62	0.91	0.62		**UCA**	**0.90**	**0.73**	**1.57**	**1.62**	**1.34**
	UUG	0.78	0.67	1.00	1.27	1.30		*UCG*	*0.30*	*0.39*	*0.31*	*0.14*	*0.21*
	CUU	0.78	0.65	1.16	0.97	1.24	Pro	CCU	1.16	1.05	1.00	1.04	1.29
	CUC	1.20	1.35	0.95	0.59	0.78		CCC	1.28	1.46	0.80	0.72	0.84
	CUA	**0.42**	**0.33**	**1.20**	**1.00**	**0.96**		**CCA**	**1.12**	**0.94**	**1.70**	**1.74**	**1.29**
	CUG	2.40	2.68	1.07	1.27	1.11		*CCG*	*0.44*	*0.56*	*0.50*	*0.49*	*0.58*
Ile	AUU	1.08	0.91	1.07	1.07	1.03	Thr	ACU	1.00	0.83	1.01	1.11	1.28
	AUC	1.41	1.67	0.77	0.78	0.89		ACC	1.44	1.68	0.79	0.96	0.72
	AUA	**0.51**	**0.42**	**1.16**	**1.16**	**1.08**		**ACA**	**1.12**	**0.92**	**1.88**	**1.74**	**1.67**
Met	AUG	1.00	1.00	1.00	1.00	1.00		*ACG*	*0.44*	*0.57*	*0.32*	*0.19*	*0.34*
Val	GUU	0.72	0.57	0.83	0.97	1.06	Ala	GCU	1.08	0.96	0.98	1.13	1.06
	GUC	0.96	1.07	0.77	0.74	0.69		GCC	1.60	1.80	0.87	0.87	0.93
	GUA	**0.48**	**0.34**	**1.12**	**1.07**	**1.02**		**GCA**	**0.92**	**0.74**	**1.87**	**1.74**	**1.73**
	GUG	1.84	2.03	1.28	1.22	1.23		*GCG*	*0.44*	*0.50*	*0.27*	*0.26*	*0.28*
Tyr	UAU	0.88	0.73	1.04	1.09	1.13	Cys	UGU	0.92	0.79	0.88	1.09	0.79
	UAC	1.12	1.27	0.96	0.91	0.87		UGC	1.08	1.21	1.12	0.91	1.21
TER	UAA	**	**	**	**	**	TER	UGA	**	**	**	**	**
	UAG	**	**	**	**	**	Trp	UGG	1.00	1.00	1.00	1.00	1.00
His	**CAU**	**0.84**	**0.70**	**1.23**	**1.05**	**1.21**	Arg	*CGU*	*0.48*	*0.44*	*0.11*	*0.24*	*0.10*
	CAC	1.16	1.30	0.77	0.95	0.79		*CGC*	*1.08*	*1.31*	*0.33*	*0.18*	*0.24*
Gln	**CAA**	**0.54**	**0.44**	**1.05**	**1.33**	**1.36**		*CGA*	*0.66*	*0.60*	*0.63*	*0.41*	*0.43*
	CAG	1.46	1.56	0.95	0.67	0.64		*CGG*	*1.20*	*1.29*	*0.43*	*0.28*	*0.57*
Asn	AAU	0.94	0.79	1.15	1.20	1.15	Ser	AGU	0.90	0.77	1.14	1.15	0.95
	GAC	1.08	1.21	0.95	0.80	0.85		AGC	1.44	1.62	1.16	1.11	1.38
Lys	AAA	0.86	0.76	1.10	1.27	1.39	Arg	**AGA**	**1.26**	**1.12**	**2.89**	**3.08**	**2.84**
	AAG	1.14	1.24	0.90	0.73	0.61		**AGG**	**1.26**	**1.23**	**1.61**	**1.81**	**1.83**
Asp	GAU	0.92	0.80	1.05	1.13	1.08	Gly	GGU	0.64	0.57	0.57	0.60	0.69
	GAC	1.08	1.20	0.95	0.87	0.92		GGC	1.36	1.46	0.62	0.55	0.62
Glu	**GAA**	**0.84**	**0.72**	**1.20**	**1.15**	**1.14**		**GGA**	**1.00**	**0.91**	**1.73**	**1.84**	**1.65**
	GAG	1.16	1.28	0.80	0.85	0.86		GGG	1.00	1.05	1.08	1.01	1.04

[a]*RSCU*, relative synonymous codon usage; *AA*, amino acid; *Cod*, codons; *HC*, human cells; *H1N1pdm*, 2009 H1N1 pdm Influenza A virus; H1N1 and H3N2, seasonal H1N1 and H3N2 Influenza A virus, respectively. Highly increased codons with respect to host cells ($\Delta \geq 0.30$) are shown in bold. Codons containing de dinucleotide CG are shown in italics. [b] RSCU codon usage of seasonal H1N1 and H3N2 according to Wong et al. (2010) [12].

dimensional factor (11.1% of the total variability) was analyzed, we found that the position of each strain significantly correlated ($r = 0.64$, $p < 0.0001$) with the respective usage of the dinucleotide CpG. Besides, although the global usage of this dinucleotide is very low, we found that the correlation is due to the differential usage of CGU (Arg) and CCG (Pro) codons, since these triplets display the most extreme values on the second dimensional factor (see Additional file 1: Table S1). Importantly, we also found that the third and the fourth dimensional factors of COA (8.7% and 5.5% of the total variability, respectively), are again mainly linked to the low usage of codons containing the dinucleotide CpG, mainly at the positions 2 and 3. Moreover, among the 16 dinucleotides, 15 are highly correlated with the first dimensional factor value in COA (Table 2). These observations indicate that the composition of dinucleotides also plays a crucial role in the variation found in synonymous codon usage among H1N1pdm IAV ORFs.

To study the possibility of codon usage variation in the H1N1pdm IAV genomes enrolled in this study, the

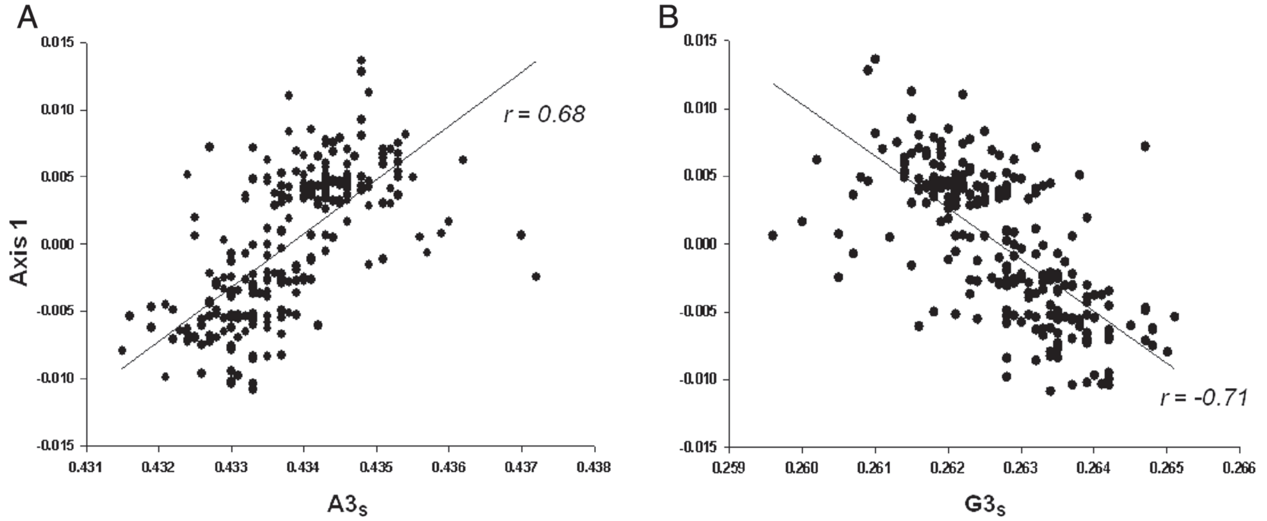


Figure 1 Association of purines at third codon positions with dimensional factor 1 generated by COA. In (**A**) and (**B**), the regression plots of the frequency of A3s and G3s *versus* the respective position of each strain in the first dimensional factor generated by the correspondence analysis on RSCU (COA-RSCU) are shown.

distribution of the 310 strains in the plane defined by the first two axes of COA was established. The results of these studies are shown in Figure 2.

Interestingly, the distribution of the H1N1pdm IAV strains in the plane defined by the first two major axes showed that the principal dimensional factor splits the strains at least three major groups: two of them discriminated by the first dimensional factor, while the third is revealed by the extreme low values on the second dimensional factor (Figure 2).

As the translation process represents a key step in the viral infection cycle, it is important to explore the strategies employed by the virus to harness the translation machinery of the cell host. Since variation at the third codon position makes possible the wobble interaction between that base and the first one of the anticodon [20], we wanted to gain further insight into the adaptation of H1N1pdm IAV strains to the respective host tRNA pool context. For this reason, the codon usage of virus (H1N1pdm IAV) was plotted against the codon usage of host (human cells) and the nucleotide that occupy the first anticodon position (wobble position) of the corresponding codon was identified. The results of these studies are shown in Figure 3.

As it can be seen in the figure, codon usage of virus and host is uncorrelated. The viral preference toward AT rich genomes and the T-headed anticodons is clear (Figure 3). This is also in agreement with the consequence of a differential usage of $A3_s$ and $G3_s$ (see also Figure 1). Comparison of these findings with the compilation of tRNAs species in the human genome [21] reveals that the virus highly preferred T-headed anticodons are not particularly adapted to the host transfer tRNA pool (Table 3). Therefore, there is no obvious correlation between the number of human host isoacceptor tRNAs and codon usage of the IAV enrolled in these studies.

Table 2 Summary of correlation analysis between the dimensional factors (DF) in COA and sixteen dinucleotides frequencies in H1N1 pdm IAV ORFs

		UU	UC	UA	UG	CU	CC	CA	CG
Mean ± SD[a]		0.893 ± 0.0054	0.814 ± 0.0050	0.736 ± 0.0009	1.215 ± 0.0009	0.797 ± 0.0056	0.672 ± 0.0033	1.326 ± 0.0042	0.319 ± 0.0020
DF 1[b]	*r*	0.43277	0.30726	0.50328	0.49116	0.16033	0.40283	0.44451	0.47789
	P	<0.0001	<0.0001	<0.0001	<0.0001	0.0048	<0.0001	<0.0001	<0.0001
		AU	**AC**	**AA**	**AG**	**GU**	**GC**	**GA**	**GG**
Mean ± SD[a]		1.281 ± 0.0046	0.926 ± 0.0039	1.804 ± 0.0071	1.327 ± 0.0037	0.682 ± 0.0076	0.703 ± 0.0009	1.472 ± 0.0012	1.040 ± 0.0018
DF 1[b]	*r*	0.44790	0.36540	0.61328	0.40489	0.08304	0.49579	0.48484	0.45555
	P	<0.0001	<0.0001	<0.0001	<0.0001	0.11880	<0.0001	<0.0001	<0.0001

[a] Mean values of 310 H1N1 pdm IAV strains' relative dinucleotide ratios ± standard deviation. [b] Correlation analysis between the first dimensional factor in COA and the sixteen dinucleotides frequencies in H1N1 pdm IAV ORF's is shown.

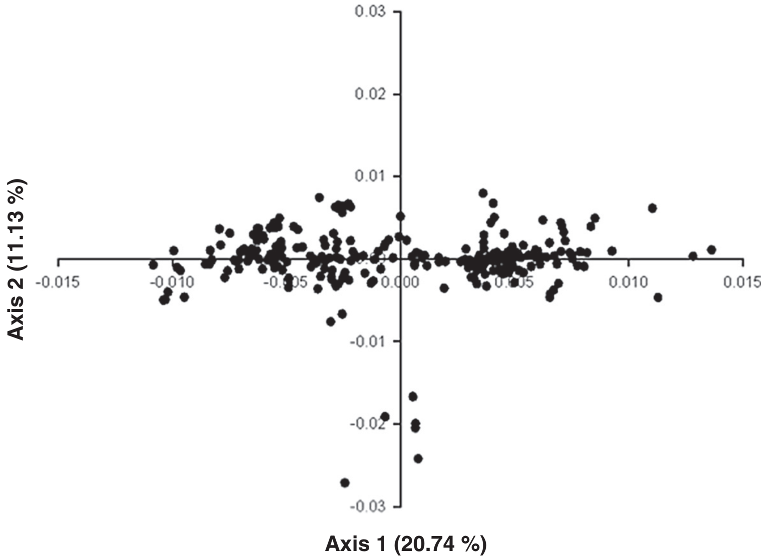


Figure 2 Position of the 310 H1N1 pdm IAV ORF's in the plane defined by the first two major axes generated by COA. The percentage of inertia of the first and second axes of COA is indicated for both axes between parentheses. The input values for COA were the RSCU values of each strain.

Discussion

As IAV relies on the host cell's machinery for its replication, codon usage bias could play a role in its adaptation to the host. The results of these studies revealed that codon usage in human IAV, including H1N1pdm, do not have the average codon usage pattern of their host's genes (see Table 1), in agreement with previous reports [12,16].

Comparisons to previous results reported for other IAV such H5N1 (mean ENC = 50.91) [16,22]; or other RNA viruses like SARS (mean ENC = 48.99) [23]; foot-and-mouth disease virus (mean ENC = 51.42) [24]; classical swine fever virus (mean ENC = 51.7) [19], Duck Enteritis virus (mean ENC =52.17) [25], Encephalomyocarditis virus (mean ENC = 54.86) [26] or Theilovirus (mean ENC = 51.08) [26], revealed that the ENC values found in this study for H1N1pdm IAV strains (mean ENC value of 52.5) are roughly similar to these previous findings, indicating that the overall extent of codon usage in these viruses are only slightly biased.

We have found a general link between codon usage bias and base composition, which is shown by the significant correlation of the position of each virus on the first dimensional factor of COA *vs.* the corresponding GC_3s, together with the opposite trends in relation to purines at third codon position (Figure 1A and B). Taken

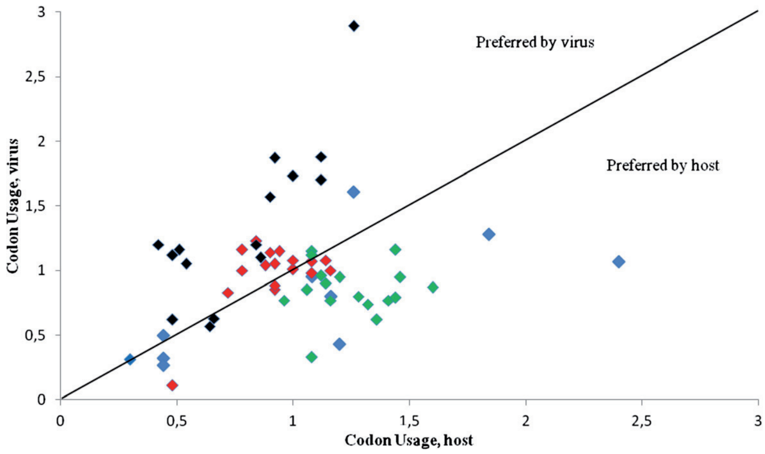


Figure 3 Codon usage of H1N1 pdm IAV plotted against the codon usage of human cells. Colors reflect the nucleotide that occupies the first anticodon position (wobble position) of the corresponding codon. A, C, G and T are indicated by red, blue, green and black diamonds, respectively.

Table 3 Frequency of tRNA genes in human cells for highly biased codons in H1N1 pdm IAV*

AA	Cod	Anticodon isotypes (tRNA count by anticodon)	Total tRNA anticodon count
Ala	**GCA**	**UGC**(**9**), AGC(29), GGC(0), CGC(5)	43
Arg	**AGA** & **AGG**	**UCU**(**6**), **CCU**(**5**), ACG(7), GCG(0), CCG(4), UCG(6)	28
Gln	**CAA**	**UUG**(**11**), CUG(21)	32
Glu	**GAA**	**UUC**(**13**), CUC(13)	26
Gly	**GGA**	**UCC**(**9**), GCC(15), CCC(7), ACC(0)	31
His	**CAU**	**AUG**(**0**), GUG(11)	11
Ile	**AUA**	**UAU**(**5**), AAU(14), GAU(8)	27
Leu	**CUA**	**UAG**(**3**), AAG(12), CAG(10), CAA(7), UAA(7), GAG(0)	39
Pro	**CCA**	**UGG**(**7**), AGG(10), GGG(0), CGG(4)	21
Ser	**UCA**	**UGA**(**5**), AGA(11), GGA(0), CGA(4), ACU(0),GCU(8)	28
Thr	**ACA**	**UGU**(**6**), AGU(10), GGU(0), CGU(6)	22
Val	**GUA**	**UAC**(**5**), CAC(16), AAC(11), GAC(0)	32

* Highly biased codons in H1N1 pdm IAV (as defined in Table 1) and their respective anticodons are shown in bold. *AA*, amino acid; *Cod*, codons.

together, our results indicate that the mutational bias is a very important trend in the evolution of H1N1pdm IAV genomes. However, this does not *per se* discards a role of other natural selection mechanisms acting in the IAV strains enrolled in these studies.

We have also found that CpG containing codons are sharply suppressed (see Table 1). This CpG deficiency was proposed to be related to the immunostimulatory properties of unmethylated CpG, which are recognized by the innate immune system of the host as a pathogen signature [24,27]. This is triggered by the intracellular Pattern Recognition Receptor (PRR) Tool-like 9 (TLR9), which activates several immune response pathways [28]. It seems reasonable to suggest that exists among vertebrates a TLR9-like mechanism acting at the RNA level [29]. Interestingly, previous studies have shown that IAV strains originated from an avian reservoir and infecting human hosts since 1918 has been selected under strong pressure to reduce the frequency of CpG in its genome [30]. Marked CpG deficiency has been observed in several other RNA viruses [24,31-35], including H1N1pdm IAV [12,30]. Then, escaping from the host antiviral response may act as another selective pressure contributing to codon usage in H1N1pdm IAV strains [36].

The distribution of the 310 H1N1 pdm IAV ORF's in the plane defined by the first two axes of COA shows the presence of at least three clusters of strains (Figure 2). Since species with a close genetic relationship always present a similar codon usage pattern [37] (see also Table 1), the results of these studies suggests that a dynamic process occurred in the H1N1pdm strains enrolled in these studies. This is reflected in the variation of codon usage observed among them (see Figure 2). These results suggest a balance of mutational bias and natural selection to shape codon usage in these strains, which allow the virus to explore and re-adapt its codon usage to different environments in a short period of time.

From the classical point of view, the preferred codons are recognized by the most abundant isoacceptors tRNAs, which implies the action of natural selection [38]. The results shown in Table 3 strongly suggest that this is not the case for H1N1pdm IAV strains. In other words, codon usage of these viruses does not seem to be adapted to the tRNA pool of the human cells but probably reflects the influence of mutational biases. Interestingly, this has been observed for some other RNA viruses, like HIV [39].

Understanding the mechanisms used by IAV to properly express its genes could suggest a novel point of intervention and drug targets. Reduced translation efficiency, particularly of structural genes that are needed for the formation of new particles, could affect viral success [40].

The results of this work suggest that synthetic attenuated virus engineering (SAVE) could play a role in creating new vaccines for IAV. By deoptimization of codon usage (replacing wild-type codons with codons and codon combinations whose sequences impair replication and/or expression), it might be possible to attenuate a virus [41]. Moreover, as the codon changes do not alter the protein sequence, the antigenicity should not differ from the wild-type virus. Besides, codon changes tend to have individually small fitness effects, so many nucleotide changes will be required to restore wild-type fitness, itself requiring 100 s or more generations [42-45]. This “death by a thousand cuts” strategy may provide an alternative method of attenuation [46]. Interestingly, it has been show that replacement of natural codons with synonymous triplets with increased frequencies of CpG gives rise to inactivation of Poliovirus infectivity [47]. Very recent studies revealed that this strategy can be applied to IAV [48].

Owing to known genome sequences, modern strategies of DNA synthesis have made it possible to recreate in principle all known viruses independent of natural templates [48]. Recoding of IAV to develop new vaccine candidates taking into account codon bias, base composition and adaptation to host tRNA by gene synthesis may provide important clues to elucidate virulence factors, identify targets for future drug intervention, and to develop new and appropriate vaccines [49].

Methods

Sequences and dataset

Sequences from H1N1pdm IAV strains, isolated from April to December of 2009, were obtained from The Influenza Virus Resource at the National Center for Biotechnological Information [50]. The data set comprised the complete genome sequences (eight segments) of 310 strains. For each strain the ORFs were concatenated (PB2 + PB1 + PA + HA + NP + NA + MP + NS) and aligned using the MUSCLE program [51]. The alignment is available upon request.

Codon usage analysis

Codon usage, base dinucleotide composition, G + C at synonymous variable third position codons (GC_3s), the relative synonymous codon usage (RSCU) [14] and the effective number of codons (ENC) [17] were calculated using the program CodonW (written by John Peden and available at http://sourceforge.net/projects/codonw/) as implemented in the Mobile server (http://mobyle.pasteur.fr). Codon usage data of influenza viral hosts, human (*Homo sapiens*) and domestic swine (*Sus scrofa*) were obtained from the codon usage database (available at: http://www.kazusa.or.jp/codon) [52]. The frequencies of tRNAs in human cells were retrieved from the GtRNAdb database [21].

Correspondence analysis(COA)

COA is an ordination technique that identifies the major trends in the variation of the data and distributes genes along continuous axes in accordance with these trends. COA creates a series of orthogonal axes to identify trends that explain the data variation, with each subsequent dimensional factor explaining a decreasing amount of the variation [18]. Each ORF is represented as a 59-dimensional and each dimension is related to the RSCU value of each triplet (excluding AUG, UGG and stop codons). This was done using the CodonW program.

Statistical analysis

Correlation analysis was carried out using Spearman's rank correlation analysis method [53].

Additional file

Additional file 1: Table S1. Each codon included in the correspondence analysis is represented by a row. Factor 1 and 2 columns contain the coordinate of the codon on the respective generated axis.

Competing interests

The authors declare that they do not have competing interests.

Authors' contributions

JC conceived of the study, and participated in its design and coordination. NG and AI have made substantial contributions to the design of the study, acquisition of data and analysis. VC, MS, GM, GC, have been involved in revising the manuscript critically for important intellectual content. JC wrote the paper. HM helped to draft the manuscript and made substantial and fundamental contributions to the interpretation and discussion of the results found in this work. All authors read and approved the final manuscript.

Acknowledgements

We acknowledge support by International Atomic Energy Agency, through Research Contract No. 15792 and Comisión Sectorial de Investigación Científica (CSIC), Universidad de la República, Uruguay, through I + D Project "Variabilidad genetica y evolución viral de virus Influenza A en Uruguay". Authors acknowledge support by Agencia Nacional de Investigación e Innovación (ANII) through project PE_ALI_2009_1_1603 and PEDECIBA, Uruguay. We also thank the support of Fondo Clemente Estable, 2007_ 722 to HM.

Author details

[1]Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares, Facultad de Ciencias, Universidad de la República, Iguá 4225, Montevideo 11400, Uruguay. [2]Laboratorio de Organización y Evolución del Genoma, Instituto de Biología, Facultad de Ciencias, Universidad de la República, Iguá 4225, Montevideo 11400, Uruguay. [3]Laboratorio de Evolución, Instituto de Biología, Facultad de Ciencias, Universidad de la República, Iguá 4225, Montevideo 11400, Uruguay. [4]Unidad de Proteínas Recombinantes, Institut Pasteur de Montevideo, Mataojo 2020, Montevideo 11400, Uruguay. [5]Unidad de Biofísica de Proteínas, Institut Pasteur de Montevideo, Mataojo 2020, Montevideo 11400, Uruguay.

Received: 23 February 2012 Accepted: 2 November 2012
Published: 8 November 2012

References

1. Neumann G, Brownlee GG, Fodor E, Kawaoka Y: **Orthomyxovirus: replication, transcription, and polyadenylation.** *Curr Top Microbiol Immunol* 2004, **283:**121–143.
2. Ahn I, Son HS: **Comparative study of the hemagglutinin and neuraminidase genes of Influenza A virus H3N2, H9N2 and H5N1 subtypes using bioinformatics techniques.** *Can J Microbiol* 2007, **53:**830–839.
3. Wolf YL, Viboud C, Holmes EC, Koonin EV, Lipman DJ: **Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus.** *Biol Direct* 2006, **1:**34.
4. Hillerman MR: **Realities and enigmas of human viral influenza: pathogenesis, epidemiology and control.** *Vaccine* 2002, **20:**3068–3087.
5. De Jong JC, Rimmelzwaan GF, Fouchier RA, Osterhaus AD: **Influenza virus: a master of metamorphosis.** *J Infection* 2000, **40:**218–228.
6. Ferguson NM, Galvani AP, Bush RM: **Ecological and immunological determinants of influenza evolution.** *Nature* 2003, **422:**428–433.
7. World Health Organization: *Pandemic (H1N1). Influenza-like illness in the United States and Mexico. 24 April 2009*. 2009. Available: http://www.who.int/csr/don/2009_04_24/en/index.html.
8. Centers for Disease Control and Prevention: **Update: infections with a swine-origin influenza A (H1N1) virus – United States and other countries, April 28th, 2009.** *Morb Mortal Wkly Rep* 2009, **58:**431–433.
9. Smith GJD, Vijaykrishna D, Bahl J, Lycett SJ, Worobey M, Pybus OG, Ma SK, Cheung CL, Raghwani J, Bhatt S, Peiris JSM, Guan Y, Rambaut A: **Origins and evolutionary genomics of the 2009 swine-origin H1N1 Influenza A epidemic.** *Nature* 2009, **459:**1122–1125.

10. Gorman OT, Bean WJ, Kawaoka Y, Donatelli I, Guo YJ, Webster RG: **Evolution of influenza A virus nucleocapsid genes: implications for the origins of H1N1 human and classical swine viruses.** *J Virol* 1991, **65**:3704–3714.
11. Stoletzki N, Eyre-Walker A: **Synonymous codon usage in Escherichia coli: selection for translational accuracy.** *Mol Biol Evol* 2007, **24**:374–381.
12. Wong E, Smith DK, Rabadan R, Peiris M, Poon L: **Codon usage bias and the evolution of Influenza A viruses. Codon usage biases of Influenza virus.** *Evol Biol* 2010, **10**:253.
13. Ikemura T: **Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs.** *J Mol Biol* 1982, **158**:573–597.
14. Sharp PM, Li WH: **An evolutionary perspective on synonymous codon usage in unicellular organisms.** *J Mol Evol* 1986, **24**:28–38.
15. Kryazhimskiy S, Bazykin GA, Dushoff J: **Natural selection for nucleotide usage at synonymous and non-synonymous sites in the influenza A genes.** *J Virol* 2008, **82**:4938–4945.
16. Zhou T, Gu W, Ma J, Sun X, Lu Z: **Analysis of synonymous codon usage in H5N1 virus and other influenza A viruses.** *Biosystems* 2005, **81**:77–86.
17. Comeron JM, Aguade M: **An evaluation of measures of synonymous codon usage bias.** *J Mol Evol* 1998, **47**:268–274.
18. Greenacre M: *Theory and applications of correspondence analysis*. London: Academic; 1984.
19. Tao P, Dai L, Luo M, Tang F, Tien P, Pan Z: **Analysis of synonymous codon usage in classical swine fever virus.** *Virus Genes* 2009, **38**:104–112.
20. Crick FHC: **Codon-anticodon pairing – Wobble hypothesis.** *J Mol Biol* 1966, **19**:548–555.
21. Chan PP, Lowe TM: **GtRNAdb: a database of transfer RNA genes detected in genomic sequence.** *Nucleic Acids Res* 2009, **37**:D93–D97.
22. Li ZP, Ying DQ, Li P, Li F, Bo XC, Wang SQ: **Analysis of synonymous codon usage bias in 09H1N1.** *Vir Sin* 2010, **25**:329–340.
23. Woo PCY, Wong BHL, Huang Y, Lau SKP, Yuen K: **Cytosine deamination and selection of CpG suppressed clones are the two major independent biological forces that shape codon usage bias in Coronaviruses.** *Virology* 2007, **369**:431–442.
24. Zhong J, Li Y, Zhao S, Liu S, Zhang Z: **Mutation pressures shapes codon usage in the GC-rich genome of foot-and-mouth disease virus.** *Virus Genes* 2007, **35**:767–776.
25. Jia R, Cheng A, Wang M, Xin H, Guo Y, Zhu D, Qi X, Zhao L, Ge H, Chen X: **Analysis of synonymous codon usage in the UL24 gene of duck enteritis virus.** *Virus Genes* 2009, **38**:96–103.
26. Liu WQ, Zhang J, Zhang YQ, Zhou JH, Chen HT, Ma LN, Ding YZ, Liu Y: **Compare the differences of synonymous codon usage between the two species within cardiovirus.** *Virology J* 2011, **8**:325.
27. Shackelton LA, Parrish CR, Holmes EC: **Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses.** *J Mol Evol* 2006, **62**:551–563.
28. Dorn A, Kippenberger S: **Clinical application of CpG-, non-CpG, and antisense oligodeoxynucleotides as immunomodulators.** *Curr Opin Mol Ther* 2008, **10**:10–20.
29. Lobo FP, Mota BEF, Pena SDJ, Azevedo V, Macedo AM, Tauch A, Machado CR, Franco GR: **Virus-host coevolution: common patterns of nucleotide motif usage in Flaviviridae and their hosts.** *PLoS One* 2009, **4**:6282.
30. Greenbaum BD, Levine AJ, Bhanot G, Rabadan R: **Patterns of evolution and host gene mimicry in influenza and other RNA viruses.** *PLoS Pathog* 2008, **4**:1000079.
31. Rabadan R, Levine AJ, Robins H: **Comparison of avian and human influenza A viruses reveals a mutational bias on the viral genomes.** *J Virol* 2006, **80**:11887–11891.
32. Rothberg PG, Wimmer E: **Mononucleotide and dinucleotide frequencies and codon usage in poliovirus RNA.** *Nucleic Acids Res* 1981, **9**:6221–6229.
33. Karlin S, Doerfler W, Cardon LR: **Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses?** *J Virol* 1996, **68**:2889–2897.
34. Martínez-Gómez M, López-Tort F, Volotao-Ede M, Recarey R, Moratorio G, Musto H, Leite JP Cristina J: **Analysis of human P[4]G2 rotavirus strains isolated in Brazil reveals codon usage bias and strong compositional constraints.** *Infec Genet Evol* 2011, **11**:580–586.
35. D'Andrea L, Pintó RM, Bosch A, Musto H, Cristina J: **A detailed comparative analysis on the overall codon usage patterns in hepatitis A virus.** *Virus Res* 2011, **157**:19–24.
36. Vetsigian K, Goldenfeld N: **Genome rhetoric and the emergence of compositional bias.** *Proc Nat Acad Sci USA* 2009, **106**:215–220.
37. Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F: **Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens: a review of the considerable within-species diversity.** *Nucleic Acids Res* 1988, **16**:8207–8211.
38. Ikemura T: **Codon usage and tRNA content in unicellular and multicellular organisms.** *Mol Biol Evol* 1985, **2**:13–34.
39. van Weringh A, Ragonnet-Cronin M, Pranckeviciene E, Pavon-Eternod M, Kleiman L, Xia X: **HIV-1 modulates the tRNA pool to improve translation efficiency.** *Mol Biol and Evol* 2011, **28**:1827–1834.
40. Ngumbela KC, Ryan KP, Sivamurthy R, Brockman MA, Gandhi RT, Bhardwaj N, Kavanagh DG: **Quantitative effect of suboptimal codon usage on translational efficiency of mRNA encoding HIV-1 gag in intact T cells.** *PLoS ONE* 2008, **3**:2356.
41. Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Muller S: **Virus attenuation by genome-scale changes in codon pair bias.** *Science* 2008, **320**:1784–1787.
42. Burns CC, Shaw J, Campagnoli R, Jorba J, Vincent A, Quay J, Kew O: **Modulation of poliovirus replicative fitness in HeLa cells by deoptimization of synonymous codon usage in the capsid region.** *J Virol* 2006, **80**:3259–3272.
43. Mueller S, Papamichail D, Coleman JR, Skiena S, Wimmer E: **Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity.** *J Virol* 2006, **80**:9687–9696.
44. Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S: **Virus attenuation by genome-scale changes in codon pair bias.** *Science* 2008, **320**:1784–1787.
45. Bull JJ, Molineux IJ, Wilke CO: **Slow fitness recovery in a codon-modified viral genome.** *Mol Biol Evol* 2012, doi:10.1093/molbev/mss119.
46. Mueller S, Coleman JR, Papamichail D, Ward CB, Nimnaual A, Futcher B, Skiena S, Wimmer E: **Live attenuated Influenza virus vaccines by computer-aided rational design.** *Nature Biotech* 2010, **28**:723–726.
47. Burns CC, Campagnoli R, Shaw J, Vincent A, Jorba J, Kew O: **Genetic inactivation of poliovirus infectivity by increasing the frequencies of CpG and UpA dinucleotides within and across synonymous capsid region codons.** *J Virol* 2009, **83**:9957–9969.
48. Wimmer E, Paul AV: **Synthetic poliovirus and other designer viruses: what have we learned from them?** *Ann Rev Microbiol* 2011, **65**:583–609.
49. Wimmer E, Mueller S, Tumpey TM, Taubenberger JK: **Synthetic viruses: a new opportunity to understand and prevent viral disease.** *Nature Biotech* 2009, **27**:1163.
50. Bao Y, Bolotov D, Dernovoy B, Kiryutin L, Zaslavsky L, Tatusova T, Ostell J, Lipman D: **The Influenza Virus Resource at the National Center for Biotechnology Information.** *J Virol* 2008, **82**:596–601.
51. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5**:113.
52. Nakamura Y, Gojobori T, Ikemura T: **Codon usage tabulated from international DNA sequence databases: status for the year 2000.** *Nucleic Acids Res* 2000, **28**:292.
53. Wessa P: *Free Statistics Software*, Office for Research Development and Education, version 1.1.23-r7. URL: http://www.wessa.net.

doi:10.1186/1743-422X-9-263
Cite this article as: Goñi *et al.*: Pandemic influenza A virus codon usage revisited: biases, adaptation and implications for vaccine strain development. *Virology Journal* 2012 **9**:263.

Supplementary Table 1.

Each codon included in the correspondence analysis is represented by a row. Factor 1 and 2 columns contain the coordinate of the codon on the respective generated axis.

Codon	Factor 1	Codon	Factor 2
CGA	-0.0172	CGU	-0.0253
GUC	-0.0142	UUA	-0.0212
UUA	-0.0096	ACG	-0.0060
AUA	-0.0083	CUG	-0.0057
CAA	-0.0053	CCA	-0.0049
GAU	-0.0042	GUA	-0.0039
ACC	-0.0038	CGA	-0.0037
ACA	-0.0038	CAC	-0.0028
GCA	-0.0033	GGC	-0.0020
CGC	-0.0028	ACC	-0.0020
UGC	-0.0023	AUC	-0.0020
UCC	-0.0021	CAA	-0.0019
UUG	-0.0021	AAG	-0.0017
GCU	-0.0019	AAU	-0.0017
CCA	-0.0017	AAA	-0.0009
GUG	-0.0017	GGG	-0.0008
UAC	-0.0013	ACA	-0.0007
AAA	-0.0013	GCC	-0.0007
AGU	-0.0011	UUU	-0.0007
GGU	-0.0011	GAA	-0.0006
GGG	-0.0010	GAC	-0.0006
CUA	-0.0009	UGC	-0.0005
UCG	-0.0009	UCC	-0.0005
CAU	-0.0007	CCC	-0.0004
CUU	-0.0003	UAC	-0.0003
AAC	-0.0003	CUC	-0.0003
UUC	-0.0002	UCA	-0.0001
GAA	-0.0001	AUU	-0.0001
UCU	0.0000	CUU	-0.0001
GGC	0.0001	AGA	-0.0001
AGA	0.0002	UAU	0.0001
AGC	0.0003	GCA	0.0003
AAG	0.0005	GAU	0.0005
AUU	0.0006	GGU	0.0005
GAG	0.0006	GUU	0.0005
ACU	0.0006	AGC	0.0005
GGA	0.0007	GGA	0.0006
GCC	0.0008	UUC	0.0006
UAU	0.0009	GCU	0.0007
AGG	0.0009	UCU	0.0007
UUU	0.0009	CAU	0.0007
CCU	0.0009	UGU	0.0008
ACG	0.0009	AGG	0.0011
CUC	0.0012	CCU	0.0011
GAC	0.0015	AGU	0.0013
CCC	0.0017	GAG	0.0015
AAU	0.0023	ACU	0.0015

CCG	0.0028	GUG	0.0016
UGU	0.0028	CAG	0.0020
GUU	0.0028	AUA	0.0024
CAC	0.0029	UUG	0.0029
CGU	0.0029	GCG	0.0036
UCA	0.0050	AAC	0.0037
CAG	0.0056	GUC	0.0040
AUC	0.0092	CGG	0.0045
CUG	0.0109	CGC	0.0103
GUA	0.0129	CUA	0.0108
CGG	0.0255	UCG	0.0113
GCG	0.0297	CCG	0.0131

Infection, Genetics and Evolution 14 (2013) 396–400

Contents lists available at SciVerse ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

A detailed comparative analysis on the overall codon usage patterns in West Nile virus

Gonzalo Moratorio [a,b,1], Andrés Iriarte [c,d,1], Pilar Moreno [a,e], Héctor Musto [c], Juan Cristina [a,*]

[a] *Laboratorio de Virología Molecular, Centro de Investigaciones Nucleares, Facultad de Ciencias, Universidad de la República, Iguá 4225, 11400 Montevideo, Uruguay*
[b] *Unidad de Biofísica de Proteínas, Institut Pasteur de Montevideo, Mataojo 2020, 11400 Montevideo, Uruguay*
[c] *Laboratorio de Organización y Evolución del Genoma, Instituto de Biología, Facultad de Ciencias, Universidad de la República, Iguá 4225, 11400 Montevideo, Uruguay*
[d] *Laboratorio de Evolución, Instituto de Biología, Facultad de Ciencias, Universidad de la República, Iguá 4225, 11400 Montevideo, Uruguay*
[e] *Unidad de Proteínas Recombinantes, Institut Pasteur de Montevideo, Mataojo 2020, 11400 Montevideo, Uruguay*

ARTICLE INFO

Article history:
Received 21 May 2012
Received in revised form 2 January 2013
Accepted 3 January 2013
Available online 16 January 2013

Keywords:
West Nile virus
Codon usage
Evolution

ABSTRACT

West Nile virus (WNV) is a member of the family *Flaviviridae* and its genome consists of an 11-kb single-stranded, positive-sense RNA. WNV is maintained in an enzootic cycle between mosquitoes and birds, but can also infect and cause disease in horses and humans, which serve as incidental dead-end hosts. Understanding the extent and causes of biases in codon usage is essential to the comprehension of viral evolution. In this study, we performed a comprehensive analysis of 449 WNV strains, for which complete genome sequences are available. Effective number of codons (ENC) indicates that the overall codon usage among WNV strains is only slightly biased. Codon adaptation index (CAI) values found for WNV genes are different from the CAI values found for human genes. The relative synonymous codon usage among WNV strains isolated from birds, equines, humans and mosquitoes are roughly similar and are influenced by the relative dinucleotide frequencies. Taking together, the results of this work suggest that WNV genomic biases are the result of the evolution of genome composition, the need to escape the antiviral cell responses and a dynamic process of mutation and selection to re-adapt its codon usage to different environments.

1. Introduction

West Nile virus (WNV) is a member of the family *Flaviviridae* and belongs to the genus *Flavivirus*, which consists of more than 70 species. Among these are several arthropod-transmitted viruses with clinical importance, most prominently dengue virus (DENV*), Yellow fever virus* (YFV), *Tick-borne encephalitis virus* (TBEV) and *Japanese encephalitis virus* (JEV). Flaviviruses cause severe health problems in nearly all parts of the world. Within flaviviruses, WNV is classified into the in JEV serogroup, which includes also *Murray Valley encephalitis virus* (MVEV), *St. Louis encephalitis virus* (SLEV) and *Usutu virus* (USUV) (Calisher et al., 1989; May et al., 2011; Ulbert, 2011). Flaviviruses have an 11-kb single-stranded, positive-sense RNA genome which is translated into a single polyprotein upon infection of the host cell. This polypeptide is enzymatically processed by both viral and host cell proteases, yielding the three structural proteins (C, prM and E) and seven non-structural proteins (NS1, 2A, 2B, 3, 4A, 4B and 5) (Roosendaal et al., 2006). WNV is maintained in an enzootic cycle between mosquitoes and birds, but can also infect and cause disease in horses and humans, which serve as incidental dead-end hosts (Pesko and Ebel, 2012; Lim et al., 2011). WNV is endemic in some regions of Africa, Europe, the Middle East and Asia (Dauphin et al., 2004). As with other vector-borne diseases, the warmer temperature in the tropics facilitates longer transmission seasons and sometimes increased transmission intensity through faster mosquito and virus development and increased biting rates. Following its emergence in the United States in 1999, it has rapidly spread across North America, and has been recently reported in South America and the Caribbean (Komar and Clark, 2006). WNV is currently the most widely distributed of the encephalitic flaviviruses (May et al., 2011).

Compared to most mosquito-borne viruses, WNV has an enormous vector and host range. More than 300 avian species are susceptible and many of these develop high viral serum titers during the acute phase of infection, although members of the family Passeriformes (particularly *Passer domesticus, Turdus migratorius, Sturnus vulgaris, Cyanocitta cristata* and *Carpodacus mexicanus*) are presumed to be the most important avian hosts in both Europe and the Americas. On the other hand, most isolations from mosquitoes come from *Culex* species (particularly *Culex pipiens* and *Culex pipiens quinquefasciatus*) (Hamer et al., 2009; Komar et al., 2003).

[*] Corresponding author. Tel.: +598 2 525 09 01; fax: +598 2 525 08 95.
E-mail address: cristina@cin.edu.uy (J. Cristina).

[1] These authors contributed equally to this work.

1567-1348/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.meegid.2013.01.001

The redundancy of the genetic code, in which most of the amino acids can be translated by more than one codon, offers evolution the opportunity to tune the efficiency and accuracy of protein production to various levels while maintaining the same amino-acid sequence (Stoletzki and Eyre-Walker, 2007). The various codons that correspond to the same amino acid are often considered 'synonymous,' yet their corresponding tRNAs might differ in their amounts in cells and thus also in the speed in which they will be recognized by the ribosome. The alternative nucleotide sequences of the various codon choices for a protein might give rise to transcripts with different secondary structure and stability, which may affect translation and even folding (Ikemura, 1982). The number of alternative nucleotide sequences that could still code for the same protein is astronomical, leaving many degrees of freedom that evolution could use for achieving control without affecting the protein sequence. While the non-random usage of synonymous codons is often correctly assumed to reflect the action of neutral drift, in an increasing number of cases it now turns out to reflect the result of natural selection, perhaps mainly for tuning efficiency and accuracy of translation (Gingold and Pilpel., 2011).

The differential usage of the synonymous codons (among other aspects of genome evolution) might be important for the comprehension of the viral biology, in particular, the interplay between viruses and the immune response (Shackelton et al., 2006). Indeed, as is well known, synonymous triplets are generally not used randomly, and the main forces that drive this bias from equal usage are natural selection (which is related to translation efficiency at two different levels: speed and accuracy) and mutational biases. In spite of recent efforts to understand codon usage biases in viruses (Liu et al., 2011, 2012; Zhou et al., 2012; Pandit and Sinha, 2011; D'Andrea et al., 2011), more studies are needed in order to address the evolutionary forces that influence the observed patterns. Because of the comparative small genome size and some other features generally associated with viruses (for example recursive bootlenecks), very probably they are submitted to different constraints in relation to prokaryotic and eukaryotic genomes. Since their replication and protein synthesis depends on the host's machinery, the interplay of codon usage in the virus and the host is expected to affect viral fitness, survival and evolution. For these reasons, a detailed understanding of the evolution of WNV codon usage in relation to the host is much needed.

In order to gain insight into these matters, we performed comprehensive analyses of codon usage and composition of 449 WNV strains, which represent all the complete genome sequences available in the databases, and investigated the possible key evolutionary determinants of the biases found.

2. Materials and methods

2.1. Sequences

Complete genome sequences of 449 WNV strains were obtained from DDBJ database (available at: http://arsa.ddbj.nig.ac.jp/). For strain names and accession numbers see Supplementary Material Table 1. The data set comprised a total of 1541,388 codons.

2.2. Data analyses

Codon usage, dinucleotide frequencies, base composition and the relative synonymous codon usage (RSCU) (Sharp and Li, 1986) were calculated using the program CodonW (written by John Peden and available at http://sourceforge.net/projects/codonw/). Moreover, for each strain, we computed the effective number of codons (ENC) (Novembre, 2002) using INCA2.1 (Supek and Vlahovicek, 2004). This index can vary between 20 and 61, and a low value indicates a strong bias in codon usage. Dinucleotide expected values were calculated assuming random association of bases from the observed frequencies of each base for every sequence. In order to study codon usage preferences in WNV in relation to the codon usage of WNV hosts, we employed the codon adaptation index (CAI) (Sharp and Li, 1987). CAI was calculated using the approach of Puigbo et al. (2008a) (available at: http://genomes.urv.es/CAIcal) for WNV, human and *Culex pipiens quinquefasciatus* (http://www.kazusa.or.jp/codon/). This method allows to compare a given codon usage (in our case, WNV) to a predefined reference set (human or mosquitoes). In order to show whether the WNV genes are or not better adapted to the codon usage of the reference sets (human or mosquitoes) than the genes that define the reference set itself, as measured by CAI, we constructed two different datasets: one composed of 322 human genes selected at random from Ensembl Database (available at: http://www.ensembl.org); and 77 *Culex pipiens quinquefasciatus* genes obtained from ARSA at DNA Database of Japan (available at: arsa.ddbj.nig.ac.jp). Statistically significant differences among CAI values obtained in different comparisons was addressed by means of the use of a Student *t*-test and a Wilcoxon & Mann–Whitney test (Wessa, 2012).

In order to discern if the statistically significant differences in the CAI values arise from codon preferences, we used E-CAI (Puigbo et al., 2008b) to calculate the expected value of CAI (eCAI) at the 95% confident interval. A Kolmogorov–Smirnov test for the excepted CAI was also performed (Puigbo et al., 2008b). Total G+C genomic content, as well as G+C content at first, second and third codon positions were also calculated using the approach of Puigbo et al. (2008a).

2.3. Multivariate analysis

The relationship between variables and samples can be obtained using multivariate statistical analysis. COA is a type of multivariate analysis that allows a geometrical representation of the sets of rows and columns in a dataset (Wong et al., 2010; Greenacre, 1984). Each ORF is represented as a 59-dimensional vector and each dimension correspond to the RSCU value of one codon (excluding AUG, UGG and stop codons).Major trend within a dataset can be determined using measures of relative inertia and genes ordered according to their position along the axis of major inertia (Tao et al., 2009). COA was performed on the RSCU values using the CodonW program. Principal component analysis (PCA) was also carried out using INCA2.1 (Supek and Vlahovicek, 2004).

3. Results and discussion

3.1. Codon usage variation among WNV strains

In order to investigate if these WNV strains display similar codon usage biases, the ENC' values were calculated for the 449 strains enrolled in this study. A mean value of ENC of 53.81 ± 0.11 was obtained. This suggests that the overall codon usage among these strains is similar and only slightly biased.

In order to gain insight into these matters, a COA was performed on the RSCU values for all WNV strains enrolled in these studies. The first axis generated by the analysis accounts for 27.7% of the total variation, while the next three main axes account for 12.4%, 9.3% and 7.8%, respectively. The first one is highly correlated with the frequencies of some dinucleotides (Table 1) and negatively correlated with the GC content at the second codon positions ($r = -0.42$). The principal component (PC) analysis show nearly identical results (data not shown).

Table 1
Correlations between the dinucleotide frequencies in the viruses strains against the respective position of the four main axes generated by COA.[a]

Dinucleotide	Correlation coefficient (r)			
	Axis 1 (27.7%)	Axis 2 (12.4%)	Axis 3 (9.0%)	Axis 4 (8.3%)
TT	−0.66	−0.43	−0.12	0.14
TC	0.38	0.24	−0.08	−0.31
TA	0.46	−0.25	−0.35	0.06
TG	−0.08	−0.11	0.17	−0.33
CT	0.70	0.27	−0.22	−0.18
CC	0.32	0.34	0.05	0.02
CA	−0.04	−0.15	0.60	−0.13
CG	−0.35	0.33	−0.29	0.35
AT	0.22	−0.16	−0.03	−0.55
AC	−0.12	−0.16	0.37	0.11
AA	0.32	−0.11	0.22	−0.30
AG	−0.04	−0.13	−0.08	0.45
GT	−0.09	−0.12	−0.10	0.41
GC	−0.61	0.31	−0.05	−0.03
GA	−0.84	0.05	−0.20	0.19
GG	0.53	0.45	0.09	−0.25

[a] In parentheses, the % of the total variation explained by each axis is shown. Values greater than ±0.2 are significant ($p < 0.01$).

3.2. General codon usage pattern in WNV

Base composition analyses show that the average G+C content in the genome sequence of WNV strains is 50.43 ± 0.14%. Discriminated by codon positions, the GC contents are 47.70 ± 0.21, 46.50 ± 0.11% and 56.89 ± 0.46% for the first, second and third positions, respectively, (see Table 2). This suggests that compositional bias is not very strong in this virus.

In order to compare the codon usage preferences of WNV with those of its hosts, CAI values for all triplets were calculated, using human and *Culex pipiens quinquefasciatus* codon usage as reference sets. The results of these studies are shown in Table 2.

CAI index ranges from 0 to 1, being 1 if the frequency of codon usage by WNV equals the frequency of usage of the reference set. A mean CAI of 0.779 were obtained for WNV genes in relation to human codon usage reference set, while a mean CAI of 0.809 were obtained for a human genes dataset in relation to the same reference codon usage (see Table 2). In order to observe if the differences in CAI values among the two comparisons were statistically significant, we performed a Student *t*-test and a Wilcoxon & Mann–Whitney test. The results of these tests revealed that the differences in CAI values are statistically significant ($t = -8.0822$, *p*-value = 0 and $T = 26954$, *p*-value = 0, respectively). Similar studies were carried out to observe CAI values of WNV in relation to mosquito codon usage reference set, where a mean CAI of 0.704 were obtained, while a mean CAI value of 0.731 was obtained using a mosquito gene dataset in relation to the same codon usage referent set (see Table 2). Again, Sudent *t*-test and Wilcoxon & Mann–Whitney test revealed that the differences in CAI are statistically significant ($t = -5.72363$, *p*-value = 0 and $T = 25596$, *p*-value = 0, respectively). Besides, comparable G+C content among the three species was found (see Table 2). Unfortunately, no suitable codon usage frequencies were available for comparison among virus and *Passer domesticus* or other bird hosts.

In order to discern if the statistically significant differences in the CAI values arise from codon preferences (Puigbo et al., 2008b), the expected CAI (e-CAI) values were calculated for WNV sequences in relation to human codon usage set. The e-CAI algorithm (Puigbo et al., 2008b) generated 500 random sequences with the same nucleotide and amino acid composition as the sequences of interest (in this case WNV sequences), calculated the CAI values for all of them, and applied a Kolmogorov–Smirnov test for the e-CAI of these random sequences in order to show whether the generated sequences follow a normal distribution. For these random sequences the 95% coverage is retrieved. The results of these studies revealed an e-CAI value of 0.758 ($p < 0.05$) and Kolmogorov–Smirnov test revealed a normal distribution of the generated sequences (Kolmogorov–Smirnov test of e-CAI value of 0.037, bellow of critical value of 0.061). The CAI values obtained for the actual WNV sequences (0.779) are above the values obtained for the 95% coverage. Besides, the values obtained for the actual WNV sequences have a lower CAI value than the average human genes (see Table 2). The results of these studies revealed that the CAI values for WNV genes are different from the CAI values obtained for human ones. These differences seem to be due to codon preferences in their codon usage.

Similar studies carried out in WNV sequences in relation with *Culex pipiens quinquefasciatus* codon usage set revealed an e-CAI value of 0.705 and a Kolmogorov–Smirnov test for e-CAI value of 0.030 (< of critical value of 0.061). The CAI values obtained for the actual WNV sequences in this case (0.704) are roughly situated below the values obtained for the 95% coverage. The CAI values obtained for WNV are in the range of the CAI values obtained for *Culex pipiens quinquefasciatus* genes (0.731 ± 0.053). Although we used all known (77) *Culex pipiens quinquefasciatus* genes sequences as a dataset, these results revealed that more studies will be needed in order to address the possible differences in compositional biases among WNV and *Culex pipiens quinquefasciatus* codon usage.

When WNV codons are sorted according to their position on the four major axes of COA, it is evident that the most extreme values are displayed by rarely used triplets, almost all of them containing the dinucleotides CpG and UpA. These codons are: for axis 1 CAU (His) and GUA (Val), axis 2 (CGU (Arg) and UUA (Leu), axis 3 CCG (Pro) and UUA (Leu) and for axis 4 CGA (Arg) and CAU (His). This fact reinforces the importance of CpG and UpA for the overall pattern of codon usage among WNV strains.

It has been suggested that dinucleotide frequencies can affect codon bias (Tao et al., 2009). To study this potential effect, the relative abundances of the 16 dinucleotides were calculated. The histogram of the ratio observed frequency/expected frequency for each dinucleotide in each codon position (namely 1–2, 2–3 and 3–1) is displayed in Fig. 1. When global values are considered it can be seen that UpC, CpC, ApU, ApC, ApA, ApG, GpC and GpG are around the expected values. The most divergent dinucleotides are UpA and CpG (depleted) and UpG and CpA (overrepresented) (Fig. 1). When the codon position of the dinucleotide is considered,

Table 2
Codon adaptation of West Nile virus genes in relation to Human and *C. pipiens* codon usage, displayed as CAI[a] values.

	CAI-Hs	CAI-Cp	%GC	%GC(1)	%GC(2)	%GC(3)
WNV genes	0.779 ± 0.003	0.704 ± 0.003	50.43 ± 0.14	47.70 ± 0.21	46.50 ± 0.11	56.89 ± 0.46
Human genes	0.809 ± 0.038	NA	56.47 ± 7.72	58.37 ± 6.88	44.35 ± 7.71	66.69 ± 14.81
C. pipiens genes	NA	0.731 ± 0.053	52.08 ± 7.78	52.20 ± 9.14	43.14 ± 6.90	60.91 ± 14.54

[a] CAI, codon adaptation index; CAI-Hs and CAI-Cp, codon adaptation index in relation to *Homo sapiens* and *Culex pipiens quinquefasciatus* reference codon usage set, respectively.%GC, percentage of G+C genomic content,%GC(1) through (3), percentage of G+C genomic content in codon positions 1 through 3, respectively. NA, not applicable. In all cases, mean ± standard deviation values are shown.

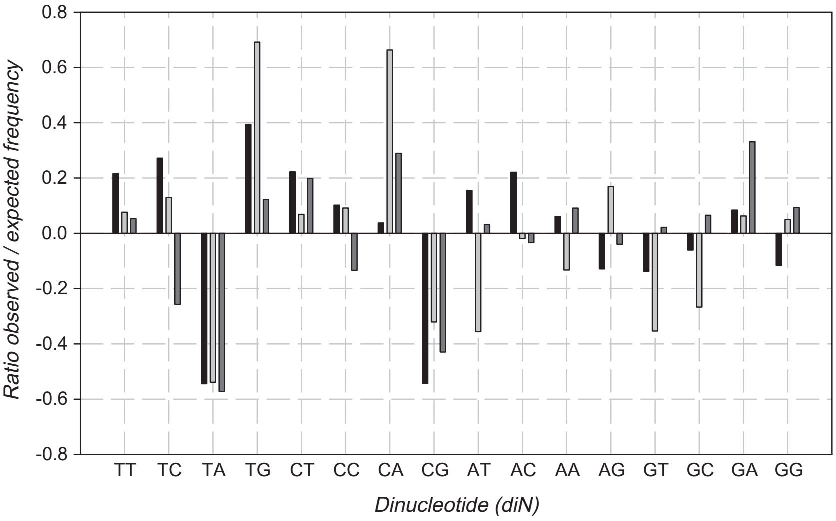


Fig. 1. Mean dinucleotide frequency among WNV strains. Histogram of the ratios of observed frequency/expected frequency for each dinucleotide in each codon position (namely 1–2, 2–3 and 3–1) is shown.

the pattern is rather similar to the global values (Fig. 1). This means that depletion and overrepresentation of dinucleotides is expected to affect not only synonymous codon preferences but also amino acid frequencies and the choice of the synonymous codon usage according to the first base in the next triplet. Furthermore, the depletion of some synonymous codons is directly linked to the presence of CpG and UpA in the second and third codon positions. This is the case of UUA, CUA, AUA, GUA and UCG, CCG, ACG and GCG (see Table 1).

These biases can be marginally noted in the amino acid usage (first and second codon positions). This can be easily explained considering the structure of the genetic code. Since Arg is coded by six different synonymous triplets, the depletion of CGN codons does not affect the usage of this amino acid given the increment of AGA and AGG. On the other hand, Tyr (encoded by UAU and UAC) is not affected by the depletion of UpA, since its observed/expected ratio is 0.99. (Note that the other two UpA- containing triplets, namely UAA and UAG are stop codons, and this virus uses only one by genome, since WNV RNA genome is translated into a single polyprotein upon infection of the host cell).

We also analyzed the synonymous codon usage biases according to the first base in the next codon position (Supplementary Material Fig. 1). The results confirm the avoidance of UpA and CpG. In the case of the latter dinucleotide, when the first base of the next triplet is a G, C is avoided in the third codon position of the previous triplet and A is preferred in the same position. Interestingly, the usage of the C-ending triplets is preferred if there is an A in the first position of the next codon. A similar trend is observed in the case of UpA (i.e., U tends to be avoided in third codon position when the next triplet begins with A). This suggests some kind of compensatory strategy among synonymous codon usage in order to avoid changes in global genomic GC content. Indeed, this specific base composition assures a minimum possible frequency of both negatively selected dinucleotides and highligths the importance of dinucleotide frequencies for synonymous codon usage in WNV.

Previous studies have proposed that the deficiency of the dinucleotide CpG is related to the immunostimulatory properties of unmethylated CpG, which is recognized by the host's innate immune system as a pathogen signature (Shackelton et al., 2006; Woo et al., 2007). This is now known to be triggered by the intracellular Pattern recognition receptor (PRR) Toll-like 9 (TLR9), which recognizes unmethylated CpG in DNA, and triggers several immune response pathways (Dorn and Kippenberger, 2008). Since the vertebrate immune system relies on unmethylated CpG recognition in DNA molecules as a sign of infection, and CpG under-representation in RNA viruses is exclusively observed in vertebrate viruses (Lobo et al., 2009), it is reasonable to suggest that a TLR9-like mechanism exists in the vertebrate immune system which recognizes CpG in RNA molecules, triggering immune responses. Interestingly, there is an absence of CpG depletion in flavivirus which replicate only in mosquitoes, suggesting that this depletion in WNV is induced by their adaptation to vertebrate hosts (Lobo et al., 2009).

An UpA recognition system in viral RNA sequences has been previously described as a vertebrate immune response mechanism. As part of dsRNA-activated antiviral pathway, most vertebrates possess a latent intracellular interferon-induced ribonucease, named Ribonuclease L (RNase L), which degrades this kind of RNA molecules and activates apoptotic pathways (Bisbal and Silverman, 2007; Player and Torrence, 1998). WNV are known to be recognized by RNase L preferentially at UpA or UpU sites (Scherbik et al., 2006). Nevertheless, UpA under-representation has also been found in flaviviruses which replicate only in insects (Lobo et al., 2009). Since mosquito immune system has no interferon-based antiviral mechanisms and no known nucleotide motif recognition signaling pathways, the roll of mosquito vector immune system in shaping virus dinucleotide frequencies remains an open question.

In order to observe if differences in codon usage can be found among WNV strains isolated from different sources, the RSCU values for each codon were determined for strains isolated from birds, equines, humans and mosquitoes. No significant differences were detected in the RSCU values among any of the groups studied (see Supplementary Material Table 2). The same results are obtained using all WNV strains enrolled in these studies (see Supplementary Material Table 3).

4. Conclusions

The results of these studies revealed different codon preferences in WNV genes in relation to codon usage of human genes (see Table 2). We show that codon usage bias in this virus is relatively low. This is in agreement with previous results found from other RNA viruses such H5N1 Influenza A Virus (Ahn et al., 2006; Zhou et al., 2005); SARS (Zhao et al., 2008); FMDV (Zhong et al., 2007); classical swine fever virus (Tao et al., 2009); Duck Enteritis virus (Jia et al., 2009) or Theilovirus (Liu et al., 2011; Hu et al.,2011). The relative synonymous codon usage among WNV strains isolated from birds, equines, humans and mosquitoes are roughly similar and are influenced by the relative dinucleotide frequencies. These findings suggest that codon usage in WNV strains is dynamic, probably reflecting a process of mutation and selection. Interestingly, very recent results suggest that WNV evolution is characterized by alternative states of weak purifying selection in mosquitoes (which leads to an increase in genetic diversity) and strong purifying selection in birds (decreasing diversity) (Dearforff et al., 2011; Brackney et al., 2011). As all RNA viruses, WNV replicate as mutant distributions termed viral quasispecies (Domingo et al., 2001). Selection of particular variants (quasispecies) from the mutant spectrum has been previously demonstrated for WNV adaptation and replication (Ciota et al., 2007; Ciota et al., 2012; Jerzak et al., 2007) This is also in agreement with the results found in this work.

Acknowledgements

We acknowledge support by International Atomic Energy Agency, through Research Contract no. 15792. Authors acknowledge support by Agencia Nacional de Investigación e Innovación (ANII) through project PE_ALI_2009_1_1603 and PEDECIBA, Uruguay. We also thank the support of Fondo Clemente Estable, 2007_ 722 to HM. We thank anonymous reviewers for important insights into the improvement of the quality of this work.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.meegid.2013.01.001.

References

Ahn, I. et al., 2006. Genomic analysis of Influenza a viruses, including Avian Flu (H5N1) strains. Eur. J. Epidemiol. 21, 511–519.

Bisbal, C., Silverman, R.H., 2007. Diverse functions of RNase L and implications in pathology. Biochimie 89, 789–798.

Brackney, D.E. et al., 2011. West Nile virus genetic diversity is maintained during transmission by *Culex pipiens quinquefasciatus* mosquitoes. PLoS ONE 6, e24466.

Calisher, C.H. et al., 1989. Antigenic relationships between flaviviruses as determined by cross neutralization tests with polyclonal antisera. J. Gen. Virol. 70, 37.

Ciota, A.T. et al., 2007. Role of the mutant spectrum in adaptation and replication of West Nile virus. J. Gen. Virol. 88, 865–874.

Ciota, A.T. et al., 2012. Quantification of intrahost bottlenecks of West Nile virus in Culex pipiens mosquitoes using an artificial mutant swarm. Infect. Genet. Evol. 12 (3), 557–564.

Dearforff, E.R., Fitzpatrick, K.A., Jerzak, G.V.S., Shi, P.Y., Kramer, L.D., Ebel, G.D., 2011. West Nile virus experimental evolution in vivo and the trade-off hypothesis. PLoS Pathog. 7, e1002335.

Dauphin, G. et al., 2004. West Nile: wordwide current situation in animals and humans. Comp. Immunol. Microbiol. Infect. Dis. 27, 343–355.

Domingo, E. et al., 2001. Quasispecies and RNA virus evolution: principles and consequences. Asutin, Landes Bioscience.

Dorn, A., Kippenberger, S., 2008. Clinical application of CpG-, non-CpG, and antisense oligodeoxynucleotides as immunomodulators. Curr. Opin. Mol. Ther. 10, 10–20.

D'Andrea, L. et al., 2011. A detailed comparative analysis on the overall codon usage patterns in hepatitis A virus. Virus Res. 157, 19–24.

Gingold, H., Pilpel, Y., 2011. Determinants of translation efficiency and accuracy. Mol. Syst. Biol. 7, 481.

Greenacre, M., 1984. Theory and Applications of Correspondence Analysis. Academic Press, London.

Hamer, G.L. et al., 2009. Host selection by *Culex pipiens* mosquitoes and West Nile Virus amplification. Am. J. Trop. Med. Hyg. 80, 268–278.

Hu, J.S. et al., 2011. The characteristic of codon usage pattern and its evolution of hepatitis C virus. Infect. Genet. Evol. 11, 2098–2102.

Ikemura, T., 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer RNAs. J. Mol. Biol. 158, 73–597.

Liu, X.S. et al., 2012. Patterns and influencing factions of synonymous codon usage in porcine circovirus. Virol. J. 9, 68.

Liu, W.Q. et al., 2011. Compare the differences of synonymous codon usage between the two species within cardiovirus. Virology J. 8, 325.

Jerzak, G.V. et al., 2007. The West Nile virus mutant spectrum is host-dependant and a determinant of mortality in mice. Virology 360, 469–476.

Jia, R. et al., 2009. Analysis of synonymous codon usage in the UL24 gene of duck enteritis virus. Virus Genes 38, 96–103.

Komar, N. et al., 2003. Experimental infection of North American birds with the New York 1999 strain of West Nile virus. Emerg. Infect. Dis. 9, 311–322.

Komar, N., Clark, G.G., 2006. West Nile activity in Latin America and the Caribbean. Rev. Panam. Salud Publica 19, 112–117.

Supek, F., Vlahovicek, K., 2004. INCA: synonymous codon usage analysis and clustering by means of self-organizing map. Bioinformatics 20, 2329–2330.

Lim, S.M. et al., 2011. West Nile Virus: immunity and pathogenesis. Viruses 3, 811–828.

Lobo, F.P. et al., 2009. Virus-host coevolution: common patterns of nucleotide motif usage in *Flaviviridae* and their hosts. PLoS ONE 4, e6282.

May, F.J. et al., 2011. Phylogeography of West Nile Virus: from the cradle of evolution in Africa to Eurasia, Australia and the Americas. J. Virol. 85, 2964–2974.

Novembre, J.A., 2002. Accounting for background nucleotide composition when measuring codon usage bias. Mol. Biol. Evol. 19, 1390–1394.

Pandit, A., Sinha, S., 2011. Differential trends in the codon usage patterns in HIV-1 genes. PLoS One 6, 28889.

Pesko, K.N., Ebel, G.D., 2012. West Nile virus population genetics and evolution. Infect. Genet. Evol. 12, 181–190.

Player, M.R., Torrence, P.F., 1998. The 2–5A system: modulation of viral and cellular processes through acceleration of RNA degradation. Pharmacol. Ther. 78, 55–113.

Puigbo, P., Bravo, I.G., Garcia-Vallve, S., 2008a. CAIcal: a combined set of tools to assess codon usage adaptation. Biology Direct 3, e38.

Puigbo, P., Bravo, I.G., Garcia-Vallve, S., 2008b. E-CAI: a novel server to estimated an expected value of Codon Adaptation Index (eCAI). BMC Bioinformatics 9, e65.

Roosendaal, J. et al., 2006. Regulated cleavages at the West Nile virus NS4A-2K-NS4B junctionsplay a major role in rearranging cytoplasmic membranes and Golgi trafficking of the NS4A protein. J. Virol. 80, 4623–4632.

Scherbik, S.V. et al., 2006. RNase L plays a role in the antiviral response to West Nile virus. J. Virol 80, 2987–2999.

Shackelton, L.A., Parrish, C.R., Holmes, E.C., 2006. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. J. Mol. Evol. 62, 551–563.

Sharp, P.M., Li, W.H., 1987. The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15, 1281–1295.

Sharp, P.M., Li, W.H., 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24, 28–38.

Stoletzki, N., Eyre-Walker, A., 2007. Synonymous codon usage in Escherichia coli: selection for translational accuracy. Mol. Biol. Evol. 24, 374–381.

Tao, P. et al., 2009. Analysis of synonymous codon usage in classical swine fever virus. Virus Genes 38, 104–112.

Ulbert, S., 2011. West Nile Virus: the complex biology of an emerging pathogen. Intervirology 54, 171–184.

Wessa, P., 2012. Free Statistics Software, Office for Research Development and Education, version 1.1.23-r7. URL: http://www.wessa.net

Wong, E. et al., 2010. Codon usage bias and the evolution of Influenza A viruses. Codon usage biases of Influenza virus. Evol. Biol. 10, 253.

Woo, P.C.Y. et al., 2007. Cytosine deamination and selection of CpG suppressed clones are the two major independent biological forces that shape codon usage bias in Coronaviruses. Virology 369, 431–442.

Zhao, S. et al., 2008. Analysis of synonymous codon usage in 11 Human Bocavirus isolates. Biosystems 92, 207–214.

Zhong, J. et al., 2007. Mutation pressures shapes codon usage in the GC-Rich genome of foot-and-mouth disease virus. Virus Genes 35, 767–776.

Zhou, J.H. et al., 2012. Comparative the codon usage between the three main virases in pestivirus genus and susceptible livestock. Virus Genes 44, 475–481.

Zhou, T. et al., 2005. Analysis of synonymous codon usage in H5N1 virus and other Influenza A viruses. Biosystems 81, 77–86.

Supplementary Material Table 1. Origins of the WNV strains[a].

DDBJ	
Accession Number	Name
HM488208	WNV-1/US/BID-V4204/2002
HM488227	WNV-1/US/BID-V4608/2003
JF488091	WNV-1/US/BID-V5181/2004
JF730042	WNV-1/US/BID-V5147/2007
HQ671702	WNV-1/US/BID-V4595/2003
JN183891	WNV-1/US/BID-V4344/2002
HM488140	WNV-1/US/BID-V4210/2003
HM488184	WNV-1/US/BID-V4346/2002
JF920728	WNV-1/US/BID-V4568/2003
HM488118	WNV-1/US/BID-V4107/2005
HM488196	WNV-1/US/BID-V4376/2005
HM488206	WNV-1/US/BID-V4100/2008
HM488121	WNV-1/US/BID-V4110/2005
HM488145	WNV-1/US/BID-V4217/2004
HM488244	WNV-1/US/BID-V4634/2008
HQ671710	WNV-1/US/BID-V4902/2000
HM488143	WNV-1/US/BID-V4215/2004
JN183889	WNV-1/US/BID-V4579/2003
JF920753	WNV-1/US/BID-V5225/2008
HM488186	WNV-1/US/BID-V4350/2003
JF488087	WNV-1/US/BID-V5177/2004
HM488168	WNV-1/US/BID-V4364/2008
JF920757	WNV-1/US/BID-V5230/2008
HM488246	WNV-1/US/BID-V4689/2001
HM488204	WNV-1/US/BID-V4098/2008
JF920739	WNV-1/US/BID-V5209/2007
HM488147	WNV-1/US/BID-V4219/2004
JN183886	WNV-1/US/BID-V4629/2008
HM488149	WNV-1/US/BID-V4223/2005
JF920751	WNV-1/US/BID-V5223/2008
HM756677	WNV-1/US/BID-V4530/2005
HM488187	WNV-1/US/BID-V4351/2003
HM488129	WNV-1/US/BID-V4191/2000
HM488240	WNV-1/US/BID-V4627/2008
HM756649	WNV-1/US/BID-V4354/2006
HM756675	WNV-1/US/BID-V4806/2005
JF920734	WNV-1/US/BID-V5204/2006
HM488139	WNV-1/US/BID-V4208/2003
HQ671717	WNV-1/US/BID-V4910/2001
JF920736	WNV-1/US/BID-V5206/2006
HM756663	WNV-1/US/BID-V4697/2001
HM488179	WNV-1/US/BID-V4338/2002
HQ705659	WNV-1/US/BID-V4209/2003
HM488189	WNV-1/US/BID-V4367/2004
HM488200	WNV-1/US/BID-V4092/2007
HM488238	WNV-1/US/BID-V4623/2008
HM488177	WNV-1/US/BID-V4336/2002
JF920732	WNV-1/US/BID-V5202/2006
HQ671713	WNV-1/US/BID-V4905/2001
HM488161	WNV-1/US/BID-V4356/2007
JF920745	WNV-1/US/BID-V5216/2007
HM488137	WNV-1/US/BID-V4202/2002
HM488133	WNV-1/US/BID-V4195/2001
HM488163	WNV-1/US/BID-V4359/2007
JF920743	WNV-1/US/BID-V5214/2007
HM756661	WNV-1/US/BID-V4692/2001
HM488233	WNV-1/US/BID-V4616/2003
HM488157	WNV-1/US/BID-V4231/2006
JF920747	WNV-1/US/BID-V5218/2008
HM488115	WNV-1/US/BID-V4103/2005
HQ671722	WNV-1/US/BID-V4704/2002
HM488117	WNV-1/US/BID-V4105/2005
HQ671720	WNV-1/US/BID-V4913/2002
HM488135	WNV-1/US/BID-V4199/2001
JN367277	WNV-1/US/BID-V4803/2004
HM488175	WNV-1/US/BID-V4569/2003
JN183887	WNV-1/US/BID-V4706/2002
HQ671698	WNV-1/US/BID-V4203/2002
JF920749	WNV-1/US/BID-V5220/2008
HM488181	WNV-1/US/BID-V4340/2002
JF920741	WNV-1/US/BID-V5212/2007
HQ671696	WNV-1/US/BID-V4196/2001
HM756671	WNV-1/US/BID-V4798/2004
HQ671729	WNV-1/US/BID-V4892/2006
HM488199	WNV-1/US/BID-V4090/2007
HM756665	WNV-1/US/BID-V4709/2002
HM488212	WNV-1/US/BID-V4567/2003
HM488192	WNV-1/US/BID-V4371/2005
JF488094	WNV-1/US/BID-V5150/2004
HM756673	WNV-1/US/BID-V4801/2004
HM488225	WNV-1/US/BID-V4605/2003
JF920754	WNV-1/US/BID-V5226/2008
HM488210	WNV-1/US/BID-V4565/2003
HM488150	WNV-1/US/BID-V4224/2005
HQ671704	WNV-1/US/BID-V4618/2003
HM756652	WNV-1/US/BID-V4587/2003
HM756656	WNV-1/US/BID-V4615/2003
HM488252	WNV-1/US/BID-V4805/2005
JF920760	WNV-1/US/BID-V5235/2009
HM488127	WNV-1/US/BID-V4188/1999
HQ671727	WNV-1/US/BID-V4889/2006
HM488165	WNV-1/US/BID-V4361/2007
HM488173	WNV-1/US/BID-V4562/2003
HM756658	WNV-1/US/BID-V4686/2003
HM488194	WNV-1/US/BID-V4374/2005
HM488131	WNV-1/US/BID-V4193/2000
HM488242	WNV-1/US/BID-V4631/2008
HQ671708	WNV-1/US/BID-V4900/2000
HM488223	WNV-1/US/BID-V4603/2003
HM756669	WNV-1/US/BID-V4718/2003
HQ671708	WNV-1/US/BID-V4900/2000
HM756669	WNV-1/US/BID-V4718/2003
HM488141	WNV-1/US/BID-V4212/2003
HM488152	WNV-1/US/BID-V4226/2005
HQ671723	WNV-1/US/BID-V4715/2003
HM488146	WNV-1/US/BID-V4218/2004
HM488190	WNV-1/US/BID-V4368/2004
HM756667	WNV-1/US/BID-V4712/2003
HQ671706	WNV-1/US/BID-V4898/1999
JF488096	WNV-1/US/BID-V5159/2009
HM488248	WNV-1/US/BID-V4694/2001
HM488125	WNV-1/US/BID-V4186/1999
HM488250	WNV-1/US/BID-V4717/2003
HQ671725	WNV-1/US/BID-V4885/2005
HM488214	WNV-1/US/BID-V4572/2003
HM488221	WNV-1/US/BID-V4593/2003
HQ705669	WNV-1/US/BID-V4342/2002
HM488144	WNV-1/US/BID-V4216/2004
HM488183	WNV-1/US/BID-V4345/2002
JF488090	WNV-1/US/BID-V5180/2004
HQ671707	WNV-1/US/BID-V4899/1999
JF920738	WNV-1/US/BID-V5208/2007
HM488119	WNV-1/US/BID-V4108/2005
HQ671701	WNV-1/US/BID-V4590/2003
HM488226	WNV-1/US/BID-V4607/2003
HQ671705	WNV-1/US/BID-V4620/2003
HM488207	WNV-1/US/BID-V4101/2008
JF920758	WNV-1/US/BID-V5233/2009
JN183890	WNV-1/US/BID-V4699/2003
HM488216	WNV-1/US/BID-V4574/2003
HM488120	WNV-1/US/BID-V4109/2005
HQ671724	WNV-1/US/BID-V4883/2005
HM488136	WNV-1/US/BID-V4200/2001
HM488197	WNV-1/US/BID-V4377/2005
HQ671711	WNV-1/US/BID-V4903/2000
HM488148	WNV-1/US/BID-V4220/2004
HM488185	WNV-1/US/BID-V4347/2003
HM488247	WNV-1/US/BID-V4691/2001
JF488088	WNV-1/US/BID-V5178/2004
JF730043	WNV-1/US/BID-V5170/2002
HM488224	WNV-1/US/BID-V4604/2003
HQ705660	WNV-1/US/BID-V4714/2003
HQ671730	WNV-1/US/BID-V4897/2007
HM488167	WNV-1/US/BID-V4363/2008
HM488209	WNV-1/US/BID-V4564/2003
JF920750	WNV-1/US/BID-V5222/2008
HM488169	WNV-1/US/BID-V4365/2008
JF920756	WNV-1/US/BID-V5229/2008
JF920731	WNV-1/US/BID-V5201/2006
JN183885	WNV-1/US/BID-V4626/2008

HM756648	WNV-1/US/BID-V4205/2002	HM488124	WNV-1/CTFS/BID-V4113/2006
HM488205	WNV-1/US/BID-V4099/2008	HM488142	WNV-1/US/BID-V4214/2004
HM488170	WNV-1/US/BID-V4366/2008	HQ671699	WNV-1/US/BID-V4206/2002
JF920737	WNV-1/US/BID-V5207/2006	HM488245	WNV-1/US/BID-V4635/2008
HM488188	WNV-1/US/BID-V4353/2004	HM488153	WNV-1/US/BID-V4227/2005
HM488234	WNV-1/US/BID-V4617/2003	HM488191	WNV-1/US/BID-V4369/2004
HM488241	WNV-1/US/BID-V4628/2008	HM488130	WNV-1/US/BID-V4192/2000
JF920307	WNV-1/US/BID-V4907/2001	HM488198	WNV-1/US/BID-V4378/2005
HM488138	WNV-1/US/BID-V4207/2003	HQ671728	WNV-1/US/BID-V4891/2006
JF920748	WNV-1/US/BID-V5219/2008	JF920755	WNV-1/US/BID-V5227/2008
HM488134	WNV-1/US/BID-V4198/2001	JF488097	WNV-1/US/BID-V5148/2007
HM488237	WNV-1/US/BID-V4622/2008	HM756672	WNV-1/US/BID-V4799/2004
JF488086	WNV-1/US/BID-V5176/2004	HQ671726	WNV-1/US/BID-V4887/2005
HM756676	WNV-1/US/BID-V4349/2003	HM488222	WNV-1/US/BID-V4599/2003
HM488201	WNV-1/US/BID-V4093/2007	HM488243	WNV-1/US/BID-V4632/2008
HM488164	WNV-1/US/BID-V4360/2007	HM756668	WNV-1/US/BID-V4716/2003
HM756662	WNV-1/US/BID-V4693/2001	AY848695	9317A
JF920735	WNV-1/US/BID-V5205/2006	DQ066423	9317B
HM488203	WNV-1/US/BID-V4096/2008	AF260968	Eg101
HM488162	WNV-1/US/BID-V4357/2007	DQ666448	BSL5-2004
HM488232	WNV-1/US/BID-V4614/2003	AY765264	Rabensburg isolate 97-103
JF920744	WNV-1/US/BID-V5215/2007	AY603654	EthAn4766
HM756660	WNV-1/US/BID-V4097/2008	EF429197	SPU116/89
HM488253	WNV-1/US/BID-V4553/2006	JF415928	M20122
JN183892	WNV-1/US/BID-V4379/2005	JF415925	M38488
HM488154	WNV-1/US/BID-V4228/2005	FJ411043	NY99iso-1
HQ671718	WNV-1/US/BID-V4911/2001	DQ005530	FDA-BSL5-2003
JF920733	WNV-1/US/BID-V5203/2006	GQ379161	ArEq003
HM488132	WNV-1/US/BID-V4194/2000	GQ379160	ArEq001
HM488156	WNV-1/US/BID-V4230/2006	GQ379157	DB080718-14
HM488116	WNV-1/US/BID-V4104/2005	DQ164204	CO 2003 1
JF899528	WNV-1/US/BID-V4800/2004	DQ164202	OH 2002
HM488176	WNV-1/US/BID-V4575/2003	FJ151394	NY99-crow-V76/1
HM488155	WNV-1/US/BID-V4229/2006	AY848697	385-99
HM488213	WNV-1/US/BID-V4571/2003	JF703164	CA-03 IMPR116
HM488114	WNV-1/US/BID-V4102/2002	JF415930	M6019
HM488158	WNV-1/US/BID-V4232/2006	AY278442	LEIV-Vlg00-27924
HQ671719	WNV-1/US/BID-V4912/2001	EU081844	Egypt 101
HQ671714	WNV-1/US/BID-V4906/2001	GQ379159	JPW080813-01
HQ671712	WNV-1/US/BID-V4904/2000	JF415919	M19433
JF920742	WNV-1/US/BID-V5213/2007	JF703161	CA-04 COAV689
HM756678	WNV-1/US/BID-V4095/2007	EF429199	SA381/00
HM488219	WNV-1/US/BID-V4585/2003	DQ164206	TX 2004 Harris 4
HM488217	WNV-1/US/BID-V4581/2003	DQ164198	TX 2002 1
HM488178	WNV-1/US/BID-V4337/2002	DQ164199	TX2003
HM488228	WNV-1/US/BID-V4609/2003	JF415916	TX6276
JF488093	WNV-1/US/BID-V5188/2005	DQ176637	TX 2002-HC
JF488093	WNV-1/US/BID-V5188/2005	DQ164193	NY 2002 Clinton
HM488174	WNV-1/US/BID-V4563/2003	JF719069	Spain/2010/H-1b
HQ671709	WNV-1/US/BID-V4901/2000	AY289214	TVP 8533
HM756651	WNV-1/US/BID-V4584/2003	M12294	M12294
HQ671697	WNV-1/US/BID-V4197/2001	DQ256376	804994
HM488182	WNV-1/US/BID-V4341/2002	AF260969	RO97-50
JF920730	WNV-1/US/BID-V5197/2006	DQ164191	NY 2003 Chautauqua
HM488230	WNV-1/US/BID-V4612/2003	DQ164195	NY 2002 Nassau
HM488180	WNV-1/US/BID-V4339/2002	JF415923	M37906
HQ671721	WNV-1/US/BID-V4625/2008	EF530047	3356.2.1.1
JF488089	WNV-1/US/BID-V5179/2004	DQ164203	CO 2003 2
JF488095	WNV-1/US/BID-V5157/2009	JF703163	CA-05 COAV2900
HM488236	WNV-1/US/BID-V4700/2003	JF415927	M20141
JF920740	WNV-1/US/BID-V5210/2007	DQ164187	NY 2002 Broome
HM488249	WNV-1/US/BID-V4696/2001	DQ164189	NY 2003 Albany
HM488126	WNV-1/US/BID-V4187/1999	DQ666451	BSL13-2005
HM488220	WNV-1/US/BID-V4586/2003	AY277252	LEIV-Vlg99-27889
HQ671700	WNV-1/US/BID-V4576/2003	AF404754	MQ5488
HM756664	WNV-1/US/BID-V4701/2002	DQ318020	ArB3573/82
HM756653	WNV-1/US/BID-V4588/2003	JF415929	TX5058
HM756666	WNV-1/US/BID-V4711/2003	JF415921	TX7558
HM756659	WNV-1/US/BID-V4687/2003	JF415918	TX6747
HM488172	WNV-1/US/BID-V4561/2003	GQ379156	FL2001
HM488193	WNV-1/US/BID-V4373/2005	AF404756	WN NY 2000
HM488251	WNV-1/US/BID-V4719/2003	FJ159129	101_5-06-Uu
HM488128	WNV-1/US/BID-V4189/1999	FJ159130	5_50-05-Uu
JF920729	WNV-1/US/BID-V5196/2006	FJ159131	8_1-05Uu
HM756657	WNV-1/US/BID-V4685/2003	FJ483548	15217
HM488211	WNV-1/US/BID-V4566/2003	DQ164201	AZ 2004
HM488195	WNV-1/US/BID-V4375/2005	AF481864	IS-98 STD
HM488239	WNV-1/US/BID-V4624/2008	AY688948	Sarafend
HM756670	WNV-1/US/BID-V4720/2003	DQ411035	Ast02-2-692
HM488151	WNV-1/US/BID-V4225/2005	DQ411034	Ast02-2-691
JF972636	WNV-1/US/BID-V5228/2008	DQ411030	Ast01-182
HM488166	WNV-1/US/BID-V4362/2008	DQ411032	Ast02-3-146
JF920759	WNV-1/US/BID-V5234/2009	DQ377180	Ast02-3-208
HM488215	WNV-1/US/BID-V4573/2003	DQ374651	Ast02-3-570
HM488122	WNV-1/CTFS/BID-V4111/2006	DQ374650	Ast02-3-717
HM488123	WNV-1/CTFS/BID-V4112/2006	DQ374653	Ast02-2-25

DQ411031	Ast01-187
DQ374652	Ast04-2-824A
AY278441	Ast99-901
DQ377179	Ast02-2-298
AY842931	385-99
FJ483549	15803
DQ164200	IN 2002
JF415914	M12214
DQ164196	GA 2002 1
DQ164197	GA 2002 2
DQ211652	NY99
GQ379158	ORCO0559-07
EU249803	68856
AF206518	2741
DQ164190	NY 2003 Suffolk
HQ596519	4132
JF415926	M20140
DQ164188	NY 2003 Westchester
DQ666450	GCTX2-2005
JF415922	M37012
DQ164186	NY 2002 Queens
EU155484	OK03
JF719065	Italy/2008/J-242853
JN858070	Italy/2011/AN-2
JF719067	Italy/2009/G-223184
JN858069	Italy/2011/AN-1
JF719066	Italy/2008/M-203204
GU011992	Ita09
JQ928174	Italy/2011/Livenza
JQ928175	Italy/2011/Piave
DQ164192	NY 2003 Rockland
AY277251	LEIV-Krnd88-190
JF707789	HU6365/08
AM404308	PTRoxo
AY490240	Chin-01
AY795965	ARC10
EF571854	385-99
DQ666452	BSL2-2005
EF429198	SA93/01
JF415917	TX6647
DQ164205	TX 2002 2
AY646354	AY646354
JF703162	CA-03 COAV997
AJ965628	PT5.2
DQ176636	Madagascar-AnMg798
HM147822	South Africa
JF415924	TX7827
HM147824	Congo
AY848696	385-99 hamster passage
AY532665	B956
HM051416	twn9
EF429200	H442
AF404753	MD 2000-crow265
DQ116961	Hungary/04
JN183893	WNV-1/Gallus/BID-V4954/kidney
HQ671693	WNV-1/Gallus/BID-V4961/kidney
HQ705677	WNV-1/Gallus/BID-V4960/kidney
HQ671691	WNV-1/Gallus/BID-V4958/kidney
HQ705678	WNV-1/Gallus/BID-V4962/kidney
HQ671692	WNV-1/Gallus/BID-V4959/kidney
HQ671694	WNV-1/Gallus/BID-V4963/kidney
JN183894	WNV-1/Gallus/BID-V4955/kidney
HQ671695	WNV-1/Gallus/BID-V5112/skin
JF357960	WNV-1/Gallus/BID-V5109/skin
HQ705672	WNV-1/Culex/BID-V4168/legs
HQ705674	WNV-1/Culex/BID-V4173/midgut
HQ671731	WNV-1/Culex/BID-V4171/midgut
HQ671690	WNV-1/Culex/BID-V4180/midgut
HQ671688	WNV-1/Culex/BID-V4165/legs
JN183895	WNV-1/Culex/BID-V5768/midgut
JN183897	WNV-1/Culex/BID-V5808/midgut
HQ705671	WNV-1/Culex/BID-V4166/midgut
HQ671687	WNV-1/Culex/BID-V4161/legs
HQ705673	WNV-1/Culex/BID-V4169/midgut
HQ671689	WNV-1/Culex/BID-V4174/legs
HQ705670	WNV-1/Culex/BID-V4163/legs
JN183896	WNV-1/Culex/BID-V5776/midgut
AY712948	v4369
HQ537483	Nea Santa-Greece-2010
JN819311	WNV-1/BID-V5038
JN819312	WNV-1/BID-V5039
JN819313	WNV-1/BID-V5040
JX041632	Ig2266
HQ671686	WNV-1/Mus/BID-V5017/brain
HQ671682	WNV-1/Mus/BID-V4986/spleen
HQ671733	WNV-1/Mus/BID-V5004/brain
JF730040	WNV-1/Mus/BID-V4982/serum
HQ671669	WNV-1/Mus/BID-V4730/spleen
HQ671673	WNV-1/Mus/BID-V4924/spleen
JF899531	WNV-1/Mus/BID-V4979/serum
JF357958	WNV-1/Mus/BID-V4997/brain
JF899535	WNV-1/Mus/BID-V5142/brain
HQ671671	WNV-1/Mus/BID-V4920/spleen
HQ891013	WNV-1/Mus/BID-V5122/serum
HQ671677	WNV-1/Mus/BID-V4938/spleen
HQ671675	WNV-1/Mus/BID-V4926/spleen
HQ705663	WNV-1/Mus/BID-V4936/spleen
HQ671680	WNV-1/Mus/BID-V4981/serum
HQ671679	WNV-1/Mus/BID-V4969/serum
HQ671684	WNV-1/Mus/BID-V5007/brain
JF899537	WNV-1/Mus/BID-V5146/brain
JF899533	WNV-1/Mus/BID-V4996/brain
JF730041	WNV-1/Mus/BID-V4729/spleen
HQ671668	WNV-1/Mus/BID-V4728/spleen
JF357959	WNV-1/Mus/BID-V4999/brain
JF784158	WNV-1/Mus/BID-V4974/serum
HQ671732	WNV-1/Mus/BID-V5000/brain
HQ671670	WNV-1/Mus/BID-V4731/spleen
JF899530	WNV-1/Mus/BID-V4923/spleen
HQ671674	WNV-1/Mus/BID-V4925/spleen
JF899532	WNV-1/Mus/BID-V4992/brain
HQ671672	WNV-1/Mus/BID-V4921/spleen
HQ705676	WNV-1/Mus/BID-V4971/serum
HQ891010	WNV-1/Mus/BID-V4987/spleen
HQ671678	WNV-1/Mus/BID-V4940/spleen
HQ891012	WNV-1/Mus/BID-V5118/serum
HQ671676	WNV-1/Mus/BID-V4937/spleen
HQ671681	WNV-1/Mus/BID-V4985/spleen
HQ671683	WNV-1/Mus/BID-V5006/brain
HQ671685	WNV-1/Mus/BID-V5008/brain
JF899536	WNV-1/Mus/BID-V5145/brain
JF899534	WNV-1/Mus/BID-V5002/brain
AF260967	NY99-eqhs
AF404757	Italy 1998-equine
DQ118127	goose-Hungary/03
AF404755	NY 2000-grouse3282
JX041630	LEIV-1640Az
JX041631	LEIV-3266Ukr
AY712945	Bird 1153
AY712946	Bird 1171
AY712947	Bird 1461
JN819315	WNV-1/BID-V5042
JN819318	WNV-1/BID-V5045
JN819319	WNV-1/BID-V5047
JN819320	WNV-1/BID-V5048
JX041629	LEIV-1628Az

[a] Strains isolated from birds, equines, humans and mosquitoes enrolled in the RSCU studies by group of isolation are highlighted in green, blue, red and yellow, respectively.

Supplementary Material Table 2. Synonymous codon usage bias in WNV^{a}.

		WNV isolated from:			
Amino acid	**Codon^{b}**	**Birds**	**Equine**	**Human**	**Mosquitoes**
Phe	UUU	0.91	0.89	0.90	0.90
	UUC	1.09	1.11	1.10	1.10
Leu	UUA	0.20	0.20	0.19	0.22
	UUG	1.43	1.47	1.47	1.46
	CUU	0.69	0.73	0.72	0.77
	CUC	1.20	1.16	1.17	1.20
	CUA	0.66	0.64	0.64	0.71
	CUG	1.82	1.80	1.80	1.64
Ile	AUU	1.09	1.07	1.09	1.06
	AUC	1.30	1.27	1.30	1.27
	AUA	0.61	0.66	0.61	0.67
Met	AUG	1.00	1.00	1.00	1.00
Val	GUU	0.77	0.74	0.77	0.79
	GUC	1.07	1.09	1.08	1.03
	GUA	0.31	0.33	0.33	0.35
	GUG	1.84	1.84	1.83	1.83
Tyr	UAU	0.73	0.76	0.74	0.65
	UAC	1.27	1.24	1.26	1.35
TER	UAA	**	**	**	**
	UAG	**	**	**	**

Synonymous codon usage bias in WNV[a] (Cont.).

		WNV isolated from:			
Amino acid	**Codon[b]**	**Birds**	**Equine**	**Human**	**Mosquitoes**
His	CAU	0.80	0.79	0.82	0.78
	CAC	1.20	1.21	1.18	1.22
Gln	CAA	0.87	0.89	0.87	0.94
	CAG	1.13	1.11	1.13	1.06
Asn	AAU	0.73	0.76	0.75	0.73
	AAC	1.27	1.24	1.25	1.27
Lys	AAA	0.85	0.85	0.84	0.85
	AAG	1.15	1.15	1.16	1.15
Asp	GAU	0.95	0.94	0.95	0.91
	GAC	1.05	1.06	1.05	1.09
Glu	GAA	1.03	1.02	1.02	1.00
Ser	UCU	0.70	0.66	0.66	0.64
	UCC	0.67	0.66	0.66	0.80
	UCA	1.56	1.58	1.60	1.52
	UCG	0.56	0.56	0.55	0.62
Pro	CCU	0.87	0.85	0.88	0.86
	CCC	0.90	0.91	0.89	1.02
	CCA	1.79	1.78	1.80	1.59
	CCG	0.44	0.46	0.43	0.53

Synonymous codon usage bias in WNV[a] (Cont.).

		WNV isolated from:			
Amino acid	**Codon[b]**	**Birds**	**Equine**	**Human**	**Mosquitoes**
Thr	ACU	0.84	0.86	0.84	0.88
	ACC	1.28	1.25	1.26	1.23
	ACA	1.24	1.26	1.25	1.11
	ACG	0.64	0.63	0.65	0.78
Ala	GCU	1.34	1.32	1.32	1.17
	GCC	1.16	1.18	1.18	1.22
	GCA	0.98	1.00	1.00	0.95
	GCG	0.52	0.50	0.50	0.67
Cys	UGU	0.86	0.90	0.89	0.90
	UGC	1.14	1.10	1.11	1.10
TER	UGA	**	**	**	**
Trp	UGG	1.00	1.00	1.00	1.00
Arg	**CGU**	0.56	0.52	0.55	0.55
	CGC	0.84	0.87	0.85	0.88
	CGA	0.52	0.50	0.50	0.55
	CGG	0.73	0.75	0.74	0.79
Ser	AGU	1.00	1.10	1.11	1.10
	AGC	1.40	1.40	1.40	1.37
Arg	AGA	2.00	1.99	2.01	1.96
	AGG	1.34	1.37	1.35	1.26

Synonymous codon usage bias in WNV[a] (Cont.).

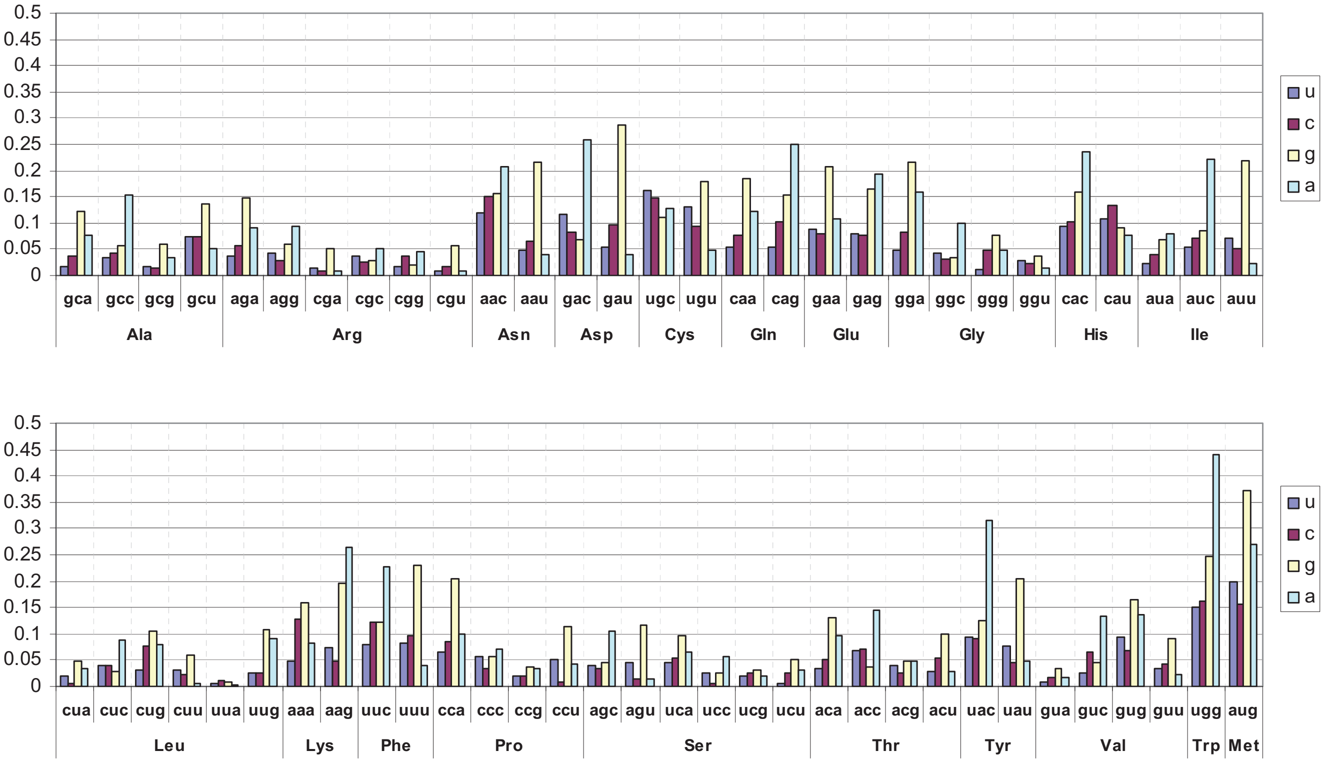
		WNV isolated from:			
Amino acid	**Codon[b]**	**Birds**	**Equine**	**Human**	**Mosquitoes**
Gly	GGU	0.43	0.43	0.43	0.48
	GGC	0.82	0.81	0.81	0.81
	GGA	2.01	2.02	2.03	1.90
	GGG	0.74	0.74	0.74	0.81

[a] For names and accession numbers of WNV strains enrolled in these studies see Supplementary Material Table 1. [b] UpA containing codons are shown underlined. CpG containing codons are shown in bold.

Supplementary Material Table 3. Global codon usage West Nile virus, displayed as RSCU[a] values.

AA	Cod	WNV	AA	Cod	WNV	AA	Cod	WNV	AA	Cod	WNV
Phe	UUU	0.90	Ser	UCU	0.67	Tyr	UAU	0.75	Cys	UGU	0.90
	UUC	1.10		UCC	0.67		UAC	1.25		UGC	1.10
Leu	UUA	0.18		UCA	1.57	TER	UAA	**	TER	UGA	**
	UUG	1.50		**UCG**	0.57		UAG	**	Trp	UGG	1.00
	CUU	0.71	Pro	CCU	0.86	His	CAU	0.82	Arg	**CGU**	0.56
	CUC	1.18		CCC	0.88		CAC	1.18		**CGC**	0.84
	CUA	0.65		CCA	1.82	Gln	CAA	0.88		**CGA**	0.51
	CUG	1.77		**CCG**	0.45		CAG	1.12		**CGG**	0.73
Ile	AUU	1.09	Thr	ACU	0.83	Asn	AAU	0.74	Ser	AGU	1.15
	AUC	1.29		ACC	1.28		AAC	1.26		AGC	1.37
	AUA	0.63		ACA	1.25	Lys	AAA	0.83	Arg	AGA	2.01
Met	AUG	1.00		**ACG**	0.64		AAG	1.17		AGG	1.36
Val	GUU	0.76	Ala	GCU	1.33	Asp	GAU	0.95	Gly	GGU	0.42
	GUC	1.08		GCC	1.16		GAC	1.05		GGC	0.83
	GUA	0.31		GCA	1.00	Glu	GAA	0.97		GGA	2.02
	GUG	1.85		**GCG**	0.51		GAG	1.03		GGG	0.73

[a]RSCU, relative synonymous codon usage; AA, amino acid; Cod, codons; WNV, West Nile virus. UpA containing codons are shown underlined. Codons containing CpG are shown in bold.



Supplementary Figure 1. Histogram of the relative frequency of synonymous codons according to the first base in the next codon (U, C, G, A). Note that the relative frequency is estimated within each codon family.

7 Capítulo IV: Reconstrucción del sesgo selectivo en el Uso de Aminoácidos en Mollicutes

Trends in amino acid usage across the class Mollicutes

Iriarte A, Baraibar JD, Diana L, Castro-Sowinski S, Romero H, Musto H.

Journal of Biomolecular Structure and Dynamics 2013, Jan 9.

Resumen:

Muchos estudios han explorado los mecanismos implicados en el uso relativo de aminoácidos (RAAU) en una gran variedad de procariotas y eucariotas. Se ha descrito un sesgo fuerte en los genes altamente expresados (HEGs) de bacterias endosimbióticas. Por medio del análisis de correspondencias, se estudiaron las principales tendencias que afectan a la variabilidad interna en el RAAU en Mollicutes. El factor principal estría relacionado con el nivel de expresión, en donde los HEGs tenderían a usar aminoácidos más pequeños, no aromáticos y codificados por codones ricos en G + C. Dada la naturaleza del código genético, estas propiedades están vinculadas entre sí. Como era de esperar, se encontró también que los HEGs presentan una tasa de evolución más lenta que el resto de los genes. Esto sería consecuencia del efecto de la selección purificadora. Por otro lado, el resto de los genes acumularían cambios rápidamente como resultado del sesgo mutacional extremo hacia A + T y la deriva genética, lo que, como consecuencia, generaría un aumento de la variabilidad interna en el RAAU en endosimbiontes. Se mapearon los cambios de reemplazo en toda la filogenia del grupo con el fin de estimar las tendencias del peso molecular medio y la aromaticidad. Por último, se comparó el RAAU dentro de Mollicutes, y entre Mollicutes y Firmicutes de vida libre.

Journal of Biomolecular Structure and Dynamics, 2013
http://dx.doi.org/10.1080/07391102.2012.748636

Trends in amino acid usage across the class Mollicutes

Andrés Iriarte[a,b,c], Juan Diego Baraibar[a], Leticia Diana[d], Susana Castro-Sowinski[d], Héctor Romero[a,e] and Héctor Musto[a]*

[a]*Laboratorio de Organización y Evolución del Genoma, Facultad de Ciencias, UDELAR. Iguá 4225, Montevideo, 11400 Uruguay;* [b]*Laboratorio de Evolución, Facultad de Ciencias, UDELAR. Iguá 4225, Montevideo, 11400 Uruguay;* [c]*Área Genética, Depto. de Genética y Mejora Animal, Facultad de Veterinaria, UDELAR. Av. A. Lasplaces y1550, Montevideo, CP 11600, Uruguay;* [d]*Sección Bioquímica, Facultad de Ciencias, UDELAR. Iguá 4225, Montevideo, 11400 Uruguay;* [e]*Centro Universitario Regional Este (C.U.R. E.), UDELAR, Maldonado, Uruguay*

Communicated by Ramaswamy H. Sarma

(*Received 3 November 2012; final version received 8 November 2012*)

Many studies have explored the mechanisms involved in relative amino acid usage (RAAU) in a variety of prokaryotes and eukaryotes. A strong bias was observed in highly expressed genes (HEGs) of endosymbiotic bacteria. By means of correspondence analysis, we studied the major trends affecting internal variability of RAAU in Mollicutes. The principal trend is related to the usage of smaller, less aromatic, and GC-rich coded amino acids in HEGs. Given the nature of the genetic code, these properties are linked among them. Expectedly, we found a slow evolutionary rate of HEGs, which is likely driven by purifying selection. On the other hand, the rest of the genes accumulate rapid changes as a result of the extreme mutational bias toward A + T of the genomes and genetic drift, increasing internal variability. Amino acid changes across the phylogeny of the group were traced in order to estimate the mean molecular weight and aromaticity trends in each branch. Finally, we compared amino acid usage bias within and between Mollicutes and the free-living Firmicutes.

Keywords: *Mycoplasma*; complete genome; amino acid usage; highly expressed genes; evolution

Introduction

Understanding the main factors affecting the relative amino acid usage (RAAU) in different genomes is one of the major goals in molecular evolution, and is usually studied using multivariate approaches like correspondence analysis (COA). In the first report of this kind, (Lobry & Gautier, 1994) found that in *Escherichia coli,* the three most important factors regarding RAAU were hydrophobicity, expressivity, and aromaticity of the proteins. They proposed that translational constraints, which are stronger in highly expressed genes (HEGs), might affect the global amino acid composition. Using the same approach, Garat and Musto (2000) found that in Giardia lamblia, the most relevant factor in RAAU is related to the mechanisms of defense against reactive oxygen species. Further, these authors found a clear tendency of HEGs to use smaller amino acids, so the cell economy seemed to be another prominent factor. From then on, many authors have explored the possible effect of selective pressures over RAAU in a variety of prokaryotes and eukaryotes (e.g. (Akashi & Gojobori, 2002; Banerjee & Ghosh, 2006; Basak, Banerjee, Gupta, & Ghosh, 2004; McDonald, 2010; Naya, Zavala, Romero, Rodriguez-Maseda, & Musto, 2004; Palacios & Wernegreen, 2002; Tekaia & Yeramian, 2006; Tekaia, Yeramian, & Dujon, 2002; Sabbia et al., 2007; Swire, 2007; Wang & Lercher, 2010; Zavala, Naya, Romero, & Musto, 2002).

Considering the particular case of endosymbiotic bacteria (Das, Paul, Chatterjee, & Dutta, 2005; Herbeck, Wall, & Wernegreen, 2003; Palacios & Wernegreen 2002; Rispe, Delmotte, van Ham, & Moya, 2004; Schaber et al., 2005), some authors postulated that RAAU bias in HEGs is driven by two factors: the avoidance of aromatic residues and the resistance to an enrichment in A+T (Rispe et al., 2004). According to this view, greater conservation in HEGs since the divergence of the group from the last free-living common ancestor may explain the observed resistance to A+T enrichment (Banerjee, Basak, Gupta, & Ghosh, 2004; Banerjee & Ghosh, 2006; Herbeck, Wall, & Wernegreen, 2003; Rispe et al., 2004).

On the other hand, it has been argued that G+C-rich amino acids are favored by natural selection (Schaber

*Corresponding author. Email: hmusto@fcien.edu.uy et al., 2005). In the case of Mollicutes, Yamao, Andachi, Muto, Ikemura, and Osawa, 1991 suggested that the amount of tRNAs is affected by the corresponding amino acid usage in ribosomal proteins. Under this scenario, neutral evolution cannot explain the observed RAAU bias.

Protein evolutionary rates are usually quantified by the estimation of the number of nonsynonymous nucleotide changes per nonsynonymous site (dN), and expression levels might be its best indicator (Duret & Mouchiroud, 2000; Pal, Papp, & Hurst, 2001, 2003; Rocha & Danchin, 2004; Subramanian & Kumar, 2004). This is reinforced by the fact that in several species, HEGs are systematically more conserved in comparison to lowly expressed genes (LEGs). The influence of this phenomenon on intragenomic RAAU variability should be particularly evident for organisms that experience substitutions due to selective or neutral forces, the latter expected to be less effective in HEGs.

Mollicutes, as well as other endosymbiotic bacteria, go through recursive bottlenecks. They typically have reduced genomes, accelerated evolutionary rates and a strong A+T mutational bias, mainly reflecting an increased effect of genetic drift and mutational pressure. Biosynthetic capabilities are reduced in these organisms so several nutrients must be acquired from their hosts (Razin, Yogev, & Naot, 1998). Based on phylogenies inferred either from 16S rRNA and/or conserved protein coding genes, Mollicutes are known to be a branch of the Firmicutes, as well as Bacilli and Clostridia (Woese, Maniloff, & Zablen, 1980; Wolf, Muller, Dandekar, & Pollack, 2004). Although the taxonomy of the group is still a matter of controversy (Boone, Castenholz, & Garrity, 2001), the phylogeny of Mollicutes shows two main groups: the AAA branch (Asteroleplasma, Anaeroplasma, Phytplasma, and Acholeplasma) and the SEM branch (Spiroplasma, Entomoplasma, Mesoplasma, Mycoplasma, and Ureaplasma) (Razin, Yogev, & Naot, 1998). Considering only the complete sequenced Mollicutes, studies usually deal with four monophyletic clusters (Oshima & Nishida, 2007): hominis, pneumoniae, and spiroplasma which belong to the SEM branch while phytoplasma-like cluster concerns all organisms in the AAA branch.

At the moment of writing this manuscript, there are enough complete sequenced genomes of Mollicutes and, as far as we know, no study has focused exclusively on RAAU in this class under a comparative and phylogenetic framework. On the other hand, many articles discussed the influence of expression level on RAAU in organisms with similar population dynamics and genomic features.

In this study, our focus is to analyze factors shaping amino acid composition of proteins within Mollicutes taking into consideration some key properties of genes (like expression levels and GC content) and features of amino acids (like molecular weight, aromaticity, and hydrophobicity). We consider and discuss the possible role of natural selection, mutational bias and genetic drift taking an intragenomic and intergenomic perspective. We also make inference of aromaticity and mean molecular weight (MMW) trends based on orthologous amino acid changes traced among the class.

Materials and methods

Sequences

Complete genome and coding sequences were downloaded from NCBI ftp site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria). Finally, 19 complete Mollicutes and 14 other Firmicutes (outgroup) were studied (Table 1).

Phylogenetic reconstruction

The reconstruction procedure is fully described elsewhere (Iriarte, Baraibar, Romero, & Musto, 2011). The tree was rooted between Clostridia and the rest of the Firmicutes based on the interactive tree of life online tool (http://itol.embl.de/).

Amino acid usage analyses

The major sources of variation among proteins were studied through COA (Greenacre, 1984) on RAAU. Base composition, amino acid usage, gravy score, aromaticity, and COA were calculated with CodonW 1.4.2 (written by John Peden and available from http://sourceforge.net/projects/codonw/).

RAAU bias (β) in HEGs was calculated in relation to the total proteome, for each residue in each species, according to the following equation:

$$\beta_i = \frac{Fr_i^{(\mathrm{HEGs})} - Fr_i^{(\mathrm{total})}}{Fr_i^{(\mathrm{total})}} \tag{1}$$

where, Fr_i is the relative frequency of amino acid *i* calculated for HEGs and for the rest of the genes in each genome, $\mathrm{Fri}^{(\mathrm{HEGs})}$ and $\mathrm{Fri}^{(\mathrm{total})}$, respectively. HEGs were defined in two different ways: (1) genes encoding ribosomal proteins + elongation factors, $\beta_{(\mathrm{REF})}$ and (2) as the top 10% of the HEGs identified by COA-RAAU for Mollicutes, $\beta_{(\mathrm{COA})}$.

A hierarchical cluster analysis of studied species was done based on the estimated $\beta_{(\mathrm{REF})}$ following the method developed by Ward. The analysis was done by means of the "pvclust" package (Suzuki & Shimodaira, 2006) from R Development Core Team (2012). Pvclust assesses the uncertainty in hierarchical cluster analysis using bootstrap resampling techniques ($n = 1000$) and provides two types of p-values: Approximately Unbiased and Bootstrap Probability.

Table 1. Summary of main results for each species analyzed

Organisms*	Axis[a]	MMW (Ref. Set)[b]	MMW (All genes)[c]	Composition (%) GC	GC3s[d]	CU[e]	Acc. N°
Acholeplasma laidlawii [Pl]	2 $^{(13.6)}$	130.0	131.5	32	19	×	NC_010163
Aster yellows phytoplasma [Pl]	1 $^{(18.7)}$	131.2	134.2	28	18		NC_007716
C. Phytoplasma australiense [Pl]	2 $^{(16.3)}$	131.4	134.2	28	19		NC_010544
C. Phytoplasma mali [Pl]	1 $^{(21.8)}$	132.4	134.6	23	9		NC_011047
Mesoplasma florum [S]	1 $^{(21.9)}$	130.3	132.1	27	11	×	NC_006055
Mycoplasma hyopneumoniae [H]	1 $^{(19.7)}$	131.3	133.4	29	18		NC_006360
Mycoplasma agalactiae [H]	1 $^{(18.8)}$	129.9	132.2	30	19	×	NC_009497
Mycoplasma arthritidis [H]	2 $^{(14.7)}$	130.6	132.4	31	21	×	NC_011025
Mycoplasma capricolum [S]	1 $^{(22.5)}$	130.4	133.0	24	7	×	NC_007633
Mycoplasma gallisepticum [Pn]	3 $^{(11.5)}$	130.7	132.0	32	23		NC_004829
Mycoplasma genitalium [Pn]	2 $^{(14.6)}$	131.5	132.2	32	22		NC_000908
Mycoplasma mobile [H]	2 $^{(18.3)}$	131.4	132.6	25	10		NC_006908
Mycoplasma mycoide [S]	1 $^{(22.3)}$	130.2	133.4	24	8	×	NC 005364
Mycoplasma penetrans [Pn]	2 $^{(14.9)}$	130.5	132.2	26	12	×	NC_004432
Mycoplasma pneumoniae [Pn]	3 $^{(11.1)}$	131.2	131.3	40	40		NC_000912
Mycoplasma pulmonis [H]	1 $^{(22.5)}$	131.2	133.4	27	14		NC_002771
Mycoplasma synoviae [H]	1 $^{(25.6)}$	130.6	132.3	29	15	×	NC_007294
Onion yelows phytoplasma [Pl]	1 $^{(20.0)}$	131.1	133.8	29	18		NC_005303
Ureaplasma urealyticum [Pn]	1 $^{(22.5)}$	130.3	133.4	26	11		NC_002162
Bacillus anthracis	–	130.1	131.1	36	23	×	NC_003997
Bacillus halodurans	–	130.1	130.6	44	40	×	NC_002570
Enterococcus faecalis	–	130.0	130.6	37	28	×	NC_017316
Lactobacillus johnsonii	–	129.9	130.8	35	22	×	NC_013504
Lactobacillus plantarum	–	129.3	129.2	45	43	×	NC_014554
Lactococcus lactis	–	129.1	130.6	36	23	×	NC_013656
Listeria innocua	–	130.0	130.2	38	26		NC_003212
Oceanobacillus iheyensis	–	130.4	130.7	36	23	×	NC_004193
Staphylococcus aureus	–	130.4	131.4	33	20	×	NC_013450
Staphylococcus epidermidis	–	130.5	131.6	32	19	×	NC_002976
Streptococcus mutans	–	129.2	130.7	37	27	×	NC_013928
Streptococcus pyogenes	–	129.0	130.1	39	29	×	NC_002737
Clostridium perfringens	–	130.2	131.0	29	14	×	NC_003366
Clostridium tetani	–	129.1	131.3	29	14	×	NC 004557

[a] Represents which axis generated by Correspondence Analysis (COA) on relative amino acid usage is related with expression. In brackets is shown the variability explained by the axis.

[b] Mean molecular weight of the protein coded by the reference set of highly expressed genes.

[c] Mean molecular weight of the protein coded by all genes.

[d] Guanine plus Cytosine content at third synonymous codon positions.

[e] Indicates strong or moderately selected codon usage bias is derived from (Sharp et al., 2005) and (Iriarte et al., 2011).

*Letters in brackets represent monophyletic groups within Mollicutes: [Pl] Phytoplasma-like, [S] Spiroplasma, [Pn] Pneumoniae and [H] Hominis.

Estimation of expression levels

For each gene, we computed MELP values as an expression-level predictor (Supek & Vlahovicek, 2005) using INCA2.1 (Supek & Vlahovicek, 2004). For a description of how it was used in this study, see Supek and Vlahovicek (2005). MELP is valid as an expression-level predictor for organisms in which natural selection for translation has been shown to shape the codon usage: eight Mollicutes (Iriarte et al., 2011; Yamao et al., 1991) and the outgroup (Musto, Romero, & Zavala, 2003; Sharp, Bailes, Grocock, Peden, & Sockett, 2005). Recently, Supek, Skunca, Repar, Vlahovicek, & Smuc, 2010 proved that translational selection in prokaryotes is practically universal. Therefore, the clustering of HEGs (including the reference set and other HEGs) at one extreme of the distribution of an axis generated by COA and the correlation of the position of the sequences along this axis with the respective MELP values were taken as evidence for the existence of a trend associated to expression level.

Energetic cost

As an index of relative cost, we calculated the MMW for each protein, and was calculated as in Zavala et al. (2002).

Trace of amino acid changes and molecular distances

Eighty-five putative orthologous sequences were identified among Mollicutes, as previously described in Iriarte et al., 2011. Protein sequences were aligned using

ClustalW (Thompson, Higgins, & Gibson, 1994). Maximum likelihood method, using REV (nr = 189) model was used to infer amino acid changes. Trace, reconstruction and global dN divergence for orthologs were estimated by means of Codeml program, included in PAML 4.0 (Yang, 2007) package.

Putative ortholog proteins were identified as in (Iriarte et al., 2011). Pairwise nonsynonymous (dN) distances among hominis, pneumoniae, spiroplasma, and "phytoplasma-like" groups were also calculated using Codeml (see Table 1).

Trends in molecular weight (ΔMW) were estimated as the differences in MW of the involved amino acids inferred to have changed in the phylogeny. Protein average molecular weight changes were calculated in each branch for orthologous sequences, 23 HEGs, and 62 LEGs (Supplementary Table 1). In order to estimate trends in aromaticity among these proteins, the number of changes to (positive) and from (negative) Phe, Tyr, and Trp were calculated. Finally, GC content variation associated with each an amino acid change (ΔGC) was estimated as the difference in the average GC content between triplet coding for the involved residues.

Results and discussion

The phylogenetic tree was in agreement with previous reconstructions for this group (Oshima & Nishida, 2007; Woese et al., 1980; Wolf et al., 2004). The bootstrap values displayed in Supplementary Figure 1 are high in almost all nodes. Therefore, amino acid changes and global dN estimations were traced on it, since we assumed that this tree reflects correctly the real relationships among the analyzed species.

Correspondence analysis on RAAU

This analysis showed that *M. synoviae*, *M. agalactiae*, *M. pulmonis*, *M. hyopneumoniae*, *Ureaplasma parvum*, *Mesoplasma florum*, *M. capricolum*, *M. mycoides*, *Candidatus Phytoplasma mali*, *Onion yellows phytoplasma*, and *Aster yellows phytoplasma* present a common pattern. Indeed, the first trend (which, depending on the species, accounted between 18.7 and 25.6% of the total variability) was related to expression levels, as shown by the fact that HEGs tended to cluster at one side of the distribution (Supplementary Figure 2). Furthermore, MMW and GC content at the first two codon positions [$GQ_{(1-2)}$] were linked with this principal trend (Table 2). With the exception of *M. synoviae*, the second axis (between 14.1 and 16.9% of the variability) was correlated with the aromaticity and hidropathy level of each protein. As hydrophobic and aromatic amino acids are energetically expensive, it was not surprising to find that this trend was often also related with MMW. In *M. synoviae*, the second axis was associated with aromaticity

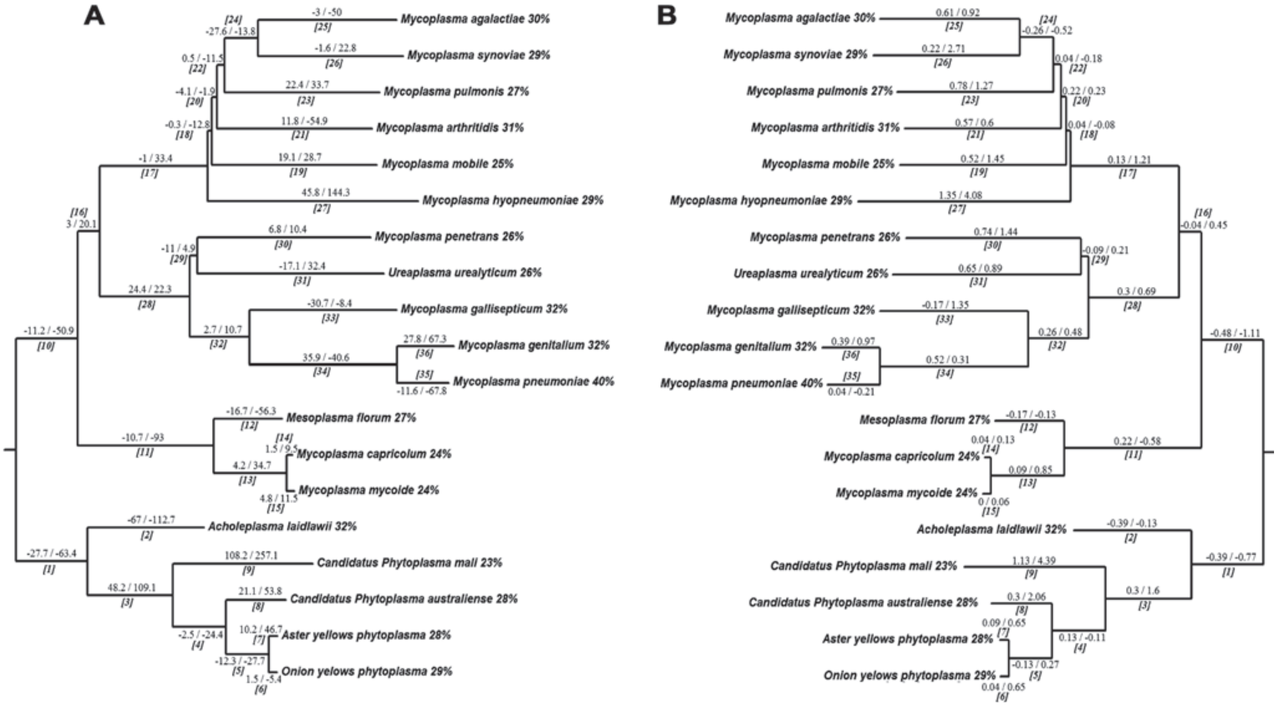
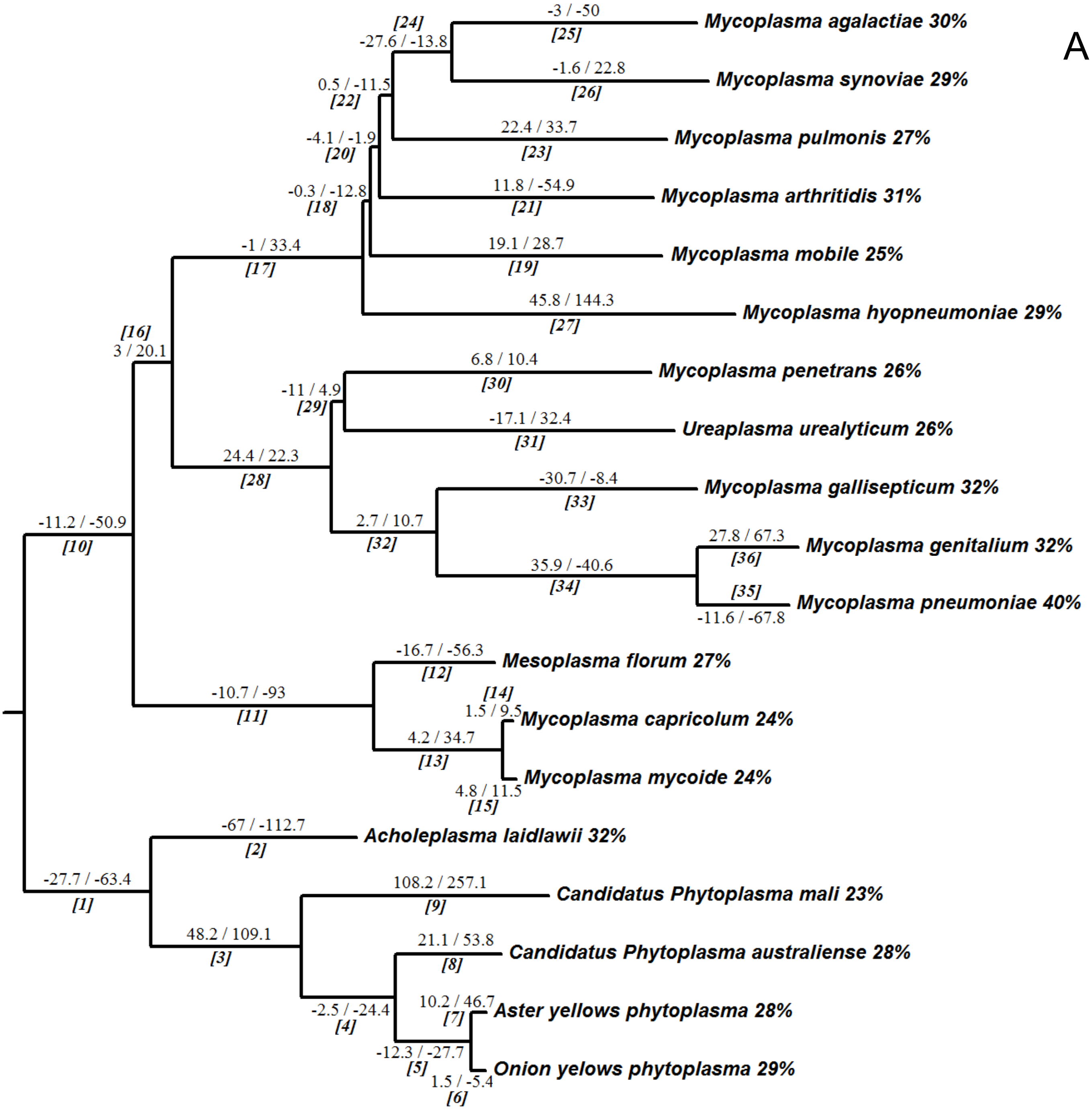
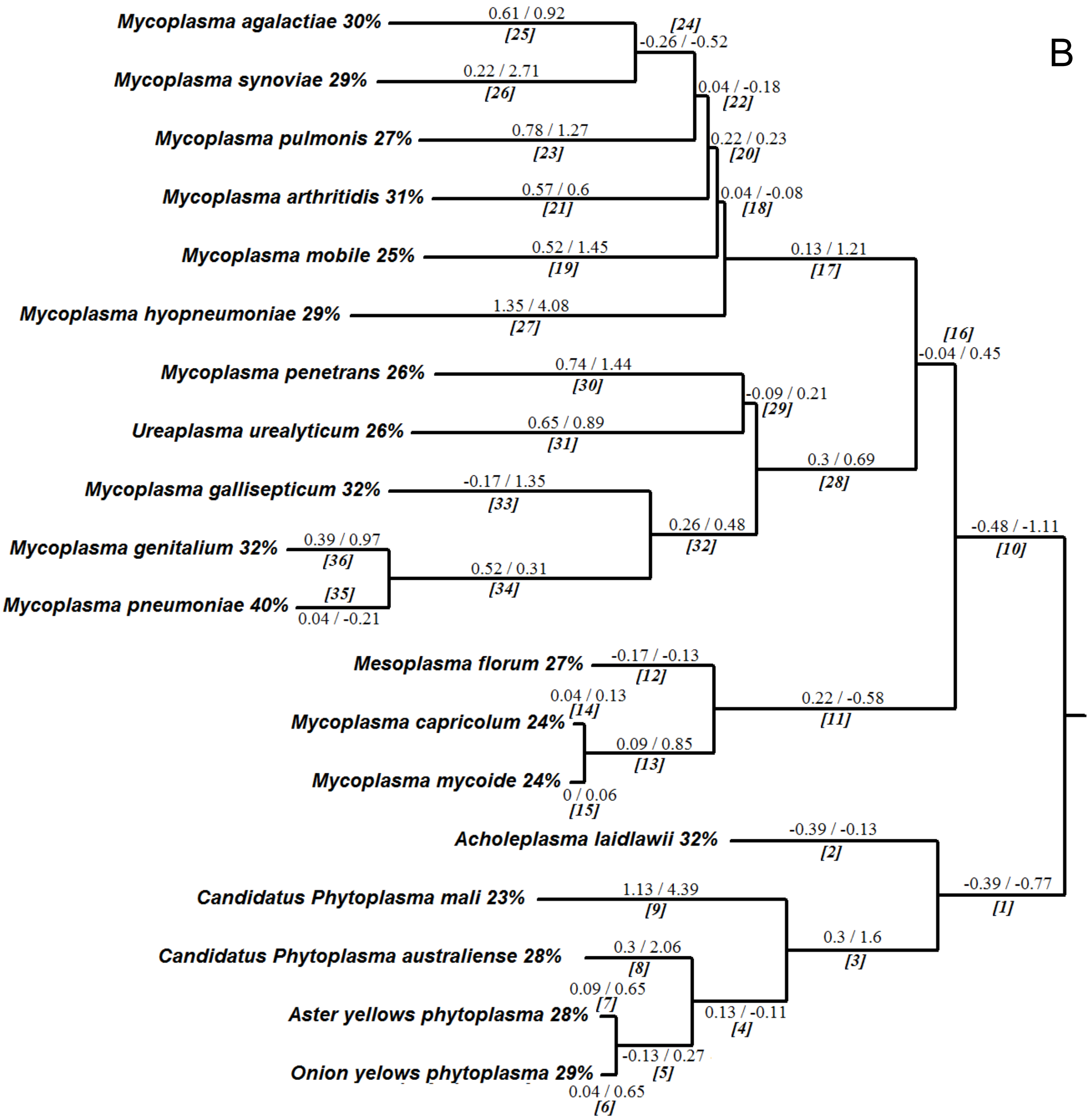


Figure 1. Trends in MW (A) and Aromaticity and (B) traced on the Mollicutes phylogenetic tree. Trends in MW were estimated for each branch as the differences in MW of the involved amino acids inferred to have changed. Trends in aromaticity were estimated as the number of changes to (positive) and from (negative) Phe, Tyr, and Trp. Averages are shown in each branch for orthologous sequences, 23 HEGs (left) and 62 LEGs (right) (Supplementary Table 1).


A
Mycoplasma agalactiae 30%
-3 / -50
[25]
[24]
-27.6 / -13.8
Mycoplasma synoviae 29%
-1.6 / 22.8
[26]
0.5 / -11.5
[22]
Mycoplasma pulmonis 27%
22.4 / 33.7
[23]
-4.1 / -1.9
[20]
Mycoplasma arthritidis 31%
11.8 / -54.9
[21]
-0.3 / -12.8
[18]
Mycoplasma mobile 25%
19.1 / 28.7
[19]
-1 / 33.4
[17]
Mycoplasma hyopneumoniae 29%
45.8 / 144.3
[27]
[16]
3 / 20.1
Mycoplasma penetrans 26%
6.8 / 10.4
[30]
-11 / 4.9
[29]
Ureaplasma urealyticum 26%
-17.1 / 32.4
[31]
24.4 / 22.3
[28]
Mycoplasma gallisepticum 32%
-30.7 / -8.4
[33]
2.7 / 10.7
[32]
Mycoplasma genitalium 32%
27.8 / 67.3
[36]
35.9 / -40.6
[34]
[35]
Mycoplasma pneumoniae 40%
-11.6 / -67.8
-11.2 / -50.9
[10]
Mesoplasma florum 27%
-16.7 / -56.3
[12]
[14]
1.5 / 9.5
Mycoplasma capricolum 24%
-10.7 / -93
[11]
4.2 / 34.7
[13]
Mycoplasma mycoide 24%
4.8 / 11.5
[15]
Acholeplasma laidlawii 32%
-67 / -112.7
[2]
-27.7 / -63.4
[1]
Candidatus Phytoplasma mali 23%
108.2 / 257.1
[9]
48.2 / 109.1
[3]
Candidatus Phytoplasma australiense 28%
21.1 / 53.8
[8]
-2.5 / -24.4
[4]
10.2 / 46.7
[7]
Aster yellows phytoplasma 28%
-12.3 / -27.7
[5]
Onion yelows phytoplasma 29%
1.5 / -5.4
[6]


B
Mycoplasma agalactiae 30%
0.61 / 0.92
[25]
[24]
-0.26 / -0.52
Mycoplasma synoviae 29%
0.22 / 2.71
[26]
0.04 / -0.18
[22]
Mycoplasma pulmonis 27%
0.78 / 1.27
[23]
0.22 / 0.23
[20]
Mycoplasma arthritidis 31%
0.57 / 0.6
[21]
0.04 / -0.08
[18]
Mycoplasma mobile 25%
0.52 / 1.45
[19]
0.13 / 1.21
[17]
Mycoplasma hyopneumoniae 29%
1.35 / 4.08
[27]
[16]
-0.04 / 0.45
Mycoplasma penetrans 26%
0.74 / 1.44
[30]
-0.09 / 0.21
[29]
Ureaplasma urealyticum 26%
0.65 / 0.89
[31]
0.3 / 0.69
[28]
Mycoplasma gallisepticum 32%
-0.17 / 1.35
[33]
0.26 / 0.48
[32]
-0.48 / -1.11
[10]
Mycoplasma genitalium 32%
0.39 / 0.97
[36]
0.52 / 0.31
[34]
Mycoplasma pneumoniae 40%
[35]
0.04 / -0.21
Mesoplasma florum 27%
-0.17 / -0.13
[12]
0.04 / 0.13
[14]
0.22 / -0.58
[11]
Mycoplasma capricolum 24%
0.09 / 0.85
[13]
Mycoplasma mycoide 24%
0 / 0.06
[15]
Acholeplasma laidlawii 32%
-0.39 / -0.13
[2]
Candidatus Phytoplasma mali 23%
1.13 / 4.39
[9]
-0.39 / -0.77
[1]
0.3 / 1.6
[3]
Candidatus Phytoplasma australiense 28%
0.3 / 2.06
[8]
0.09 / 0.65
[7]
Aster yellows phytoplasma 28%
0.13 / -0.11
[4]
-0.13 / 0.27
[5]
Onion yelows phytoplasma 29%
0.04 / 0.65
[6]

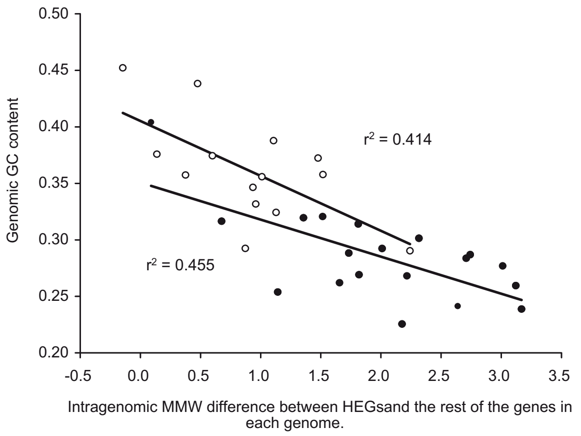


Figure 2. Linear regressions of the MMW intragenomic difference and the genomic GC content. MMW intragenomic difference between HEGs and the rest of the genes was calculated for each organisms analyzed in the present study. Coefficient of determination of the linear regression are shown for Mollicutes (black dots), $p < 0.002$ and the outgroup (white dots), $p < 0.015$.

($r = 0.52$; $p < 0.001$), while the third was significantly correlated with hidropathy ($r = 0.62$; $p < 0.001$). From now on, we shall refer to this pattern, as "pattern 1."

In the cases of Candidatus *Phytoplasma australiense*, *M. arthritidis*, *M. genitalium*, and *Acholeplasma laidlawii*, the principal trend (which accounted between 17.6 and 27.5% of the total variability) was associated with the hydropathy level of each protein, and to a lesser extent with aromaticity ($0.20 < r < 0.62$, $p < 0.001$). In these species (and in *M. penetrans*), the second main axis was related to expression levels (Supplementary Figure 2). We refer this pattern as "pattern 2" (which is almost the "mirror" of pattern 1). Moreover, in all these species, the position of the proteins along this axis was also correlated with MMW and $GC_{(1-2)}$ (Table 2). In *M. penetrans*, the third axis (11.8% of the variability) was correlated with hidropathy and to a lesser extent with aromaticity. Remarkably, in the latter species, the main axis was linked with the usage of Cys.

These findings suggest that HEGs are characterized by smaller, GC-rich, and less aromatic residues, while LEGs show the opposite trend. Due to restrictions imposed by the genetic code, it is not surprising that in each species, the "expression linked axis" was strongly correlated with $GC_{(1-2)}$, while it was independent of GC3.

Pairwise dN distances were estimated within phylogenetic groups, as a result, the distributions of the sequences according to dN showed that the most heavily expressed genes display the lower values. This fact held true for all the species analyzed here. Moreover, as expected, there were strong correlations within each species between the position of the genes along the "expression" axes and dN (Table 2 and Supplementary Table 2).

M. mobile, *M. pneumoniae*, and *M. gallisepticum* fit neither in pattern 1 nor 2. The first trend in M. mobile was not associated with expression levels, though it was significantly correlated with MMW. In addition, the second trend correlated with the usage of Phe, Asn, and Trp. In the case of M. pneumoniae and M. gallisepticum, the third trend (≈11.5% of the variability) was marginally related to the expression level. We stress, however, that the correlation with dN was significant ($0.4 < r < 0.63$, $p < 0.001$).

Interestingly, COA on RAAU did not find any trend associated with expression levels in the species used as outgroup, even in those whose GC content was in the range of Mollicutes. It seems reasonable to assume that expression levels was an important source of amino acids usage early in the evolution of Mollicutes, and this can be interpreted as a by-product of the evolutionary path toward a parasitic life style. Indeed, simple mutational bias is not enough to explain this pattern, since, as mentioned above, some organisms in the outgroup display genomic GC levels that fall within the range found in Mollicutes. Thus, genetic drift and/or a high evolutionary rate might (at least partially) explain these patterns and the similar ones observed in other endosymbiotic bacteria. We propose that, among Mollicutes, due to genetic drift, genes that do not experience a highly constrained amino acid sequence evolution (LEGs) may have suffered rapid accumulation of AT-rich amino acids generating significant internal variation. On the contrary, under this scenario, the same genes in free-living species may have more time to accumulate compensatory changes in response to mutational bias, which in turn keeps intragenomic variability at a low level.

To sum up, COA on RAAU showed that there were differences among the principal trends in amino acid usage among Mollicutes. However, two factors were widely distributed in this class. First of all, HEGs were more conserved at nonsynonymous positions, and second, expression, MMW, aromaticity, and GC content at the first and second codon positions played a primary role when explaining intragenomic RAAU variation. Similar situations have been reported in *Blochmannia floridanus* (Banerjee et al., 2004) and several endosymbiont γ-3-proteobacteria (Herbeck, Wall, & Wernegreen, 2003; Palacios & Wernegreen, 2002; Rispe et al., 2004).

Amino acids usage bias in highly expressed genes (β)

The direction of the relative amino acid biases in the reference set in comparison with the whole set of genes in each species (see Material and methods) were conserved across all studied organisms (Supplementary Table 3). We note that almost identical results were obtained when the highly expressed set of genes was

Table 2. Correlation coefficient of the positions of the genes in the COA-generated axes with protein and gene properties analyzed*.

Organisms		Axis[a]	MMW[b]	GC_1[c]	GC_2[c]	GC_3[c]	Gravy[d]	Aromo[e]	MELP[f]	dN[g]	
Pattern 1	*M. synoviae*	1 (25.6)	0.81	0.85	0.83	0.06	0.14	0.62	0.57	0.54	0.54
		2 (13.3)	0.10	0.33	0.15	0.16	0.46	0.52	0.34	0.32	0.39
		3 (11.5	0.03	0.14	0.27	0.26	0.62	0.33	0.09	0.20	0.25
	M. agalactiae	1 (18.8)	0.74	0.83	0.73	0.28	0.07	0.62	0.62	0.53	0.59
		2 (16.9)	0.00	0.39	0.17	0.22	0.91	0.61	0.10	0.10	0.21
	M. pulmonis	1 (22.5)	0.79	0.82	0.81	0.00	0.10	0.66	0.35	0.56	0.59
		2 (14.1)	0.10	0.32	0.22	0.10	0.93	0.57	0.00	0.05	0.13
	M. hyopneumoniae	1 (19.7)	0.75	0.87	0.73	0.35	0.22	0.74	0.45	0.61	0.69
		2 (16.8)	0.20	0.22	0.35	0.14	0.89	0.45	0.00	0.02	0.12
	U. urealyticum	1 (22.5)	0.73	0.76	0.88	0.00	0.10	0.52	0.48	0.58	0.67
		2 (15.5)	0.06	0.57	0.07	0.17	0.89	0.62	0.26	0.21	0.40
	M. florum	1 (21.9)	0.77	0.84	0.83	0.04	0.03	0.58	0.67	0.63	0.65
		2 (15.7)	0.24	0.33	0.30	0.08	0.87	0.53	0.14	0.15	0.19
	M. capricolum	1 (24.2)	0.79	0.79	0.85	0.02	0.14	0.49	0.56	0.32	0.66
		2 (16.4)	0.07	0.46	0.14	0.03	0.82	0.65	0.17	0.01	0.33
	M. mycoide	1 (22.3)	0.79	0.79	0.85	0.02	0.14	0.49	0.56	0.32	0.06
		2 (16.4)	0.07	0.46	0.14	0.03	0.82	0.65	0.17	0.01	0.33
	C. P. mali	1 (21.8)	0.73	0.82	0.82	0.11	0.24	0.68	0.14	0.60	0.69
		2 (15.9)	0.11	0.29	0.27	0.17	0.86	0.63	0.10	0.03	0.16
	O. y. phytoplasma	1 (20)	0.77	0.54	0.74	0.03	0.47	0.14	0.35	0.23	0.58
		2 (15.4)	0.24	0.66	0.17	0.01	0.74	0.84	0.42	0.13	0.55
	A. y. phytoplasma	1 (18.7)	0.77	0.66	0.79	0.04	0.40	0.24	0.41	0.25	0.58
		2 (15.6)	0.24	0.58	0.09	0.04	0.68	0.81	0.40	0.12	0.19
Pattern 2	*C. P. australiense*	1 (16.8	0.13	0.47	0.09	0.09	0.83	0.74	0.18	0.25	0.39
		2 (16.3)	0.82	0.76	0.73	0.15	0.22	0.16	0.48	0.52	0.60
	M. arthritidis	1 (27.5)	0.02	0.24	0.31	0.35	0.69	0.60	0.00	0.01	0.08
		2 (14.7)	0.63	0.84	0.67	0.12	0.22	0.65	0.60	0.53	0.64
	M. genitalium	1 (18.1)	0.30	0.47	0.33	0.00	0.85	0.62	0.08	0.03	0.06
		2 (14.6)	0.63	0.62	0.64	0.17	0.14	0.69	0.24	0.15	0.19
	A. laidlawii	1 (17.6)	0.56	0.17	0.42	0.05	0.88	0.35	0.02	0.01	0.07
		2 (13.6)	0.62	0.84	0.66	0.30	0.28	0.69	0.63	0.50	0.59
	M. penetrans	1 (29.1)	0.77	0.14	0.84	0.30	0.24	0.04	0.08	0.00	0.09
		2 (14.9)	0.55	0.81	0.51	0.14	0.07	0.62	0.51	0.51	0.67
		3 (11.8)	0.10	0.41	0.01	0.22	0.83	0.63	0.22	0.04	0.10
	M. mobile	1 (23.3)	0.83	0.40	0.88	0.28	0.55	0.20	0.24	0.36	0.47
		2 (18.3)	0.26	0.80	0.33	0.00	0.20	0.36	0.52	0.57	0.64
	M. pneumoniae	1 (16.5)	0.59	0.20	0.79	0.22	0.46	0.33	0.14	0.02	0.17
		2 (13.5)	0.28	0.33	0.24	0.04	0.79	0.50	0.08	0.01	0.24
		3 (11.1)	0.35	0.71	0.33	0.32	0.03	0.58	0.33	0.37	0.40
	M. gallisepticum	1 (23.7)	0.86	0.22	0.89	0.41	0.01	0.26	0.32	0.15	0.37
		2 (16.7)	0.08	0.73	0.04	0.24	0.82	0.75	0.37	0.24	0.42
		3 (11.5)	0.33	0.36	0.32	0.14	0.24	0.39	0.47	0.48	0.57

[a] Represents the axes generated by Correspondence analysis (COA). In parenthesis is shown the variability explained by each axis.

[b] Mean molecular weight of the protein.

[c] Guanine plus Cytosine composition at first, second and third codon position of the gene.

[d] General average hydropathicity score of each protein, calculated according to Kyte and Doolittle (1982).

[e] Aromaticity score of the protein, calculated as the frequency of aromatic amino acids (Phe, Tyr,Trp).

[f] Codon-based expression level predictor independent of length and composition.

[g] Minimum and Maximum correlation coefficient of the dN estimation for orthologs with their respective gene position values in each axis (see Supplementary Table 1).

[*] Italics represent non-significant correlation values ($p > 0.001$). Organisms are listed according to monophyletic groups within Mollicutes they belong to: Hominis, Pneumoniae, Spiroplasma and Phytoplasma-like.

derived from COA on RAAU (Supplementary Table 3). Interestingly, there are several differences between Mollicutes and Firmicutes (Table 3). As can be seen, 13 residues display statistically significant differences between the two groups (*U*-test, $p < 0.01$). In other words, although the direction of the biases in the Mollicutes and Firmicutes are identical, the strength of the biases are different. Two nonmutually exclusive explanations can account for this fact. The first is the existence of a strong phylogenetical inertia (see above and Supplementary Figure 3), while the second can be related to differences in lifestyles (parasitic vs. free-living organisms). This may be partly explained by the important reduction of genes involved in amino acid metabolism in the common

Table 3. Comparison of amino acid usage bias of HEGs (P($_{\text{ref}}$)) among groups^{a,b}.

Amino acid	Mollicutes	No-Mollicutes	Significance$^{(c)}$	SEM group	AAA group	Significance$^{(c)}$
Ala	0.54 (±0.20)	0.26 (±0.15)	*	−0.47 (±0.03)	−0.47 (±0.02)	
Arg	1.45 (±0.21)	1.15 (±0.20)	*	−0.25 (±0.03)	−0.25 (±0.02)	
Asn	−0.31(±−0.08)	−0.15 (±0.13)	*	−0.27 (±0.08)	−0.24 (±0.07)	
Asp	−0.29 (±0.07)	−0.19 (±0.09)	*	0.34 (±0.10)	0.12 (±0.04)	
Cys	0.15 (±0.48)	−0.12 (±0.23)		0.31 (±0.10)	0.55 (±0.19)	*
Gln	0.00 (±0.13)	−0.16 (±0.21)	*	−0.16 (±0.06)	−0.06 (±0.09)	
Glu	−0.08 (±0.06)	0.01 (±0.09)	*	0.12 (±0.11)	0.07 (±0.09)	
Gly	0.41 (±0.14)	0.19 (±0.06)	*	0.14 (±0.10)	0.08 (±0.02)	
His	0.39 (±0.24)	0.02 (±0.22)	*	0.51 (±0.21)	0.64 (±0.1)	*
Ile	−0.26 (±0.07)	−0.28 (±0.07)		−0.40 (±0.05)	−0.41 (±0.04)	
Leu	−0.25 (±0.03)	−0.31 (±0.05)	*	0.49 (±0.19)	0.13 (±0.18)	
Lys	0.33 (±0.12)	0.50 (±0.20)	*	0.02 (±0.12)	−0.08 (±0.14)	
Met	0.28 (±0.15)	−0.04 (±0.09)	*	−0.32 (±0.07)	−0.27 (±0.06)	*
Phe	−0.47 (±0.03)	−0.43 (±0.05)	*	0.36 (±0.11)	0.25 (±0.07)	
Pro	0.11 (±0.11)	0.10 (±0.15)		−0.31 (±0.05)	−0.24 (±0.06)	
Ser	−0.14 (±0.08)	−0.17 (±0.07)		−0.08 (±0.06)	−0.09 (±0.02)	
Thr	0.12 (±0.09)	0.03 (±0.08)		0.32 (±0.44)	−0.33 (±0.25)	
Trp	−0.48 (±0.07)	−0.55 (±0.08)		−0.51 (±0.05)	−0.41 (±0.08)	*
Tyr	−0.40 (±0.05)	−0.33 (±0.06)	*	1.40 (±0.19)	1.57 (±0.20)	
Val	0.37 (±0.17)	0.30 (±0.08)		0.38 (±0.13)	0.47 (±0.12)	

[a] Average $\beta_{(\text{ref})}$ for each amino acid in each phylogenetic group. In parenthesis is shown the standard deviation of the distribution.

[b] Amino acid usage comparisons were made between Mollicutes vs. Non-Mollicutes and between the AAA branch (Phytplasma and Acholeplasma) vs. the SEM branch (Mesoplasma, Mycoplasma and Ureaplasma).

[c] The statistically significant differences are marked with '*' (U-test, $p < 0.01$).

ancestor of Mollicutes (Chen, Chung, Lin, & Kuo, 2012).

Species which belong to the branches AAA (*Phytplasma* and *Acholeplasma*) and SEM (*Mesoplasma*, *Mycoplasma* and *Ureaplasma*) were also compared (Table 3). Results showed similar biases in all amino acids, and the differences became statistically significant only for Cys, His, Met, and Trp (*U*-test, $p < 0.01$). No particular feature of these amino acids seems to explain the observed pattern. We think that these results can be partially associated with differences in the main hosts (animal vs. plants) and/or the independent pathway toward parasitic lifestyle of each group (Razin, Yogev, & Naot, 1998).

We identified some particular cases that did not match the expected phylogenetic pattern (Supplementary Figure 3). This is the case of species of the *Clostridium* genus, *Mycoplasma mycoides*, and *A. laidlawii*. The idiosyncratic behavior observed in *Clostridium* may be linked with their basal phylogenetic position respective to the rest of the organisms analyzed. Similarly, *Acholeplasma* are believed to have retained many ancestral characteristics in comparison to phytoplasmas and other highly diverged Mollicutes (Bai et al., 2006; Iriarte et al., 2011; Lazarev et al., 2011) No particular feature seems to explain the observed result for *M. mycoides*.

Once again, the underrepresentation of $AT_{(1-2)}$-rich codons, and aromatic and heavy amino acids may explain the trends detected in HEGs, but none of these properties could explain the whole trend by itself. These properties, which are deeply interdependent and superimposed, might affect amino acid usage evolution in Mollicutes but also in the other Firmicutes analyzed (see the estimated $\beta_{(\text{REF})}$ in Supplementary Table 3). Given that HEGs biases (estimated with $\beta_{(\text{REF})}$) are mainly conserved in the outgroup and in several other free-living species, there is no link between these trends in HEGs and distinctive selection pressures in Mollicutes. Besides, when the MMW were estimated without considering Tyr, Phe and Trp, the significance of the correlations were lost. This suggests that the avoidance of these three aromatic residues is a main determinant for the MMW biases in HEGs and supports the interdependence among the properties studied.

As previously reported (Herbeck, Wall, & Wernegreen, 2003; Palacios & Wernegreen, 2002; Rispe et al., 2004), purifying selection acting at the amino acid usage level on HEGs in parasitic species is expected to be less effective than drive and mutational biases in relation to free-living relatives. On the other hand, we found that in Mollicutes (as well as in other symbiotic species), expression levels (plus GC_{12}, MW, etc) is a main factor, often the first one, explaining an important part of the internal variability on the usage of amino acids (11% < inertia of the "expression" axis < 25%). Examples can be found in Palacios and Wernegreen (2002), Herbeck, Wall, and Wernegreen (2003), and Rispe et al. (2004) for parasitic endosymbiotic bacteria. Considering unicellular eukaryotic parasites, similar results were reported for *Leishmania* (Chauhan, Vidyarthi, & Poddar, 2011) and for *Plasmodium falciparum* (Chanda, Pan, & Dutta, 2005). In opposition to the "symbiotic" pattern, for the rest of the firmicutes, no axis was related with expression levels. In *E. coli* and *Thermotoga maritima*, the "expression" axes only explained between 11.3 and 13% of the internal variability (Lobry & Gautier, 1994; Zavala et al., 2002). This is also true for *Saccharomyces cerevisiae* (Kahali, Basak, & Ghosh, 2007). Once again, this result is quite puzzling for parasitic bacteria, where a lower effect of purifying selection is expected. We believe that the conserved biases of HEGs is widespread in a variety of taxa, but that in many parasitic species it emerges as a very important factor explaining intragenomic variability. Note that this statement does not rule out the possibility that some other features, such as genome contraction, may be partially responsible of the observed pattern.

Molecular evolutionary rate, mutational bias and amino acids changes

Many years ago, Karlin and Bucher (1992) described correlated amino acid frequencies based on strong (GC-rich) and weak (AT-rich) codons, among other features. Nowadays, it is widely accepted that changes in nucleotide biases affect global amino acid composition and probably molecular weight and cost. Singer and Hickey (2000), among others, reported that Phe, Tyr, Met, Ile, Asp, and Lys frequencies correlate negatively with GC composition, while the opposite occurs with Gly, Ala, Arg, and Pro. Mutational bias and genetic drift effects may be always less pronounced in HEGs that are believed to evolve slower at nonsynonymous and synonymous sites (Pal, Papp, & Hurst, 2001, 2003). In these sequences, conserved amino acids biases are driven by purifying selection and may have a functional, structural, and/or energy cost basis. Nevertheless, our results show that the relative contribution of each factor is not clear, which one-by-one cannot completely explain the observed patterns.

Twenty-three (23) out of 85 putative orthologous sequences were considered HEGs (elongation factors tufA, fusA, and tsf plus 20 ribosomal proteins genes). The other 62 putative orthologous sequences were considered LEGs (tRNA synthetases, kinases, DNA-ligase, among others) (see Supplementary Table 1 for a complete list of all orthologs). This classification was conservative since all the latter group was certainly not composed only by sequences expressed at low levels.

The degree of conservation between these two groups was compared as the sum of the dN of each branch across the tree. Results showed that HEGs were significantly more conserved than LEGs (*U*-test, $p < 0.0001$). Note that β, estimated only for this group of genes, showed essentially the same trend as comparisons made considering the whole proteome (data not shown). Both results validated working with orthologous sequences and suggest that the effect of horizontal gene transfer events on the general patterns is rather marginal.

The trends in molecular weight and aromaticity, derived from the amino acid changes traced in the phylogeny, were averaged for both groups of orthologs in each branch (Figures 2 and 3). Although the reconstruction of the ancestral compositional biases was not done, a simple inspection of the tree suggested that the observed trends may be explained by the differences in genomic GC. As expected, changes in compositional biases toward an increased GC content were associated with a decreasing MMW tendency, and *vice versa* (see Figure 2). HEGs usually experience the same bias than LEGs but with a lesser degree, which may be mainly due to their higher degree of conservation. In agreement with previous results, the aromatic trends strongly correlated with average MW changes (see above), both for HEGs ($r = 0.74$) and LEGs ($r = 0.83$) (Figure 1).

As previously stated, and as can be deduced from the structure of the genetic code, there is an expected negative correlation between the GC content of a set of genes and the molecular weight of the corresponding proteins. However, since not all amino acid changes are equally frequent, it is interesting to analyze "top 20%" of the most frequent ones, which accounts for more than 70% of the inferred total changes in the phylogeny. The results confirm the idea that positive ΔGC are associated with negative ΔMW trends ($r = -0.49$, $p < 0.0001$).

The global amino acid substitution matrix of HEGs and LEGs was compared. A contingency χ^2 test was used to establish significant differences between both groups of genes. A total of 14 changes were recognized as significantly more frequent in HEGs (χ^2 $p < 0.01$), while 12 changes were more frequent in LEGs (χ^2 $p < 0.01$) (Supplementary Figure 4). Results showed that, on an average, the more frequent amino acid changes in HEGs were also characterized by a decreasing MW ($\Delta MW = -8.10$), while the opposite was true for LEGs ($\Delta MW = 1.85$). The average AMW and aromatic trend per inferred change in each branch confirmed this observation. Taken together, these results suggest that HEGs are characterized not only by a reduced nonsynonymous substitution rate but also by more conservative changes.

Identical directional trends in HEGs and LEGs were observed in some lineages (Figure 1). This probably reflects the action of a strong nucleotide compositional bias throughout these lineages. In addition, a negative correlation between the observed MW bias of HEGs (the difference between MW of HEGs in relation to the rest of the genes in each genome) and the genomic GC was observed for all the species analyzed in the present study ($r^2 = 0.62$). Note that the observed bias in MW was slightly more pronounced in Mollicutes than in the outgroup when considering the same GC content variation (compare the slopes in Figure 2). This supports the idea that under the same mutational bias, the intragenomic difference between HEGs and LEGs may increase faster in Mollicutes than in the outgroup. We believe that this is probably attributable to the noteworthy effect of genetic drift and the characteristic high mutation rate. This observation does not undermine the idea that intragenomic MMW difference between HEGs and LEGs is mainly driven by the frequency of aromatic residues, which are among the biggest amino acids. Therefore, even thought that this "aromatic" bias may also have a cell economy explanation, other functional properties cannot be completely discarded, for instance, residue-specific interactions or functions (Dougherty, 2007; Gallivan & Dougherty, 1999).

In summary, we propose that among Mollicutes, several pressures affect amino acid usage. The main factors are mutational bias, genetic drift, and purifying selection. The first two should be quantitatively very important in these organisms given their way of life, which in most cases is characterized by recursive bottlenecks and should be more effective in LEGs. On the other hand, negative selection should be very important for HEGs. Thus, the cumulative effect of these evolutionary forces acting differently among genes, might explain the observed increase of the internal variability in the genomes of Mollicutes.

Abbreviations

GC	—	molar content of guanine + cytosine
HEGs	—	highly expressed genes
LEGs	—	lowly expressed genes
COA	—	correspondence analysis
RAAU	—	relative amino acid usage
MELP	—	MILC-based expression level predictor
dN	—	Nonsynonymous substitution rate
β	—	RAAU bias
MW	—	Molecular weight

Acknowledgment

We thank Rosina Piovani for fruitful discussions. This work was partially supported by Fondo Clemente Estable 2007, FCE 722 (ANII, Uruguay) and PEDECIBA (Uruguay).

Supplementary Material

The supplementary material for this paper is available online at http://dx.doi.10.1080/07391102.2012.748636.

References

Akashi, H., & Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. *Proceedings of the National Academy of Sciences, 99*, 3695–3700.

Bai, X., Zhang, J., Ewing, A., Miller, S. A., Jancso Radek, A., Shevchenko, D. V., … Hogenhout, S. A. (2006). Living with genome instability: The adaptation of phytoplasmas to diverse environments of their insect and plant hosts. *Journal of Bacteriology, 188*, 3682–3696.

Banerjee, T., Basak, S., Gupta, S. K., & Ghosh, T. C. (2004). Evolutionary forces in shaping the codon and amino acid usages in Blochmannia floridanus. *Journal of Biomolecular Structure & Dynamics, 22*, 13–23.

Banerjee, T., & Ghosh, T. C. (2006). Gene expression level shapes the amino acid usages in Prochlorococcus marinus MED4. *Journal of Biomolecular Structure & Dynamics, 23*, 547–554.

Basak, S., Banerjee, T., Gupta, S. K., & Ghosh, T. C. (2004). Investigation on the causes of codon and amino acid usages variation between thermophilic Aquifex aeolicus and mesophilic Bacillus subtilis. *Journal of Biomolecular Structure & Dynamics, 22*, 205–214.

Boone, D. R., Castenholz, R. W., & Garrity, G. M. (2001). *Bergey's manual of systematic bacteriology* (George M. Garrity, editor-in-chief). New York, NY: Springer.

Chanda, I., Pan, A., & Dutta, C. (2005). Proteome composition in Plasmodium falciparum: Higher usage of GC-rich nonsynonymous codons in highly expressed genes. *Journal of Molecular Evolution, 61*, 513–523.

Chauhan, N., Vidyarthi, A. S., & Poddar, R. (2011). Comparative multivariate analysis of codon and amino acid usage in three Leishmania genomes. *Genomics Proteomics Bioinformatics, 9*, 218–228.

Chen, L. L., Chung, W. C., Lin, C. P., & Kuo, C. H. (2012). Comparative analysis of gene content evolution in phytoplasmas and mycoplasmas. *PLoS ONE, 7*, e34407.

Das, S., Paul, S., Chatterjee, S., & Dutta, C. (2005). Codon and amino acid usage in two major human pathogens of genus Bartonella–optimization between replicational-transcriptional selection, translational control and cost minimization. *DNA Research, 12*, 91–102.

Dougherty, D. A. (2007). Cation-pi interactions involving aromatic amino acids. *Journal of Nutrition, 137*, 1504S–1508Sdiscussion 1516S–1517S.

Duret, L., & Mouchiroud, D. (2000). Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. *Molecular Biology and Evolution, 17*, 68–74.

Gallivan, J. P., & Dougherty, D. A. (1999). Cation-pi interactions in structural biology. *Proceedings of the National Academy of Sciences, 96*, 9459–9464.

Garat, B., & Musto, H. (2000). Trends of amino acid usage in the proteins from the unicellular parasite Giardia lamblia. *Biochemical and Biophysical Research Communications, 279*, 996–1000.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Herbeck, J. T., Wall, D. P., & Wernegreen, J. J. (2003). Gene expression level influences amino acid usage, but not codon usage, in the tsetse fly endosymbiont Wigglesworthia. *Microbiology, 149*, 2585–2596.

Iriarte, A., Baraibar, J. D., Romero, H., & Musto, H. (2011). Selected codon usage bias in members of the class Mollicutes. *Gene, 473*, 110–118.

Kahali, B., Basak, S., & Ghosh, T. C. (2007). Reinvestigating the codon and amino acid usage of *S. cerevisiae* genome: A new insight from protein secondary structure analysis. *Biochemical and Biophysical Research Communications, 354*, 693–699.

Karlin, S., & Bucher, P. (1992). Correlation analysis of amino acid usage in protein classes. *Proceedings of the National Academy of sciences, 89*, 12165–12169.

Lazarev, V. N., Levitskii, S. A., Basovskii, Y. I., Chukin, M. M., Akopian, T. A., Vereshchagin, V. V., … Govorun, V. M. (2011). Complete genome and proteome of Acholeplasma laidlawii. *Journal of Bacteriology, 193*, 4943–4953.

Lobry, J. R., & Gautier, C. (1994). Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. *Nucleic Acids Research, 22*, 3174–3180.

McDonald, J. H. (2010). Temperature adaptation at homologous sites in proteins from nine thermophile-mesophile species pairs. *Genome Biology and Evolution, 2*, 267–276.

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

Naya, H., Zavala, A., Romero, H., Rodriguez-Maseda, H., & Musto, H. (2004). Correspondence analysis of amino acid usage within the family Bacillaceae. *Biochemical and Biophysical Research Communications, 325*, 1252–1257.

Oshima, K., & Nishida, H. (2007). Phylogenetic relationships among mycoplasmas based on the whole genomic information. *Journal of Molecular Evolution, 65*, 249–258.

Pal, C., Papp, B., & Hurst, L. D. (2001). Highly expressed genes in yeast evolve slowly. *Genetics, 158*, 927–931.

Pal, C., Papp, B., & Hurst, L. D. (2003). Genomic function: Rate of evolution and gene dispensability. *Nature, 421*, 496–497; discussion 497–498.

Palacios, C., & Wernegreen, J. J. (2002). A strong effect of AT mutational bias on amino acid usage in Buchnera is mitigated at high-expression genes. *Molecular Biology and Evolution, 19*, 1575–1584.

R Development Core Team. (2012). R: A Language and Environment for Statistical Computing.

Razin, S., Yogev, D., & Naot, Y. (1998). Molecular biology and pathogenicity of mycoplasmas. *Microbiology and Molecular Biology Reviews, 62*, 1094–1156.

Rispe, C., Delmotte, F., van Ham, R. C., & Moya, A. (2004). Mutational and selective pressures on codon and amino acid usage in Buchnera, endosymbiotic bacteria of aphids. *Genome Research, 14*, 44–53.

Rocha, E. P., & Danchin, A. (2004). An analysis of determinants of amino acids substitution rates in bacterial proteins. *Molecular Biology and Evolution, 21*, 108–116.

Sabbia, V., Piovani, R., Naya, H., Rodriguez-Maseda, H., Romero, H., & Musto, H. (2007). Trends of amino acid usage in the proteins from the human genome. *Journal of Biomolecular Structure & Dynamics, 25*, 55–59.

Schaber, J., Rispe, C., Wernegreen, J., Buness, A., Delmotte, F., Silva, F.J., & Moya, A. (2005). Gene expression levels influence amino acid usage and evolutionary rates in endosymbiotic bacteria. *Gene, 352*, 109–117.

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*, 1141–1153.

Singer, G. A., & Hickey, D. A. (2000). Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Molecular Biology and Evolution, 17*, 1581–1588.

Subramanian, S., & Kumar, S. (2004). Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics, 168*, 373–381.

Supek, F., Skunca, N., Repar, J., Vlahovicek, K., & Smuc, T. (2010). Translational selection is ubiquitous in prokaryotes. *PLoS Genetics, 6*, e1001004.

Supek, F., & Vlahovicek, K. (2004). INCA: synonymous codon usage analysis and clustering by means of self-organizing map. *Bioinformatics, 20*, 2329–2330.

Supek, F., & Vlahovicek, K. (2005). Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity. *BMC Bioinformatics, 6*, 182.

Suzuki, R., & Shimodaira, H. (2006). Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics, 22*, 1540–1542.

Swire, J. (2007). Selection on synthesis cost affects interprotein amino acid usage in all three domains of life. *Journal of Molecular Evolution, 64*, 558–571.

Tekaia, F., & Yeramian, E. (2006). Evolution of proteomes: Fundamental signatures and global trends in amino acid compositions. *BMC Genomics, 7*, 307.

Tekaia, F., Yeramian, E., & Dujon, B. (2002). Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: A global picture with correspondence analysis. *Gene, 297*, 51–60.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research, 22*, 4673–4680.

Wang, G. Z., & Lercher, M. J. (2010). Amino acid composition in endothermic vertebrates is biased in the same direction as in thermophilic prokaryotes. *BMC Evolutionary Biology, 10*, 263.

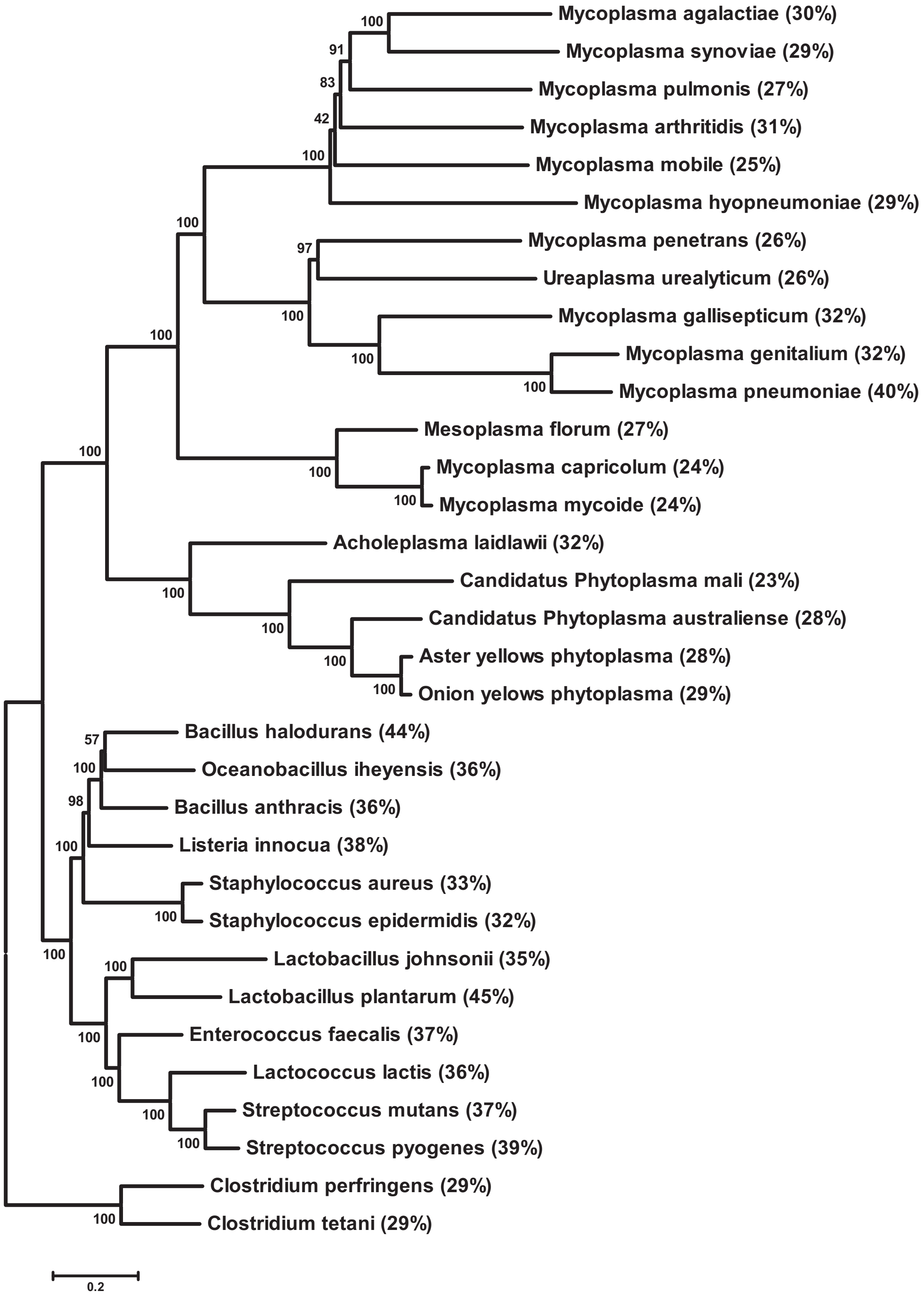
Woese, C. R., Maniloff, J., & Zablen, L. B. (1980). Phylogenetic analysis of the mycoplasmas. *Proceedings of the National Academy of Sciences, 77*, 494–498.

Wolf, M., Muller, T., Dandekar, T., & Pollack, J. D. (2004). Phylogeny of Firmicutes with special reference to Mycoplasma (Mollicutes) as inferred from phosphoglycerate kinase amino acid sequence data. *International Journal of Systematic and Evolutionary Microbiology, 54*, 871–875.

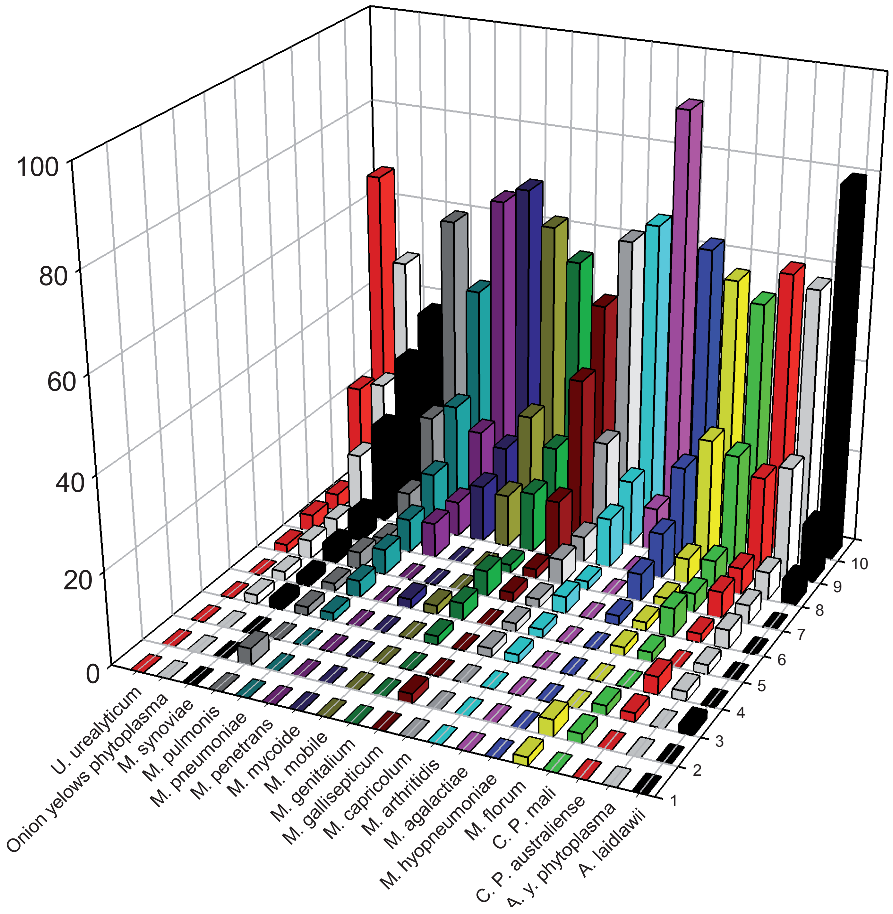
Yamao, F., Andachi, Y., Muto, A., Ikemura, T., & Osawa, S. (1991). Levels of tRNAs in bacterial cells as affected by amino acid usage in proteins. *Nucleic Acids Research, 19*, 6119–6122.

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution, 24*, 1586–1591.

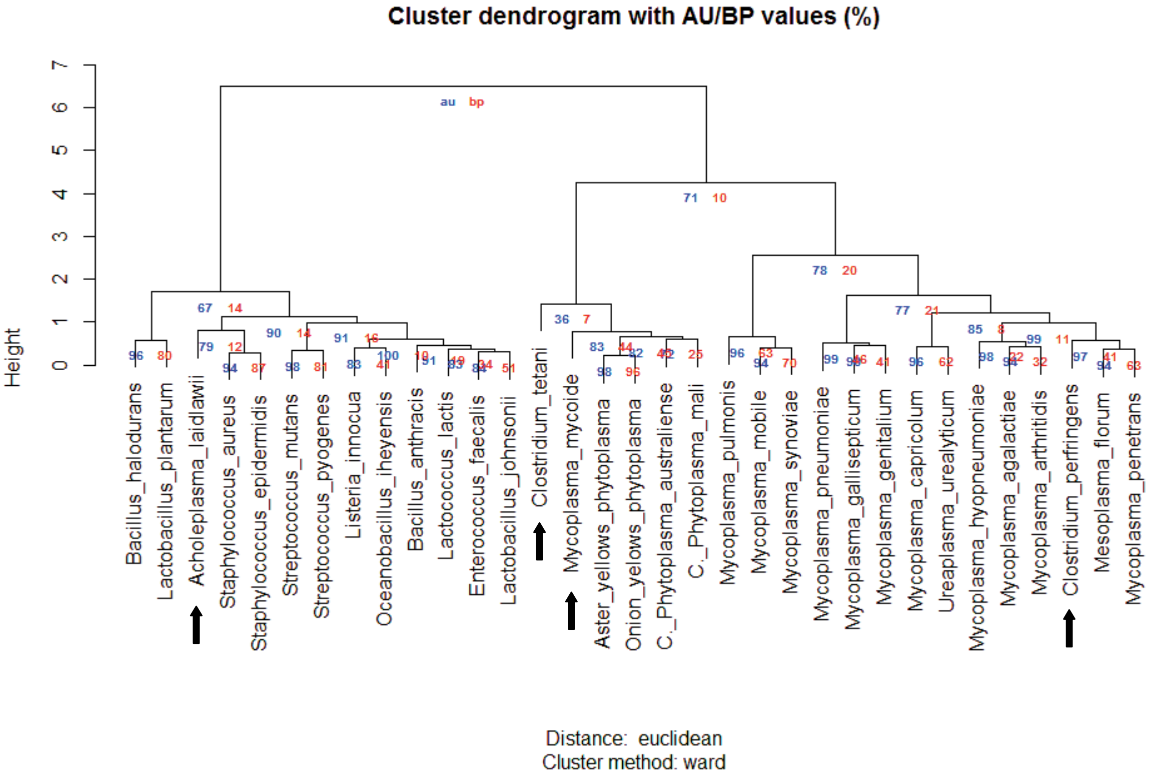
Zavala, A., Naya, H., Romero, H., & Musto, H. (2002). Trends in codon and amino acid usage in Thermotoga maritima. *Journal of Molecular Evolution, 54*, 563–568.


Mycoplasma agalactiae (30%)
Mycoplasma synoviae (29%)
Mycoplasma pulmonis (27%)
Mycoplasma arthritidis (31%)
Mycoplasma mobile (25%)
Mycoplasma hyopneumoniae (29%)
Mycoplasma penetrans (26%)
Ureaplasma urealyticum (26%)
Mycoplasma gallisepticum (32%)
Mycoplasma genitalium (32%)
Mycoplasma pneumoniae (40%)
Mesoplasma florum (27%)
Mycoplasma capricolum (24%)
Mycoplasma mycoide (24%)
Acholeplasma laidlawii (32%)
Candidatus Phytoplasma mali (23%)
Candidatus Phytoplasma australiense (28%)
Aster yellows phytoplasma (28%)
Onion yelows phytoplasma (29%)
Bacillus halodurans (44%)
Oceanobacillus iheyensis (36%)
Bacillus anthracis (36%)
Listeria innocua (38%)
Staphylococcus aureus (33%)
Staphylococcus epidermidis (32%)
Lactobacillus johnsonii (35%)
Lactobacillus plantarum (45%)
Enterococcus faecalis (37%)
Lactococcus lactis (36%)
Streptococcus mutans (37%)
Streptococcus pyogenes (39%)
Clostridium perfringens (29%)
Clostridium tetani (29%)
100
91
83
42
100
100
97
100
100
100
100
100
100
100
100
100
100
100
100
57
100
98
100
100
100
100
100
100
100
100
100
100
0.2

Supplementary Figure 1. Phylogenetic relationships among microorganisms analyzed as inferred by Neighbor Joining. Reconstruction was based on 85 concatenated orthologous proteins using the amino acid JTT+G matrix-based model. Numbers in interior branches represent the nodal support values, estimated using non-parametric bootstrap technique with 5,000 replicates.



Supplementary Figure 2. Histogram of the distribution of highly expressed genes (%) (genes encoding ribosomal proteins + elongation factors: tufB, tsf, fusA) along the axis related to expression levels (see Table 1) for each species. The axis was divided into 10 parts, each of them containing an equal number of genes.



Supplementary Figure 3. Hierarchical cluster analysis of studied species based on the amino acid usage bias (β) (Supplementary table 3). Ward's method was used for this analysis. Approximately Unbiased p-value (AU), (blue) and Bootstrap Probability (BP) value (red). Note that the observed β follows a phylogenetic pattern, when considering Mollicutes *vs.* No-Mollicutes or SEM branch *vs.* AAA branch. Exceptions are indicated with black arrows.

To From	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala					🟥		🟩													
Arg																				
Asn																				
Asp								🟩												
Cys																				
Gln			🟥																	
Glu	🟩																			
Gly																				
His												🟩							🟥	
Ile							🟥													🟩
Leu								🟩						🟥						🟩
Lys	🟩	🟩					🟥			🟥							🟩			
Met										🟥										
Phe																				
Pro	🟩																			
Ser	🟩		🟥																	
Thr																				
Trp									🟩											
Tyr											🟥	🟥		🟩						
Val					🟥															

Supplementary Figure 4. Contingency $\chi2$ tests were used to find which amino acid changes are significantly more frequent in HEGs (green) and LEGs (red) (P<0.01).

Supplementary Table 1.

List of 85 putative orthologous sequences and global dN estimation across Mollicutes phylogeny. 23 were considered highly expressed genes, HEGs$_{(orthologs)}$ and 62 representing the remaining genes of the genome, LEGs$_{(orthologs)}$.

	Product	Gene	dN
HEGs (orthologs)	30S ribosomal protein S10	rpsJ	3.32
	30S ribosomal protein S11	rps11	2.39
	30S ribosomal protein S12	rpsL	2.05
	30S ribosomal protein S13	rpsM	2.51
	30S ribosomal protein S3	rpsC	3.39
	30S ribosomal protein S4	rpsD	3.65
	30S ribosomal protein S5	rpsE	2.59
	30S ribosomal protein S7	rpsG	2.29
	30S ribosomal protein S8	rpsH	3.23
	30S ribosomal protein S9	rpsI	2.99
	50S ribosomal protein L1	rplA	3.23
	50S ribosomal protein L13	rplM	3.24
	50S ribosomal protein L16	rplP	2.07
	50S ribosomal protein L17	rplQ	3.83
	50S ribosomal protein L2	rplB	2.29
	50S ribosomal protein L27	rpmA	2.48
	50S ribosomal protein L3	rplC	3.15
	50S ribosomal protein L4	rplD	4.44
	50S ribosomal protein L5	rplE	2.83
	50S ribosomal protein L6	rplF	3.76
	elongation factor G	fusA	2.13
	elongation factor Ts	tsf	5.23
	elongation factor Tu	tuf	1.56
LEGs (orthologs)	ABC transporter ATP binding protein	-	3.90
	acetate kinase	ackA	4.83
	adenylate kinase	adk	5.97
	alanyl-tRNA synthetase	alaS	5.08
	asparaginyl-tRNA synthetase	asnC	4.12
	aspartyl-tRNA synthetase	aspS	4.63
	cell division protein	ftsY	3.58
	cysteinyl-tRNA synthetase	cysS	5.60
	cytidine monophosphate kinase	cmk	6.54
	DNA-directed RNA polymerase subunit alpha	rpoA	5.02
	DNA-directed RNA polymerase subunit beta	rpoB	3.22
	DNA-directed RNA polymerase subunit beta'	rpoC	3.23
	DNA gyrase subunit A	gyrA	3.76
	DNA gyrase subunit B	gyrB	3.14
	DNA ligase	lig	5.74
	DNA polymerase III DnaE	dnaE	6.23
	DNA primase	dnaG	10.25
	DNA topoisomerase I	topA	5.38
	elongation factor P	efp	3.18
	formamidopyrimidine-DNA glycosylase	fpg	5.98
	glutamyl-tRNA synthetase	gltX	4.44
	glycyl-tRNA synthetase	glyS	3.96
	GTPase ObgE	obgE	4.27
	GTP-binding protein EngA	engA	4.14
	GTP-binding protein LepA	lepA	2.65
	guanylate kinase	gmk	4.29
	histidyl-tRNA synthetase	hisS	6.56
	hypothetical protein	-	5.14
	hypothetical protein	-	5.32
	hypothetical protein	-	6.00
	inorganic pyrophosphatase	ppa	3.59
	isoleucyl-tRNA synthetase	ileS	4.62
	leucyl-tRNA synthetase	leuS	4.55
	methionine aminopeptidase	map	4.69
	methionyl-tRNA synthetase	metG	5.33
	molecular chaperone DnaK	dnaK	2.69
	nitrogen fixation protein class-V pyridoxal-phosphate aminotransferase	nifS	5.92
	O-sialoglycoprotein endopeptidase	gcp	4.18
	peptide chain release factor 1	prfA	3.82
	peptidyl-tRNA hydrolase	pth	5.36
	phenylalanyl-tRNA synthetase alpha chain	pheS	3.83
	prolyl-tRNA synthetase	proS	4.46
	putative serine protease, subtilase family active site	-	3.89
	ribosomal biogenesis GTPase	rbgA	6.39
	ribosomal large subunit pseudouridine synthase D	rluD	5.14
	RNA polymerase sigma factor RpoD	rpoD	4.09

	seryl-tRNA synthetase	serS	4.51
	signal recognition particle protein	ffh	3.71
	threonyl-tRNA synthetase	thrS	3.32
	thymidine kinase	tdk	4.57
	thymidylate kinase	tmk	5.45
	translation-associated GTPase	gtp1	3.97
	translation initiation factor IF-2	infB	4.05
	triosephosphate isomerase	tpiA	5.96
	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase	trmU	4.24
	tRNA (guanine-n1)-methyltransferase	trmD	3.83
	tRNA modification GTPase TrmE	trmE	5.20
	tRNA uridine 5-carboxymethylaminomethyl modification enzyme GidA	gidA	3.88
	tryptophanyl-tRNA synthetase	trpS	4.45
	tyrosyl tRNA synthetase	tyrS	5.83
	uracil-DNA glycosylase	ung	6.41
	valyl-tRNA synthetase	valS	5.22

Supplementary Table 2.

Number of analysed orthologous genes between species.

Hominis	M. agalactiae	M. synoviae	M. pulmonis	M. mobile	M. hyopneumo	M. arthritidis
Mycoplasma agalactiae						
Mycoplasma synoviae	275					
Mycoplasma pulmonis	271	239				
Mycoplasma mobile	241	220	270			
Mycoplasma hyopneumoniae	234	191	267	208		
Mycoplasma arthritidis	239	217	234	218	199	

Pneumoniae	M. penetrans	U. urealyticum	M. gallisepticu	M. pneumoniae	M. genitalium
Mycoplasma penetrans					
Ureaplasma urealyticum	202				
Mycoplasma gallisepticum	181	148			
Mycoplasma pneumoniae	184	144	233		
Mycoplasma genitalium	173	140	194	460	

Spiroplasma	M. florum	M. capricolum	M. mycoide
Mesoplasma florum			
Mycoplasma capricolum	381		
Mycoplasma mycoide	382	640	

Phytoplasma-like	A. laidlawii	A. y. phytoplasm	O. y. phytoplasm	C. P. australiense	C. P. mali
Acholeplasma laidlawii					
Aster yellows phytoplasma	222				
Onion yelows phytoplasma	230	411			
C. Phytoplasma australiense	226	265	269		
C. Phytoplasma mali	213	346	270	262	

Correlation coeficients estimated between the position of each gene along the "expression" axis generated by COA-RAAU in each species and the estimated pairwise non-synonymous distance (dN).

Hominis	**Axis**	M. agalactiae	M. synoviae	M. pulmonis	M. mobile	M. hyopneumo	M. arthritidis
Mycoplasma agalactiae	1		0.59	0.56	0.58	0.59	0.53
Mycoplasma synoviae	1	0.54		0.56	0.54	0.55	0.56
Mycoplasma pulmonis	1	0.57	0.59		0.59	0.56	0.56
Mycoplasma mobile	2	0.64	0.57	0.57		0.64	0.62
Mycoplasma hyopneumoniae	1	0.62	0.64	0.61	0.69		0.61
Mycoplasma arthritidis	2	0.54	0.53	0.60	0.64	0.59	

Pneumoniae		M. penetrans	U. urealyticum	M. gallisepticu	M. pneumoniae	M. genitalium
Mycoplasma penetrans	2		0.67	0.58	0.56	0.51
Ureaplasma urealyticum	1	0.67		0.62	0.60	0.58
Mycoplasma gallisepticum	3	0.48	0.55		0.57	0.52
Mycoplasma pneumoniae	3	0.40	0.39	0.37		0.40
Mycoplasma genitalium	2	0.15	0.17	0.19	0.18	

Spiroplasma		M. florum	M. capricolum	M. mycoide
Mesop. florum	1		0.63	0.65
Mycoplasma capricolum	1	0.66		0.32
Mycoplasma mycoide	1	0.65	0.36	

Phytoplasma-like		A. laidlawii	A. y. phytoplasm	O. y. phytoplasm	C. P. australiense	C. P. mali
Acholeplasma laidlawii	2		0.54	0.59	0.58	0.50
Aster yellows phytoplasma	1	0.58		0.25	0.50	0.44
Onion yelows phytoplasma	1	0.58	0.23		0.44	0.42
C. Phytoplasma australiense	2	0.60	0.55	0.53		0.52
C. Phytoplasma mali	1	0.69	0.61	0.60	0.63	

Supplementary Table 3.

Relative amino acid usage bias (β) estimated for organisms employed at present study.

$\beta_{(REF)}$: bias estimated in the reference set.

$\beta_{(COA)}$: bias estimated in HEGs identified by COA-RAAU, taking the 10% of the genes at one extreme of the "expression" axis.

Organisms		β(REF)																			
		Phe	Leu	Ile	Met	Val	Ser	Pro	Thr	Ala	Tyr	His	Gln	Asn	Lys	Asp	Glu	Cys	Trp	Arg	Gly
Outgroup	Bacillus anthracis	-0.47	-0.29	-0.28	-0.10	0.22	-0.22	0.03	0.11	0.33	-0.38	-0.11	-0.19	-0.12	0.46	-0.18	-0.03	-0.15	-0.61	1.31	0.15
	Bacillus halodurans	-0.44	-0.31	-0.17	-0.07	0.21	-0.17	-0.05	-0.03	0.12	-0.25	-0.23	-0.27	0.16	0.83	-0.18	-0.04	-0.10	-0.68	0.74	0.12
	Enterococcus faecalis	-0.47	-0.36	-0.24	0.04	0.26	-0.13	0.07	-0.03	0.20	-0.30	0.18	-0.30	-0.09	0.57	-0.16	-0.07	-0.14	-0.61	1.15	0.15
	Lactobacillus johnsonii	-0.48	-0.31	-0.26	0.06	0.22	-0.11	0.03	-0.04	0.28	-0.33	0.13	-0.23	-0.26	0.43	-0.12	0.07	-0.21	-0.64	1.20	0.19
	Lactobacillus plantarum	-0.33	-0.32	-0.15	-0.10	0.20	-0.13	-0.11	-0.13	0.01	-0.25	-0.06	-0.32	-0.07	0.85	0.05	0.18	0.21	-0.54	0.81	0.13
	Lactococcus lactis	-0.45	-0.30	-0.32	0.02	0.37	-0.15	0.09	0.13	0.33	-0.35	0.14	-0.24	-0.22	0.42	-0.28	-0.05	-0.07	-0.61	1.16	0.16
	Listeria innocua	-0.46	-0.32	-0.28	-0.07	0.26	-0.09	0.11	-0.01	0.06	-0.28	0.21	-0.17	-0.18	0.50	-0.21	-0.08	0.11	-0.56	1.29	0.19
	Oceanobacillus iheyensis	-0.44	-0.28	-0.30	-0.06	0.22	-0.16	-0.03	0.01	0.17	-0.24	-0.04	-0.20	-0.10	0.68	-0.13	-0.05	0.07	-0.53	1.21	0.13
	Staphylococcus aureus	-0.46	-0.32	-0.34	-0.11	0.30	-0.13	0.09	0.10	0.23	-0.40	-0.21	-0.23	-0.20	0.49	-0.29	0.15	-0.04	-0.51	1.41	0.27
	Staphylococcus epidermidis	-0.42	-0.31	-0.37	-0.20	0.31	-0.11	0.13	0.10	0.27	-0.35	-0.24	-0.22	-0.19	0.44	-0.22	0.14	0.19	-0.46	1.35	0.26
	Streptococcus mutans	-0.44	-0.37	-0.26	-0.05	0.41	-0.22	0.14	0.05	0.30	-0.39	0.03	-0.23	-0.09	0.43	-0.21	0.04	-0.40	-0.61	1.07	0.21
	Streptococcus pyogenes	-0.39	-0.36	-0.23	-0.10	0.35	-0.17	0.04	0.00	0.28	-0.33	-0.07	-0.30	-0.09	0.53	-0.27	-0.04	-0.59	-0.52	1.13	0.22
	Clostridium perfringens	-0.38	-0.28	-0.38	0.17	0.37	-0.36	0.33	0.12	0.55	-0.42	0.59	0.37	-0.34	0.29	-0.30	-0.07	-0.08	-0.44	1.34	0.17
	Clostridium tetani	-0.33	-0.18	-0.34	0.00	0.46	-0.22	0.48	0.07	0.49	-0.34	-0.02	0.30	-0.36	0.09	-0.18	0.00	-0.43	-0.41	0.96	0.32
Mollicutes	Acholeplasma laidlawii	-0.43	-0.28	-0.33	-0.15	0.25	-0.23	0.16	0.10	0.54	-0.48	-0.02	0.17	-0.16	0.46	-0.34	-0.01	-0.20	-0.50	1.40	0.19
	Aster yellows phytoplasma	-0.47	-0.25	-0.29	0.16	0.76	-0.03	0.07	0.10	0.80	-0.42	0.09	-0.17	-0.33	0.16	-0.16	-0.08	-0.39	-0.38	1.68	0.58
	C. Phytoplasma australiense	-0.51	-0.21	-0.18	0.16	0.60	-0.03	-0.02	0.06	0.63	-0.35	0.15	-0.08	-0.33	0.18	-0.22	-0.11	-0.30	-0.43	1.28	0.53
	C. Phytoplasma mali	-0.50	-0.27	-0.14	0.21	0.43	0.01	0.06	0.08	0.54	-0.35	0.32	-0.13	-0.19	0.25	-0.27	-0.13	-0.39	-0.31	1.76	0.59
	Mesoplasma florum	-0.48	-0.24	-0.31	0.13	0.31	-0.21	0.21	0.10	0.55	-0.39	0.49	0.15	-0.31	0.32	-0.31	-0.11	0.18	-0.53	1.55	0.30
	Mycoplasma hyopneumoniae	-0.50	-0.23	-0.29	0.39	0.31	-0.14	0.05	0.32	0.52	-0.38	0.77	0.08	-0.30	0.29	-0.29	-0.13	0.17	-0.50	1.39	0.35
	Mycoplasma agalactiae	-0.40	-0.25	-0.25	0.23	0.24	-0.17	0.04	0.21	0.51	-0.46	0.41	0.23	-0.31	0.28	-0.37	-0.11	0.28	-0.57	1.30	0.43
	Mycoplasma arthritidis	-0.46	-0.23	-0.27	0.28	0.33	-0.19	0.15	0.27	0.38	-0.38	0.57	0.12	-0.35	0.31	-0.37	-0.10	0.48	-0.57	1.10	0.32
	Mycoplasma capricolum	-0.45	-0.28	-0.29	0.46	0.41	-0.17	0.19	0.11	0.80	-0.45	0.39	-0.01	-0.40	0.25	-0.36	-0.01	0.11	-0.52	1.75	0.56
	Mycoplasma gallisepticum	-0.47	-0.29	-0.20	0.39	0.26	-0.20	-0.03	0.02	0.41	-0.43	0.46	-0.07	-0.28	0.48	-0.29	0.02	0.26	-0.52	1.17	0.34
	Mycoplasma genitalium	-0.52	-0.21	-0.17	0.34	0.16	-0.18	0.03	0.14	0.17	-0.33	0.41	-0.06	-0.22	0.49	-0.31	-0.14	0.09	-0.44	1.30	0.30
	Mycoplasma mobile	-0.47	-0.27	-0.28	0.49	0.26	-0.21	0.14	0.11	0.48	-0.32	0.69	0.15	-0.35	0.40	-0.19	-0.10	0.93	-0.56	1.41	0.31
	Mycoplasma mycoide	-0.44	-0.27	-0.29	0.41	0.46	-0.21	0.22	0.11	0.91	-0.44	0.28	-0.03	-0.39	0.20	-0.32	0.07	-0.52	-0.52	1.49	0.67
	Mycoplasma penetrans	-0.50	-0.25	-0.27	0.21	0.24	-0.15	0.36	0.02	0.57	-0.38	0.52	0.12	-0.37	0.47	-0.34	-0.07	-0.05	-0.45	1.54	0.35
	Mycoplasma pneumoniae	-0.47	-0.20	-0.09	0.33	0.18	-0.20	-0.08	-0.02	0.14	-0.30	0.31	-0.18	-0.15	0.57	-0.26	-0.14	0.31	-0.51	1.10	0.12
	Mycoplasma pulmonis	-0.47	-0.27	-0.26	0.34	0.31	-0.10	0.11	0.22	0.51	-0.44	0.42	-0.07	-0.31	0.33	-0.30	-0.10	1.20	-0.56	1.43	0.41
	Mycoplasma synoviae	-0.46	-0.28	-0.38	0.44	0.38	-0.16	0.16	0.14	0.43	-0.44	0.91	0.03	-0.43	0.39	-0.32	-0.10	0.90	-0.41	1.49	0.38
	Onion yelows phytoplasma	-0.46	-0.25	-0.25	0.22	0.73	-0.05	0.07	0.09	0.68	-0.44	0.08	-0.20	-0.34	0.21	-0.19	-0.12	-0.36	-0.44	1.73	0.48
	Ureaplasma urealyticum	-0.52	-0.25	-0.39	0.25	0.47	0.01	0.17	0.20	0.74	-0.44	0.24	-0.11	-0.33	0.28	-0.39	-0.03	0.14	-0.46	1.57	0.53

Organisms		β(COA)																			
		Phe	Leu	Ile	Met	Val	Ser	Pro	Thr	Ala	Tyr	His	Gln	Asn	Lys	Asp	Glu	Cys	Trp	Arg	Gly
Mollicutes	Acholeplasma laidlawii	-0.32	-0.30	-0.27	-0.06	0.33	-0.19	0.20	0.11	0.88	-0.44	-0.20	0.07	-0.21	0.12	-0.13	0.04	-0.33	-0.26	0.51	0.41
	Aster yellows phytoplasma	-0.32	-0.17	-0.23	0.06	1.14	0.11	0.09	0.02	1.00	-0.40	-0.18	-0.29	-0.41	-0.14	0.04	-0.10	-0.02	-0.34	0.80	1.07
	C. Phytoplasma australiense	-0.42	-0.15	-0.20	0.17	0.87	0.00	0.20	0.08	1.10	-0.42	-0.06	-0.13	-0.36	-0.17	-0.05	-0.09	-0.30	-0.39	0.82	0.85
	C. Phytoplasma mali	-0.49	-0.26	-0.24	0.13	0.74	0.01	0.22	0.15	1.09	-0.31	0.07	-0.26	-0.32	0.04	-0.07	-0.08	-0.61	-0.04	1.70	0.85
	Mesoplasma florum	-0.44	-0.23	-0.23	0.26	0.43	-0.23	0.38	0.23	0.91	-0.40	0.20	-0.03	-0.42	-0.03	-0.09	-0.09	-0.32	-0.57	0.66	0.73
	Mycoplasma hyopneumoniae	-0.50	-0.25	-0.17	0.57	0.52	-0.12	0.18	0.22	0.88	-0.35	0.36	-0.01	-0.35	-0.05	-0.08	-0.04	-0.22	-0.33	0.79	0.74
	Mycoplasma agalactiae	-0.38	-0.18	-0.13	0.32	0.45	-0.22	0.17	0.20	0.64	-0.40	0.29	0.00	-0.42	-0.07	-0.26	-0.07	0.04	-0.39	0.65	0.93
	Mycoplasma arthritidis	-0.47	-0.30	-0.25	0.39	0.47	-0.23	0.61	0.19	0.46	-0.44	0.17	-0.03	-0.40	0.15	-0.17	0.04	0.00	-0.61	0.79	0.63
	Mycoplasma capricolum	-0.38	-0.28	-0.23	0.60	0.50	-0.14	0.58	0.20	1.14	-0.47	0.05	-0.09	-0.45	-0.02	-0.28	0.08	-0.05	-0.42	0.67	1.00
	Mycoplasma gallisepticum	-0.33	-0.18	-0.05	0.72	0.32	-0.24	0.16	0.02	0.43	-0.38	0.59	-0.25	-0.42	0.15	-0.26	0.00	1.05	-0.36	0.63	0.53
	Mycoplasma genitalium	-0.53	-0.26	-0.06	0.51	0.31	-0.15	0.19	0.10	0.39	-0.24	0.34	-0.29	-0.39	0.24	-0.20	0.01	0.07	-0.59	0.86	0.61
	Mycoplasma mobile	-0.46	-0.31	-0.28	0.70	0.38	-0.29	0.11	0.04	0.63	-0.22	0.89	0.03	-0.45	0.25	0.05	0.04	1.57	-0.47	0.92	0.57
	Mycoplasma mycoide	-0.26	-0.17	-0.15	0.67	0.54	-0.14	0.48	0.16	1.24	-0.43	0.02	-0.11	-0.44	-0.16	-0.28	-0.07	-0.25	-0.21	0.52	1.12
	Mycoplasma penetrans	-0.38	-0.16	-0.12	0.45	0.47	-0.25	0.53	-0.02	0.87	-0.44	0.18	-0.02	-0.43	0.09	-0.16	-0.04	0.07	-0.31	0.79	0.59
	Mycoplasma pneumoniae	-0.45	-0.23	-0.06	0.61	0.26	-0.22	0.74	-0.04	0.27	-0.31	0.42	-0.16	-0.32	0.08	-0.15	-0.03	0.68	-0.17	0.59	0.22
	Mycoplasma pulmonis	-0.48	-0.20	-0.22	0.59	0.42	-0.11	0.32	0.17	0.91	-0.38	0.17	0.04	-0.34	-0.07	-0.16	-0.12	0.24	-0.44	0.82	0.97
	Mycoplasma synoviae	-0.40	-0.22	-0.51	0.06	0.50	-0.05	0.58	0.71	0.73	-0.36	-0.23	0.25	-0.25	-0.19	-0.03	-0.15	-0.79	0.01	0.49	0.66
	Onion yelows phytoplasma	-0.14	-0.10	-0.23	0.24	1.02	0.10	0.06	-0.02	0.89	-0.34	-0.18	-0.39	-0.38	-0.14	0.00	-0.21	-0.05	-0.32	0.63	0.98
	Ureaplasma urealyticum	-0.39	-0.22	-0.32	0.39	0.54	0.01	0.36	0.20	0.96	-0.45	0.13	-0.14	-0.43	-0.02	-0.30	-0.08	-0.10	-0.17	1.15	0.93

8. Discusión I: Evolución del UCS en procariotas y eucariotas unicelulares

8.1 Sesgo mutacional, deriva, selección e inercia filogenética

El modelo más aceptado para explicar el patrón de variabilidad observado predice que el efecto de la selección operando sobre el uso de codones sinónimos es relativamente menor (Bulmer, 1991). A su vez establece que ciertos codones mayores u óptimos serán fijados por la selección, principalmente en genes de alta expresión donde el efecto a nivel celular es mayor. Por otro lado el sesgo mutacional y la deriva genética definirían la distribución de codones sinónimos en el resto de los genes (ver arriba).

Los resultados obtenidos en los Capítulos I y II confirman que la selección operando sobre el UCS está ampliamente distribuida y que los sesgos son más pronunciados en genes de alta expresión (Gouy & Gautier, 1982; Supek *et al.*, 2010); aunque notamos que existen algunos grupos que presentan sesgos selectivos muy menores o nulos, lo que sería más común en el caso de los parásitos o simbiontes (Iriarte *et al.*, 2011; Palacios & Wernegreen, 2002; Rispe *et al.*, 2004). Se cree que el sesgo mutacional extremo y los cuellos de botella recurrentes experimentados a nivel poblacional por estas especies son responsables de mitigar el efecto de la selección. Confirmamos las observaciones previamente reportadas, en donde ciertos linajes parecen reducir el sesgo como producto de su transición hacia nichos más específicos (Botzman & Margalit, 2011; Retchless & Lawrence, 2012; Sharp *et al.*, 2010; Sharp *et al.*, 2005; Supek *et al.*, 2010). Nuevamente, el ejemplo más aceptado es la transición que se observa al comparar especies de vida libre con especies emparentadas que han transitado hacia la simbiosis o el parasitismo.

Es interesante la comparación de las tendencias en especies que son altamente divergentes de sus ancestros (ej. *Buchnera, Blochmania* y *Wigglesworthia*) con aquellas especies también simbiontes que retienen muchas de las características de sus ancestros de vida libre. La familia Enterobacteriaceae presenta varios casos de convergencia hacia la simbiosis (Husnik *et al.*, 2011). Por ejemplo el caso de *Sodalis glossinidius vs*. *Wigglesworthia glossinidia*. Ambas especies son simbiontes de la mosca *tse-tse*. Nuestros resultados sugieren que *Sodalis* y *Wigglesworthia* no forman un grupo monofilético, sin embargo la posición de los simbiontes en la filogenia de la familia ha sido recientemente discutida (Husnik *et al.*, 2011; Philippe & Roure, 2011). Según estos trabajos, *Wigglesworthia* y *Sodalis* formarían parte de un linaje basal de la familia Enterobacteriaceae, en donde todas las especies son simbiontes (géneros *Sodalis*, *Baumannia*, *Blochmannia* y *Wigglesworthia*). Estos géneros y sus ancestros han evolucionado en simbiosis con insectos durante millones de años (Akman *et al.*, 2002; Banerjee *et al.*, 2004a). *Wigglesworthia* es usualmente presentado como ejemplo típico de endosimbiosis bacteriana. Las especies simbiontes han perdido gran parte de su genoma y de sus capacidades biosintéticas (Moran & Wernegreen, 2000). Junto con estas características a nivel genómico, presentan muchas otras, como son la alta tasa de evolución molecular o el sesgo composicional marcado hacia Adenina y Timina (A + T).

Nuestros resultados sugieren que, si bien estas especies tienen un nicho muy parecido, la historia filogenética de ambas condiciona los sesgos selectivos observados en la actualidad. Según trabajos recientes el género *Sodalis* presenta una posición basal dentro de su grupo de simbiontes (Husnik *et al.*, 2011), manteniendo muchas de las características de los parientes de vida libre.

El concepto de *inercia filogenética* hace referencia a la influencia que tiene un ancestro sobre un descendiente. La idea central es que existe un conjunto importante de características en los organismos que se mantienen desde el ancestro y que no necesariamente “los hacen mejores”. En evolución, el concepto es usualmente utilizado como una alternativa a la selección natural a la hora de explicar la persistencia de características que parecen ser sub-óptimos (Shanahan, 2011). Trabajos previos sugieren que el decaimiento del sesgo selectivo en el uso de codones es un proceso lento (Sharp *et al.*, 2010). Nuestros resultados, al estudiar la relación entre la distancia filogenética y el sesgo selectivo, sugieren que el sesgo tiene un componente de inercia filogenética importante, también la preferencia por algunos codones óptimos. En suma, nuestros resultados sugieren que este es un factor importante a considerar cuando ser analiza el sesgo selectivo en el uso de codones desde una perspectiva evolutiva-comparativa. En este caso la inercia filogenética explicaría en parte el alto grado conservación en la preferencia de codones óptimos observada en linajes que han experimentado cambios importantes en el *pool* de tRNAs, en los sesgos mutacionales o en las preferencias de hábitat (cambios del tipo eco-fisiológicos). Al mismo tiempo las propiedades de los sesgos analizados, relacionados e interdependientes con varios procesos a nivel celular y derivados de las preferencias de varios genes, podrían explicar también esta dinámica evolutiva.

8.2 El papel del *pool* de tRNAs sobre el sesgo selectivo en el UCS

Varias revisiones y trabajos publicados de forma relativamente reciente han abordado este tema y han aportado nuevas perspectivas sobre cómo la abundancia de los tRNAs afecta y es afectada por el UCS (Hershberg & Petrov, 2008; Higgs & Ran, 2008; Novoa *et al.*, 2012; Rocha, 2004; Sharp *et al.*, 2010; Supek *et al.*, 2010). Se ha prestado especial atención a los mecanismos de regulación y las modificaciones postraduccionales de las secuencias que conforman el anticodón. Sin embargo, la idea intuitiva y la comprobación del papel fundamental de estas moléculas en los sesgos selectivos en el UCS data de los años 80's (Bulmer, 1987; Ikemura, 1985; Yamao *et al.*, 1991).

Se han observado fuertes correlaciones entre la magnitud de los sesgos selectivos y el número de operones de RNAs ribosomales (rRNAs) y de genes de tRNAs, también entre estos sesgos y la tasa de duplicación poblacional o de crecimiento (Dong *et al.*, 1996; dos Reis *et al.*, 2004; Sharp *et al.*, 2010). En general los codones óptimos son reconocidos por los tRNAs isoaceptores más abundantes, los que muchas veces también presentan el anticodón con apareamiento canónico (Watson y Crick) (Ikemura, 1981, 1985; Kanaya *et al.*, 1999; Kanaya *et al.*, 2001; Yamao *et al.*, 1991). Todas estas observaciones difícilmente puedan explicarse en un marco neutralista. De la misma forma, las presiones mutacionales no pueden explicar cómo las modificaciones en las bases de tambaleo en el anticodón de los tRNA, mediados por la presencia de genes específicos en el genoma, se correlacionan con las variaciones en los sesgos selectivos (Roth, 2012). Esto es especialmente cierto en las familias de codones sinónimos de 2 y 4 miembros (duetos y cuartetos, respectivamente) (Ran & Higgs, 2010). La importancia relativa de los tRNAs en los sesgos ha sido recientemente discutido (Satapathy *et al.*, 2012).

Nuestros resultados sugieren que efectivamente las preferencias de los genes de alta expresión y de las regiones conservadas tienden a coincidir, en gran medida, con los tRNA isoaceptores más abundantes. También encontramos que las reglas de apareamiento por tambaleo (*wobbling rules*) ayudan a explicar las correlaciones encontradas (Crick, 1966). De esta forma nuestros resultados van en línea con las hipótesis más aceptadas sobre este tema. Sin embargo, también sugieren que en algunos organismos, de forma más pronunciada que en otros, y para algunas familias de tripletes, existe un "desacoplamiento" entre los codones óptimos y los anticodones "presentes" en el genoma (los genes de tRNAs con anticodones predichos) (Kahali *et al.*, 2008). Varias explicaciones se han propuesto para este desacoplamiento aparente (Planteadas en los Capítulos I y II) (Satapathy *et al.*, 2012; Supek *et al.*, 2010; Withers *et al.*, 2006). En el trabajo desarrollado en el Capítulo I y en la sección "Sesgo mutacional, deriva, selección e inercia filogenética", se plantea la posibilidad de que las preferencias observadas en los codones óptimos sean en realidad el reflejo de una adaptación a un *pool* ancestral y conservado de tRNAs (Withers *et al.*, 2006). Esto explicaría la alta conservación observada en las preferencias de los genes de alta expresión, en presencia de cambios importantes en el contenido de tRNAs e incluso de importantes cambios en los nichos ecológicos. También apoyaría la idea de que los cambios en los codones óptimos, los cambios en los patrones generados por los sesgos selectivos, se dan de forma lenta (Sharp *et al.*, 2010; Shields, 1990). No es claro cómo ocurren los cambios en los sesgos selectivos en la evolución ni qué determina la existencia de determinados codones óptimos (Hershberg & Petrov, 2008). Si bien existen algunas propuestas al respecto (Bulmer, 1987; Dong *et al.*, 1996; Hershberg & Petrov, 2009; Ikemura, 1985; Shields, 1990), la más aceptada predice una co-evolución con los tRNAs presentes en el genoma (Shields, 1990), hacia un estado estable (Higgs & Ran, 2008) (ver más arriba).

Desde una perspectiva evolutiva, la presencia de un conjunto de genes de tRNAs conservado y presente desde un ancestro común, podría favorecer la existencia de un estado estable mantenido por selección todo a lo largo de la evolución de un grupo filogenético. Esto ya se ha sugerido para el caso de Mollicutes, un grupo que por su estilo de vida particular se caracteriza por cambios en las dinámicas poblacionales, lo que ha su vez genera cambios importantes en el efecto de la deriva y por consiguiente en la magnitud del sesgo selectivo a lo largo de la evolución (Iriarte *et al.*, 2011). Si los cambios en los codones óptimos se dan luego de un proceso de reducción y posterior aumento del efecto de la selección (Hershberg & Petrov, 2008; Sharp *et al.*, 2010) y si asumimos que la población de tRNAs condiciona la elección de los codones óptimos, las preferencias por ciertos codones mayores u óptimos no cambiarán mientras se mantenga un núcleo conservado y estable de tRNAs. De esta forma, los codones óptimos reflejarían mejor este conjunto de tRNAs (ancestral, conservado y estable) independientemente de los cambios recientes en el número de genes. En este sentido la aproximación tomada en los trabajos aquí presentados, el estudio de la evolución de los sesgos y de los codones óptimos en un marco filogenético, permitió tener una idea de las dinámicas evolutivas de los caracteres reconstruidos y evaluar el impacto de la inercia filogenética sobre ellos (ver arriba).

8.3 Velocidad *vs.* Fidelidad en la traducción

Como ya se planteó en la introducción, la velocidad y la fidelidad de la traducción son los componentes del sesgo selectivo más ampliamente estudiados y más aceptados (Hershberg & Petrov, 2008; Sharp *et al.*, 2010; Supek *et al.*, 2010). Los efectos esperados sobre el sesgo en el UCS de cada una de estas fuerzas están claramente planteados. Hay que tener en cuenta que recientemente se ha demostrado que otros factores pueden jugar un papel importante a la hora de generar y explicar algunas observaciones en el uso de codones sinónimos, en ciertos genes y en ciertos organismos (Ej. ordenamiento de codones, cambios en la expresión de los tRNAs en el ciclo celular, preferencia por codones "lentos o no óptimos", etc.) (Ver sección 1 "Introducción al Uso de Codones Sinónimos y Aminoácidos").

En breve, la razón fundamental por la que la selección operaría sobre el UCS es porque al incrementar el uso de codones óptimos en una secuencia se mejora la eficiencia (mayor velocidad y liberación de ribosomas) y la exactitud de la traducción (Hershberg & Petrov, 2008).

Ahora bien, no es clara la contribución relativa de la selección actuando a nivel de la velocidad (también se utiliza el término eficiencia) y de la fidelidad (Gingold & Pilpel, 2011; Hershberg & Petrov, 2008; Ran & Higgs, 2012; Sharp *et al.*, 2010). La observación de que la variación en el sesgo selectivo en el UCS entre las bacterias se correlaciona con la tasa de crecimiento apoya la idea de un rol protagónico para la velocidad (Sharp *et al.*, 2010). En el otro lado tenemos los trabajos que proponen que la fidelidad juega un rol muy importante porque sugiere que los codones no óptimos tienden a favorecer la formación de agregados tóxicos, que resultan de producir malos plegamientos en las proteínas codificadas (Drummond & Wilke, 2008). En el mismo sentido, se plantea que la selección en la fidelidad es más importante porque aparece de forma independiente de la expresión y porque existen patrones en el UCS que parecen evitar la codificación de codones *stop* prematuros (Stoletzki & Eyre-Walker, 2007).

Los trabajos desarrollados durante la tesis abordan esta pregunta fundamental y desembocan en algunos resultados interesantes. Intentamos cuantificar el sesgo selectivo y localizar el efecto de la selección sobre el UCS operando en estos dos niveles. Estimamos los sesgos específicos de cada codón sinónimo, para cada gen, dentro de los distintos genomas analizados. Como está claramente establecido en los manuscritos presentados en los Capítulos I y II, los resultados muestran que la selección opera simultáneamente en ambos niveles aunque la magnitud de los sesgos experimentados por los genes de alta expresión (con respecto al resto de los genes del genoma) (velocidad) es más importante que los que se estiman para las regiones conservadas (respecto a las no conservadas) (fidelidad) (Ran & Higgs, 2012; Sharp *et al.*, 2010).

Los análisis multivariados muestran que la expresión es el factor más importante a la hora de explicar la variabilidad intra-genómica. Dentro de los genes de alta expresión, los sesgos a favor de la velocidad, que operarían todo a lo largo del gen, podrían homogeneizar el sesgo hacia codones óptimos, reduciendo las diferencias entre las regiones conservadas respecto a las no conservadas. Es importante tener en cuenta que los genes de alta expresión, son en muchos casos, al mismo tiempo, los más conservados [ver sección 1.2 "Uso de Aminoácidos (UAA)"] . Por lo tanto, una parte del sesgo que se estima en los genes de alta expresión (con respecto al resto de los genes) incluiría el sesgo selectivo operando a nivel de la fidelidad. Pensamos que esta contribución sería relativamente menor dado que incluso en los genes de baja expresión existe un sesgo selectivo operando también a nivel de la fidelidad, por lo que en parte se cancelarían. En resumen, creemos que la inercia de la variabilidad capturada en los ejes principales de los análisis multivariados corresponde principalmente a un efecto de la selección operando a nivel de la velocidad.

Encontramos que en el caso del género *Aspergillus*, los genes de alta expresión muestran una utilización mayor de codones óptimos en las regiones conservadas en comparación con los genes de baja expresión. Este resultado confirma las observaciones previamente reportadas para algunos organismos modelos eucariotas y procariotas (Drummond & Wilke, 2008). Estos autores sostienen que la selección operando para evitar el mal plegamiento de las proteínas sería más importante en los genes de alta expresión. Sin embargo estas diferencias no parecen ser tan claras al comparar los grupos de genes con distintos niveles de expresión en Enterobacteriaceae.

Los resultados también muestran, claramente en el caso de Enterobacteriaceae, que existen diferencias en los codones significativamente más utilizados al comparar las regiones conservadas y los genes de alta expresión, fenómeno que para algunos codones está totalmente conservado en las distintas especies de esta familia. Podría deducirse entonces que, para algunos grupos de codones sinónimos, los codones que se traducen más rápidamente (óptimos para velocidad) no son los que producen menos errores (óptimos para fidelidad). Obviamente esta afirmación debería ser comprobada experimentalmente. Resulta también interesante que, en algunos casos, los codones óptimos para fidelidad resultan más conservados que los de velocidad.

9. Discusión II: UCS en virus de RNA

En el Capítulo III se presentan dos estudios sobre el UCS en especies de virus con genoma de RNA. Por un lado se trabaja con el virus de Influenza A (IAV); este virus es miembro de la familia Orthomyxoviridae, contiene un genoma compuesto por ocho segmentos de RNA de hebra simple de polaridad negativa. El segundo, el Virus del Nilo Occidental o *West Nile virus* (WNV), es miembro de la familia Flaviviridae y su genoma esta compuesto por una molécula simple de RNA de polaridad positiva con un largo de 11.000 bases.

En términos generales encontramos que ambas especies se caracterizan por presentar sesgos globales moderados en el UCS, sin embargo es claro que algunos codones particulares están desviados de los valores esperados. Los sesgos observados en estos codones sinónimos parece relacionarse con la presencia de algunos dinucleótidos en el triplete, los que a su vez están sub- o sobrerrepresentados. Las desviaciones en la frecuencia de UpA y CpG y sus efectos sobre el UCS ya han sido descriptos para los virus de RNA (Rima & McFerran, 1997) y para otras secuencias de organismos procariotas y eucariotas (Gentles & Karlin, 2001; Karlin, 1998). Este efecto es patente en la 2da y 3ra posición del codón, pero también en la 1ra y 2da y la 3ra y 1ra. Esto sugiere que hay un sesgo en el UCS en relación al siguiente codón e incluso un efecto sobre el uso de aminoácidos. Se ha propuesto que la distribución observada se debe a la mutabilidad de estos dinucleótidos, aunque se ha planteado que también la selección puede operar para mantener su frecuencia baja (Al-Saif & Khabar, 2012; Gaffney & Keightley, 2008; Rima & McFerran, 1997). En el caso particular de los virus de RNA, la reducción en estos dinucleótidos podría vincularse con la evasión de la respuesta inmune, por ejemplo la estabilidad de la molécula de RNA frente a endonucleasas antivirales (Dorn & Kippenberger, 2008; Washenberger *et al.*, 2007). Es interesante que algunos genomas presenten una subrepresentación de CpG y otros de CpG y UpA (Rima & McFerran, 1997), esto podría vincularse con la estrategia del virus, aunque no se ha estudiado en profundidad. En este contexto, el UCS asociado a la frecuencia de dinucleótidos sería una consecuencia de esto último y no a la inversa.

Al analizar el UCS en relación al *pool* de tRNAs del hospedero, encontramos que no hay una adaptación en el sentido clásico, es decir, no encontramos que los codones sinónimos más usados sean los reconocidos por los tRNA isoaceptores más abundantes. Esto se muestra en el caso de IAV. Es interesante notar que, salvo para el aminoácido Histidina, el codón sin anticodon específico en el genoma (el codón sinónimo que no tiene un tRNA que lo reconozca por apareamiento canónico) nunca es preferido. Este resultado va en línea con un trabajo reciente que sugiere que IAV redirige la expresión de tRNAs del hospedero en vez de ajustar su UCS al mismo (Pavon-Eternod *et al.*, 2013). El virus podría "preferir" ciertos tripletes por sobre otros independientemente de la distribución de tRNAs en la célula, "elegir" entre cualquiera de los tRNAs codificados en el genoma. Es posible aumentar la expresión de un gen de tRNA con un anticodón específico, siempre y cuando este exista en el genoma.

En relación a los genes del huésped, encontramos que en ambos virus el UCS se encuentran en el promedio de los genes humanos. Ahora nos preguntamos si esta falta de sesgo puede ser mantenido por la selección. Según un trabajo reciente la estrategia "óptima", especialmente en eucariotas, incluiría un uso proporcional de acuerdo a la disponibilidad de tRNAs en la célula (Qian *et al.*, 2012). Esta disponibilidad depende de cuánto se use un codón en relación a la cantidad de tRNA. Según los autores un uso "balanceado" de codones podría ser mantenido por selección. Si bien es una propuesta interesante, otros análisis son necesarios para probar esta hipótesis en relación al UCS en virus.

Al comparar entre cepas encontramos que la variabilidad en el UCS es relativamente menor. Los análisis multivariados sugieren que el uso de ciertos codones poco frecuentes explica la mayor parte de la variabilidad observada. Los codones poco frecuentes son principalmente aquellos que presentan los dinucleótidos CpG y UpA. Es posible que pequeñas diferencias en el uso de estos codones generen diferencias fenotípicas entre las cepas, nuevamente el UCS sería una consecuencia de la composición y no la causa de las diferencias entre las cepas. En el caso particular de WNV no se encontró una asociación clara entre las cepas y hospedero del cual se realizó el aislamiento.

10. Discusión III: Fuerzas evolutivas que afectan la evolución del sesgo en el Uso de Aminoácidos (UAA)

10.1 Propiedades fisicoquímicas de los aminoácidos, el código genético y la conservación diferencial de los genes

Las propiedades fisicoquímicas y biosintéticas de los aminoácidos parecen jugar un papel importante en la organización y evolución del código genético. Las hipótesis del origen del código genético consideran, de una u otra forma, fundamentales estas propiedades (Di Giulio, 1997; Szathmary, 1993; Wong, 1975). Varias revisiones se han hecho sobre el tema (Chechetkin & Lobzin, 2011; Di Giulio & Medugno, 1998). La polaridad, los volúmenes moleculares y la hidrofobicidad de los aminoácidos están estrechamente relacionados con la posición de los tripletes codificantes respectivos en el código genético (D'Onofrio *et al.*, 1999; Di Giulio, 1989; Houen, 1999). De esta forma se acepta que la organización del código, y por tanto la composición nucleotídicas en las distintas posiciones de los tripletes (principalmente 2da y 1ra posición del codón), se asocian a las propiedades de los aminoácidos codificados.

El sesgo composicional global, usualmente medido como el contenido relativo de Guanina más Citosina (G + C) en la secuencia del genoma varia significativamente de una especie a otra (aprox. del 25% al 75%) (Bernardi, 1986; Muto & Osawa, 1987). Estas observaciones han sido confirmadas en la actualidad por la secuenciación del genoma completo de cientos de organismos, procariotas y eucariotas. Algunos trabajos líderes en la temática demostraron que la evolución del proteoma es afectada por los sesgos composicionales globales del genoma (Sueoka, 1961). Existe una correlación significativa entre la composición nucleotídica (contenido de G + C) del genoma y el UAA del proteoma de distintas especies (Bernardi, 1986; D'Onofrio *et al.*, 1991; Lafay *et al.*, 1999; Sueoka, 1961). El sesgo composicional, por lo tanto, determina la frecuencia de ciertos aminoácidos en el proteoma tanto en procariotas como en eucariotas (Lightfield *et al.*, 2011; Singer & Hickey, 2000). Esto puede interpretarse como un proceso mutuamente dependiente. Por un lado los sesgos mutacionales afectan el uso de aminoácidos global (y también el UCS). Paralelamente las restricciones funcionales, producto de la relación de las propiedades de ciertos aminoácidos con la composición nucleotídica de sus codones (ver arriba), limitan los efectos del sesgo mutacional, principalmente en la 2da y 1ra posición del codón. En otras palabras, la necesidad de incluir en el proteoma el uso de ciertos aminoácidos restringe el contenido de G + C mínimo y máximo en la 2da y 1ra posición del codón, debido a su situación específica en el código genético. Esto puede observarse claramente al comparar la variación composicional de las bases a nivel intra e inter-genómico en las tres posiciones del codón.

Las funciones de las proteínas y el nivel de expresión de las mismas restringirían la evolución de las secuencias a nivel aminoacídico (y en el uso de codones sinónimos, ver arriba). Como consecuencia, y en relación a lo antes mencionado, se espera que los genes de alta expresión (que a su vez son los más conservados a nivel sinónimo y no sinónimo) respondan de forma menor a los cambios ocasionados por el sesgo composicional. Lo mismo ocurriría con las proteínas cuyas funciones son específicas, lo que condicionaría su uso de aminoácidos (ej. proteínas de membrana). La variación en el UAA en los distintos genes dentro de un genoma ha sido estudiada en una variedad de organismos mediante varios métodos. Una de las ideas más interesantes al respecto, es la que asocia las restricciones energéticas con la evolución de las proteínas, hipótesis de minimización del costo.

10.2 Hipótesis de la minimización del costo

La frecuencia de aminoácidos en la composición de las proteínas se correlaciona negativamente con el peso molecular (Barrai *et al.*, 1995), lo que sugiere que el sesgo en el uso de aminoácidos estaría relacionado con una minimización del costo de la producción de proteínas (Dufton, 1997). Se corroboró que en *E. coli* y *Bacillus subtilis* el uso de aminoácidos energéticamente costosos se relaciona inversamente con el nivel de expresión de las proteínas (Akashi & Gojobori, 2002). Posteriormente se extendió esta observación para otros organismos optimizando la estimación del costo y más recientemente considerando las auxotrofias (Raiford *et al.*, 2012; Seligmann, 2003). Mediante análisis de correspondencias se ha comprobado que uno de los factores principales a la hora de explicar la variación en el UAA a nivel intra-genómico es el nivel de expresión (Lobry & Gautier, 1994). Esto se ha comprobado para un gran conjunto de organismos, incluido algunas especies de parásitos. El factor "expresión" no siempre es el más importante, aunque en la gran mayoría de especies analizadas se correlaciona con el peso molecular (Chauhan *et al.*, 2011; Kahali *et al.*, 2007; Rispe *et al.*, 2004; Schaber *et al.*, 2005; Seligmann, 2003; Zavala *et al.*, 2002). Es decir, existe un sesgo selectivo en el UAA a favor de aquello aminoácidos más baratos. Debido al número de proteínas sintetizadas en la célula el efecto debería ser más pronunciado en genes de alta expresión, siendo éste uno de los factores principales a la hora de explicar la variabilidad intra-genómica en el UAA.

Además de esto, se enumeran otros resultados observados que complementarían o apoyarían la "hipótesis de minimización del costo": *i*) a igual costo, se evitan más aminoácidos que tienen poco impacto en la estructura de las proteínas, esto sugiere que las restricciones funcionales predominan sobre las restricciones de economía, *ii*) la tasa de evolución de las proteínas se correlaciona negativamente con el peso molecular promedio de dichas proteínas, *iii*) el peso molecular promedio se reduciría al aumentar el tamaño genómico (Seligmann, 2003). El mismo autor propone que en los organismos parásitos los sesgos son menos pronunciados que en los organismos de vida libre y que este resultado iría a favor de la hipótesis de minimización del costo.

El artículo presentado en el Capítulo IV discute esta posibilidad y establece una relación (por lo menos para los Firmicutes) entre el contenido de GC de los genomas y la diferencia interna en el peso molecular.

10.3 El caso particular de los parásitos y simbiontes

Los parásitos y simbiontes presentan particulares dinámicas poblacionales, en parte como consecuencia de ellas tendríamos un patrón de convergencia hacia ciertas propiedades típicas de estos genomas (sesgo mutacional pronunciado hacia A + T, baja variabilidad genética, reducción en el tamaño del genoma, alta tasa de evolución molecular, reducción de las vías biosintéticas, entre otras). En el caso de las bacterias simbiontes o parásitas estos patrones son muy claros (Moran & Wernegreen, 2000). Estas dinámicas deberían reducir el efecto de la selección operando sobre un conjunto importante de caracteres, dándole un papel mayor a la deriva genética y al sesgo mutacional. En particular hemos estudiado el caso del sesgo selectivo en el uso de aminoácidos en un conjunto de organismos de este tipo. El papel del sesgo selectivo a este nivel es discutido para parásitos y simbiontes (Palacios & Wernegreen, 2002; Rispe *et al.*, 2004; Schaber *et al.*, 2005). El problema es que resulta difícil separar, si es que no son parte del mismo proceso, el efecto del enriquecimiento diferencial en Adenina y Timina (A + T) en los distintos genes del genoma, del efecto producto del sesgo selectivo.

Es decir, debido al sesgo mutacional intenso que sufren estos genomas, la mayoría de los genes incrementan su contenido de A + T. Debido a las restricciones impuestas por el código genético esto lleva necesariamente a un aumento del costo promedio de las proteínas y también a un aumento de la aromaticidad promedio. Recordar que los residuos aromáticos (Phe, Tyr y Trp) están dentro de los más "caros" y son codificados por tripletes ricos en A y T. En nuestro trabajo sugerimos que los genes de baja expresión han sido a lo largo de la evolución del grupo menos conservados a nivel no-sinónimo, por lo tanto debido a la alta tasa de evolución molecular y al efecto de la deriva no hay posibilidad de que estos genes fijen cambios compensatorios para minimizar el efecto del sesgo mutacional sobre el costo energético. Los genes de alta expresión, por otro lado, están bajo un mayor efecto de la selección purificadora operando para mantener su función y/o el bajo costo de sus proteínas y como consecuencia resisten este enriquecimiento. Como resultado se incrementa la variabilidad interna asociada al nivel de expresión, en el uso de ciertos aminoácidos (aquellos codificados por tripletes ricos en A + T, aromáticos y energéticamente caros). Observamos que la diferencia en el peso molecular entre los genes de alta y baja expresión aumenta con el sesgo composicional hacia A + T, tendencia que es más pronunciada en los organismos parásitos.

El análisis multivariado sirvió para determinar qué factores explican la mayor parte de la variabilidad interna en el uso de aminoácidos. Paralelamente analizamos la evolución de las tendencias a lo largo de la clase Mollicutes para intentar separar los efectos de la aromaticidad, el costo y el contenido de GC de los tripletes. Encontramos que en estas bacterias el nivel de expresión es el primer o segundo factor, al contrario de los parientes de vida libre estudiados, en donde el nivel de expresión nunca aparece asociado a uno de los dos primeros factores. Como se estableció en el manuscrito, esto se observa también al analizar la bibliografía y al comparar los resultados de organismos de vida libre y parásitos o endosimbiontes, eucariotas y procariotas. Notar, como lo muestran nuestros resultados, que esto no implica que el sesgo hacia aminoácidos menos costosos por parte de los genes de alta expresión en organismos de vida libre no exista, sino que no aparece como un factor importante a la hora de explicar la variabilidad interna.

La aproximación utilizada permitió analizar la evolución de los sesgos. Se intentó contraponer las hipótesis alternativas para explicar el sesgo intra-genómico en los genes de alta expresión en los organismos parásitos. En este sentido nuestros resultados mostraron que en general las proteínas de estos organismos a lo largo de la evolución se han caracterizado por acumular cambios aumentando el peso molecular promedio y la aromaticidad, como “obligaría” el sesgo mutacional. Esto lleva a que los organismos de vida libre presenten en promedio valores más bajos en el peso molecular promedio de sus proteomas (Seligmann, 2003). Esto de alguna forma podría asociarse al hecho de que estos organismos toman todos sus aminoácidos del hospedero, sin embargo es importante considerar que aumentar el costo para el huésped puede nos ser la mejor estrategia para el parásito. Este concepto ha sido discutido en profundidad y va más allá de los objetivos de esta tesis; sólo a modo de ejemplo ver (Alizon & van Baalen, 2008; Alizon & Michalakis, 2011; Smith, 2011; Williams, 2012).

Observamos a nivel global que las tendencias hacia el aumento del peso molecular (como estimador del costo) y la aromaticidad son dependientes. Además, estas tendencias son siempre mucho menos pronunciadas en los genes de alta expresión, lo que va a favor de la hipótesis de minimización del costo.

Por otro lado, al observar el uso particular de cada uno de los aminoácidos notamos que no es posible explicar los patrones observados en los genes de alta expresión solamente por selección a favor de minimización del costo, enriquecimiento diferencial de A + T o la resistencia al aumento de la aromaticidad. Ninguno por si solo puede explicar todas las tendencias. Por ejemplo, en particular el uso de residuos aromáticos parece ser en gran parte responsable del sesgo en el costo. Esto abre la posibilidad de considerar además del costo, otras propiedades de estos residuos que podrían ser contra-seleccionadas en los genes de alta expresión.

Finalmente, los cambios eco-fisiológicos de algunos linajes dentro del grupo no parecen generar diferencias en el patrón general encontrado, tampoco el considerar las transferencias horizontales. El análisis de *clustering*, parece indicar que el sesgo en los genes de alta expresión sigue un patrón filogenético con algunas pocas excepciones. En este sentido, el análisis estadístico muestra que existen diferencias significativas en la magnitud del sesgo para la mayoría de los residuos, al comparar entre Firmicutes de vida libre y parásitos. También se encuentran algunas diferencias menores entre parásitos del grupo SEM y del grupo AAA. Dado que los Firmicutes de vida libre, los parásitos del grupo SEM y los del grupo AAA forman grupos naturales, no es posible estudiar el efecto de la inercia filogenética de forma independiente del producto de la selección.

Queda también por explicar por qué en el grupo “hermano” de los Mollicutes, es decir los Firmicutes de vida libre, la expresión no aparece como un factor principal a la hora de explicar la variabilidad interna, ni siquiera en los organismos de vida libre que presentan un contenido bajo de G + C, en el rango de los parásitos. La causa más probable es que la diferencia en el peso molecular promedio entre los genes de alta y baja expresión en el uso de aminoácidos pesados es mayor en Mollicutes (ver arriba). Esto se ve claramente en la comparación de los pesos moleculares en el set de referencia y el proteoma completo para las especies del género *Clostridium* y *Micoplasma hyopneumoniae* o *Micoplasma sinoviae* en el artículo presentado en el

Capítulo IV, tabla 1. Es importante tomar en cuenta que estas especies presentan un GC total y GC_3 muy similar.

En este sentido los genes de baja expresión pueden ser la clave (ver en Tabla 1, Capítulo IV las diferencias entre Mollicutes y Firmicutes de vida libre en el peso molecular promedio del set de referencia y del resto del proteoma). En organismos de vida libre, frente a un sesgo composicional hacia A + T, los genes de baja expresión fijarían cambios compensatorios que reducirían el efecto mutacional sobre el costo energético de utilizar aminoácidos más caros, lo que a su vez mantiene la variabilidad asociada a la expresión (la diferencia en el peso molecular promedio entre genes de alta y baja expresión), en niveles bajos. Como se ha dicho más arriba, frente al mismo sesgo y por efecto de la deriva (cuellos de botella recursivos) los genes de baja expresión en organismos endosimbiontes o parásitos fijarían muchos más cambios hacia residuos energéticamente costosos, aumentando la variabilidad interna asociada a la expresión. Esto no implica que otros aspectos, como ser la reducción del número de genes (de algunos genes en particular) podrían relacionarse con el patrón observado en los Mollicutes.

11. Reconstrucción de estados ancestrales

La reconstrucción de estados ancestrales se ha convertido en una herramienta estándar para entender los procesos que moldean la evolución (Pagel, 1999). La idea simple detrás del procedimiento es estimar los estados (o valores) de caracteres en los nodos internos de un árbol filogenético basado en los estados de los mismos caracteres en las especies actuales y las relaciones filogenéticas entre las mismas (incluyendo la topología y las distancias de las ramas). Esta aproximación se ha utilizado en una variedad de datos: biogeográficos, ecológicos, vías metabólicas, secuencias, comportamientos, morfológicos, entre muchos otros (Litsios & Salamin, 2012; Ronquist, 2004).

Varios factores relacionados a la exactitud del procedimiento han sido investigados, particularmente focalizando en el marco estadístico (Ekman *et al.*, 2008) y la selección del modelo de evolución (Cunningham *et al.*, 1998). Otros elementos también estudiados incluyen la topología del árbol filogenético y el muestreo de taxones. Considerando todos estos aspectos y otros no incluidos, parece haber un consenso general con la idea de que las aproximaciones desarrolladas hasta el momento todavía deben ser mejoradas (Litsios & Salamin, 2012). Actualmente los métodos más utilizados son la Máxima Verosimilitud (ML o *Maximum Likelihood*) y el Bayesiano (BY). En base a revisiones bibliográficas hemos elegido los mejores modelos y métodos. Como estrategia general hemos decidido contraponer los resultados de los métodos y modelos que parecen ser igualmente confiables.

En el Capítulo IV hemos implementado una reconstrucción de secuencias de proteínas ancestrales exclusivamente por el método de ML, con un modelo relativamente complejo, REV (nr=189). El modelo REV, modelo general reversible de las sustituciones aminoacídicas, establece que la probabilidad de cambio entre un aminoácido *a* a uno *b* es igual que de uno *b* a uno *a*. Por lo tanto el numero de parámetros libres se reduce a 189 (190 – 1) o 208 ([190 – 1] + [20 – 1]) si se consideran como parámetros la frecuencia de cada aminoácido. La probabilidad de cambios recíprocos para cada par de aminoácidos se estima por Máxima Verosimilitud. Este es un modelo rico en parámetros y puede compararse con modelos más simples, como el alternativo REV0 (que elimina la posibilidad de que existan cambios entre aminoácidos en que sus codones codificantes se diferencian en más de un nucleótido), reduciendo significativamente el número de parámetros (Yang, 2007; Yang *et al.*, 1998). Los resultados del Capítulo IV (generados a partir del modelo REV (nr = 189) fueron comparados con los generados con el modelo REV0, sin diferencias significativas (datos no mostrados).

Para analizar la evolución de los codones óptimos, en el Capítulo I, hemos establecido la preferencias de los genes de alta expresión de una forma simple y confiable. Es decir los estados de cada codón sinónimo se codificaron como caracteres binarios (existe o no existe una preferencia significativa ($\chi 2$, $p < 0{,}01$). También se estimaron los valores de los coeficientes de selección ancestrales. En ambos casos se implementó con preferencia, pero no en exclusividad, el método BY, tanto para los estados continuos como para los binarios.

Un importante paso en la investigación MCMC (métodos *Markov chain Monte Carlo* para muestrear distribuciones de probabilidades) fue el desarrollo de los métodos de salto reversible (RJMCMC) (Green, 1995). Esto permite explorar los modelos alternativos, pudiendo variar en el número de parámetros (dimensiones del modelo) y en el valor de los mismos. En otras palabras, se exploran distintos modelos y se calcula la probabilidad de cada uno. Este método es implementado para caracteres binarios en el programa *BayesTraits* (Pagel *et al.*, 2004). En el caso de los codones óptimos esto es muy importante, ya que no se conoce con claridad las dinámicas evolutivas que gobiernan los cambios en estas preferencias (Hershberg & Petrov, 2008; Sharp *et al.*, 2010). Debido a esto no es posible establecer *a priori* ni un modelo evolutivo, ni los parámetros para hacer la reconstrucción. El coeficiente de selección operando sobre el uso de codones fue también reconstruido sobre el árbol filogenético. En este caso se utilizó también el programa *BayesTaits*. Se generaron modelos y distribuciones posteriores de parámetros de modelos alternativos, de los más simples a los más complejos. El test de factor Bayesiano (o

Bayes Factor test) se utilizó para elegir el mejor modelo con el menor número de parámetros. Paralelamente se desarrolló el mismo análisis utilizando el método ML implementando el modelo Ornstein-Uhlenbeck (OU) (Butler & King, 2004) en R (R Development Core Team, 2012).

La suma de estos procedimientos para analizar las tendencias de los sesgos selectivos, tanto en el UCS, como en el UAA, es novedosa. Si bien es importante aclarar que la reconstrucción de proteínas ancestrales para inferir propiedades de los organismos ancestrales (más allá de las secuencias *per se*) ya ha sido aplicada en varias trabajos (Gaucher *et al.*, 2008; Perez-Jimenez *et al.*, 2011).

12. Conclusiones generales

i) El sesgo selectivo en el UCS y UAA está ampliamente distribuido en procariotas y eucariotas unicelulares, incluso en organismos simbiontes o que han experimentado importantes sesgos mutacionales globales. La magnitud de los sesgos operando a nivel de la velocidad parece ser mayor y restringida fundamentalmente a los genes de alta expresión. Los sesgos en las regiones conservadas, respecto a las regiones no conservadas, son menores que los producidos por el efecto de la velocidad, pero están distribuidos en gran cantidad de genes, incluso en algunos que presentan bajos niveles de expresión.

ii) En los virus analizados existe un sesgo moderado en el UCS. Sin embargo, algunos tripletes parecen ser significativamente menos utilizados que los sinónimos, lo que podría ser consecuencia de una presión selectiva para evitar la presencia de algunos dinucleótidos que generarían una reacción de respuesta a nivel celular. La variación entre las cepas tiende a ser menor y vinculada al uso de codones poco frecuentes.

iii) La historia evolutiva parece jugar un factor importante a la hora de explicar las diferencias en la magnitud de los sesgos selectivos, al comparar entre especies que presentan nichos o sesgos composicionales similares.

iv) La aproximación de reconstrucción de estados ancestrales aplicados a estos problemas es novedosa y permite estudiar algunos aspectos importantes de las dinámicas evolutivas de los sesgos selectivos y las tendencias en la evolución de los grupos.

v) Los codones óptimos parecen reflejar mejor el *pool* de tRNA ancestral más que el actual, lo que sugiere que los cambios en los sesgos mutacionales y en el número de tRNAs en el genoma tienen efectos menores sobre la identidad de estos tripletes preferidos en los genes de alta expresión.

vi) Las regiones conservadas de los genes presentan preferencias por ciertos codones. En la mayoría de los casos los codones óptimos para velocidad también lo son para velocidad. Encontramos, sin embargo, que en algunos casos existe un desacoplamiento de las preferencias.

vii) La magnitud de los sesgos observados en el UAA en genes de alta expresión respecto al resto del genoma en Mollicutes, y posiblemente en otros endosimbiontes, podría explicarse por un efecto de la selección purificadora actuando de forma diferencial en los distintos genes del genoma. No es claro que el único factor "contra seleccionado" en genes de alta expresión sea el costo energético, ya que la aromaticidad es otro posible factor. La deriva genética y el sesgo mutacional explicarían el aumento del costo energético promedio en genes de baja expresión.

14. Bibliografía

Ahn, I. & Son, H. S. (2012). Evolutionary analysis of human-origin influenza A virus (H3N2) genes associated with the codon usage patterns since 1993. *Virus Genes* **44**, 198-206.

Akashi, H. (1994). Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics* **136**, 927-935.

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

Akman, L., Yamashita, A., Watanabe, H., Oshima, K., Shiba, T., Hattori, M. & Aksoy, S. (2002). Genome sequence of the endocellular obligate symbiont of tsetse flies, Wigglesworthia glossinidia. *Nat Genet* **32**, 402-407.

Al-Saif, M. & Khabar, K. S. (2012). UU/UA dinucleotide frequency reduction in coding regions results in increased mRNA stability and protein expression. *Mol Ther* **20**, 954-959.

Alizon, S. & van Baalen, M. (2008). Multiple infections, immune dynamics, and the evolution of virulence. *Am Nat* **172**, E150-168.

Alizon, S. & Michalakis, Y. (2011). The transmission-virulence trade-off and superinfection: comments to Smith. *Evolution* **65**, 3633-3638; discussion 3639-3641.

Andersson, S. G. & Sharp, P. M. (1996). Codon usage and base composition in Rickettsia prowazekii. *J Mol Evol* **42**, 525-536.

Aragones, L., Guix, S., Ribes, E., Bosch, A. & Pinto, R. M. (2010). Fine-tuning translation kinetics selection as the driving force of codon usage bias in the hepatitis A virus capsid. *PLoS Pathog* **6**, e1000797.

Bahir, I., Fromer, M., Prat, Y. & Linial, M. (2009). Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Mol Syst Biol* **5**, 311.

Bailly-Bechet, M., Vergassola, M. & Rocha, E. (2007). Causes for the intriguing presence of tRNAs in phages. *Genome Res* **17**, 1486-1495.

Banerjee, S., Hess, D., Majumder, P., Roy, D. & Das, S. (2004a). The Interactions of Allium sativum leaf agglutinin with a chaperonin group of unique receptor protein isolated from a bacterial endosymbiont of the mustard aphid. *J Biol Chem* **279**, 23782-23789.

Banerjee, T., Basak, S., Gupta, S. K. & Ghosh, T. C. (2004b). Evolutionary forces in shaping the codon and amino acid usages in Blochmannia floridanus. *J Biomol Struct Dyn* **22**, 13-23.

Barrai, I., Volinia, S. & Scapoli, C. (1995). The usage of oligopeptides in proteins correlates negatively with molecular weight. *Int J Pept Protein Res* **45**, 326-331.

Basak, S., Banerjee, T., Gupta, S. K. & Ghosh, T. C. (2004). Investigation on the causes of codon and amino acid usages variation between thermophilic Aquifex aeolicus and mesophilic Bacillus subtilis. *J Biomol Struct Dyn* **22**, 205-214.

Bernardi, G. (1986). Compositional constraints and genome evolution. *J Mol Evol* **24**, 1-11.

Bishal, A. K., Mukherjee, R. & Chakraborty, C. (2013). Synonymous codon usage pattern analysis of Hepatitis D virus. *Virus Res*.

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

Bragg, J. G., Quigg, A., Raven, J. A. & Wagner, A. (2012). Protein elemental sparing and codon usage bias are correlated among bacteria. *Mol Ecol* **21**, 2480-2487.

Bulmer, M. (1987). Coevolution of codon usage and transfer RNA abundance. *Nature* **325**, 728-730.

Bulmer, M. (1991). The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**, 897-907.

Butler, M. A. & King, A. A. (2004). Phylogenetic comparative analysis: a modeling approach for adaptive evolution. . *American Naturalist* **164**, 683.

Cannarozzi, G., Schraudolph, N. N., Faty, M., von Rohr, P., Friberg, M. T., Roth, A. C., Gonnet, P., Gonnet, G. & Barral, Y. (2010). A role for codon order in translation dynamics. *Cell* **141**, 355-367.

Carbone, A. (2008). Codon bias is a major factor explaining phage evolution in translationally biased hosts. *J Mol Evol* **66**, 210-223.

Carlini, D. B. (2004). Experimental reduction of codon bias in the Drosophila alcohol dehydrogenase gene results in decreased ethanol tolerance of adult flies. *J Evol Biol* **17**, 779-785.

Carlini, D. B. & Stephan, W. (2003). In vivo introduction of unpreferred synonymous codons into the Drosophila Adh gene results in reduced levels of ADH protein. *Genetics* **163**, 239-243.

Castillo-Davis, C. I. & Hartl, D. L. (2002). Genome evolution and developmental constraint in Caenorhabditis elegans. *Mol Biol Evol* **19**, 728-735.

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

Cunningham, C. W., Omland, K. E. & Oakley, T. H. (1998). Reconstructing ancestral character states: a critical reappraisal. *Trends Ecol Evol* **13**, 361-366.

Curran, J. F. & Yarus, M. (1989). Rates of aminoacyl-tRNA selection at 29 sense codons in vivo. *J Mol Biol* **209**, 65-77.

Chan, C. T., Dyavaiah, M., DeMott, M. S., Taghizadeh, K., Dedon, P. C. & Begley, T. J. (2010). A quantitative systems approach reveals dynamic control of tRNA modifications during cellular stress. *PLoS Genet* **6**, e1001247.

Chauhan, N., Vidyarthi, A. S. & Poddar, R. (2011). Comparative multivariate analysis of codon and amino acid usage in three Leishmania genomes. *Genomics Proteomics Bioinformatics* **9**, 218-228.

Chechetkin, V. R. & Lobzin, V. V. (2011). Stability of the genetic code and optimal parameters of amino acids. *J Theor Biol* **269**, 57-63.

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

D'Onofrio, G., Jabbari, K., Musto, H. & Bernardi, G. (1999). The correlation of protein hydropathy with the base composition of coding sequences. *Gene* **238**, 3-14.

D'Onofrio, G., Mouchiroud, D., Aissani, B., Gautier, C. & Bernardi, G. (1991). Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J Mol Evol* **32**, 504-510.

Das, S., Paul, S., Chatterjee, S. & Dutta, C. (2005). Codon and amino acid usage in two major human pathogens of genus Bartonella--optimization between replicational-transcriptional selection, translational control and cost minimization. *DNA Res* **12**, 91-102.

Di Giulio, M. (1989). The extension reached by the minimization of the polarity distances during the evolution of the genetic code. *J Mol Evol* **29**, 288-293.

Di Giulio, M. (1997). On the origin of the genetic code. *J Theor Biol* **187**, 573-581.

Di Giulio, M. & Medugno, M. (1998). The historical factor: the biosynthetic relationships between amino acids and their physicochemical properties in the origin of the genetic code. *J Mol Evol* **46**, 615-621.

Domingo, E. (1997). Rapid evolution of viral RNA genomes. *J Nutr* **127**, 958S-961S.

Dong, H., Nilsson, L. & Kurland, C. G. (1996). Co-variation of tRNA abundance and codon usage in Escherichia coli at different growth rates. *J Mol Biol* **260**, 649-663.

Dorn, A. & Kippenberger, S. (2008). Clinical application of CpG-, non-CpG-, and antisense oligodeoxynucleotides as immunomodulators. *Curr Opin Mol Ther* **10**, 10-20.

dos Reis, M., Savva, R. & Wernisch, L. (2004). Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res* **32**, 5036-5044.

Drummond, D. A. & Wilke, C. O. (2008). Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* **134**, 341-352.

Dufton, M. J. (1997). Genetic code synonym quotas and amino acid complexity: cutting the cost of proteins? *J Theor Biol* **187**, 165-173.

Duret, L. (2002a). Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev* **12**, 640-649.

Duret, L. (2002b). Detecting genomic features under weak selective pressure: the example of codon usage in animals and plants. *Bioinformatics* **18 Suppl 2**, S91.

Duret, L. & Mouchiroud, D. (1999). Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc Natl Acad Sci U S A* **96**, 4482-4487.

Duret, L. & Mouchiroud, D. (2000). Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol* **17**, 68-74.

Ekman, S., Andersen, H. L. & Wedin, M. (2008). The limitations of ancestral state reconstruction and the evolution of the ascus in the Lecanorales (lichenized Ascomycota). *Syst Biol* **57**, 141-156.

Firth, A. E. & Brierley, I. (2012). Non-canonical translation in RNA viruses. *J Gen Virol* **93**, 1385-1409.

Fraser, H. B., Hirsh, A. E., Wall, D. P. & Eisen, M. B. (2004). Coevolution of gene expression among interacting proteins. *Proc Natl Acad Sci U S A* **101**, 9033-9038.

Gaffney, D. J. & Keightley, P. D. (2008). Effect of the assignment of ancestral CpG state on the estimation of nucleotide substitution rates in mammals. *BMC Evol Biol* **8**, 265.

Garat, B. & Musto, H. (2000). Trends of amino acid usage in the proteins from the unicellular parasite Giardia lamblia. *Biochem Biophys Res Commun* **279**, 996-1000.

Gaucher, E. A., Govindarajan, S. & Ganesh, O. K. (2008). Palaeotemperature trend for Precambrian life inferred from resurrected proteins. *Nature* **451**, 704-707.

Gentles, A. J. & Karlin, S. (2001). Genome-scale compositional comparisons in eukaryotes. *Genome Res* **11**, 540-546.

Gingold, H. & Pilpel, Y. (2011). Determinants of translation efficiency and accuracy. *Mol Syst Biol* **7**, 481.

Goni, N., Iriarte, A., Comas, V., Sonora, M., Moreno, P., Moratorio, G., Musto, H. & Cristina, J. (2012). Pandemic influenza A virus codon usage revisited: biases, adaptation and implications for vaccine strain development. *Virol J* **9**, 263.

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

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711-732.

Grosjean, H. & Fiers, W. (1982). Preferential codon usage in prokaryotic genes: the optimal codon-anticodon interaction energy and the selective codon usage in efficiently expressed genes. *Gene* **18**, 199-209.

Heizer, E. M., Jr., Raymer, M. L. & Krane, D. E. (2011). Amino acid biosynthetic cost and protein conservation. *J Mol Evol* **72**, 466-473.

Heizer, E. M., Jr., Raiford, D. W., Raymer, M. L., Doom, T. E., Miller, R. V. & Krane, D. E. (2006). Amino acid cost and codon-usage biases in 6 prokaryotic genomes: a whole-genome analysis. *Mol Biol Evol* **23**, 1670-1680.

Herbeck, J. T., Wall, D. P. & Wernegreen, J. J. (2003). Gene expression level influences amino acid usage, but not codon usage, in the tsetse fly endosymbiont Wigglesworthia. *Microbiology* **149**, 2585-2596.

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

Hershberg, R. & Petrov, D. A. (2009). General rules for optimal codon choice. *PLoS Genet* **5**, e1000556.

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**, e49060.

Higgs, P. G. & Ran, W. (2008). Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. *Mol Biol Evol* **25**, 2279-2291.

Houen, G. (1999). Evolution of the genetic code: the nonsense, antisense, and antinonsense codes make no sense. *Biosystems* **54**, 39-46.

Husnik, F., Chrudimsky, T. & Hypsa, V. (2011). Multiple origins of endosymbiosis within the Enterobacteriaceae (gamma-Proteobacteria): convergence of complex phylogenetic approaches. *BMC Biol* **9**, 87.

Ikemura, T. (1981). 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. *J Mol Biol* **151**, 389-409.

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

Iriarte, A., Baraibar, J. D., Romero, H. & Musto, H. (2011). Selected codon usage bias in members of the class Mollicutes. *Gene* **473**, 110-118.

Iriarte, A., Baraibar, J., Romero, H., Castro, S. & Musto, H. (2013a). Evolution of optimal codon choices in the family Enterobacteriaceae. *Microbiology*.

Iriarte, A., Sanguinetti, M., Fernandez-Calero, T., Naya, H., Ramon, A. & Musto, H. (2012). Translational selection on codon usage in the genus Aspergillus. *Gene* **506**, 98-105.

Iriarte, A., Baraibar, J. D., Diana, L., Castro-Sowinski, S., Romero, H. & Musto, H. (2013b). Trends in amino acid usage across the class Mollicutes. *J Biomol Struct Dyn*.

Jenkins, G. M. & Holmes, E. C. (2003). The extent of codon usage bias in human RNA viruses and its evolutionary origin. *Virus Res* **92**, 1-7.

Kahali, B., Basak, S. & Ghosh, T. C. (2007). Reinvestigating the codon and amino acid usage of S. cerevisiae genome: a new insight from protein secondary structure analysis. *Biochem Biophys Res Commun* **354**, 693-699.

Kahali, B., Basak, S. & Ghosh, T. C. (2008). Delving deeper into the unexpected correlation between gene expressivity and codon usage bias of Escherichia coli genome. *J Biomol Struct Dyn* **25**, 655-661.

Kanaya, S., Yamada, Y., Kudo, Y. & Ikemura, T. (1999). Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene* **238**, 143-155.

Kanaya, S., Yamada, Y., Kinouchi, M., Kudo, Y. & Ikemura, T. (2001). Codon usage and tRNA genes in eukaryotes: correlation of codon usage diversity with translation efficiency and with CG-dinucleotide usage as assessed by multivariate analysis. *J Mol Evol* **53**, 290-298.

Karlin, S. (1998). Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol* **1**, 598-610.

Karlin, S., Doerfler, W. & Cardon, L. R. (1994). Why is CpG suppressed in the genomes of virtually all small eukaryotic viruses but not in those of large eukaryotic viruses? *J Virol* **68**, 2889-2897.

Lafay, B., Lloyd, A. T., McLean, M. J., Devine, K. M., Sharp, P. M. & Wolfe, K. H. (1999). Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res* **27**, 1642-1649.

Li, G. W., Oh, E. & Weissman, J. S. (2012). The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* **484**, 538-541.

Lightfield, J., Fram, N. R. & Ely, B. (2011). Across bacterial phyla, distantly-related genomes with similar genomic GC content have similar patterns of amino acid usage. *PLoS One* **6**, e17677.

Lithwick, G. & Margalit, H. (2003). Hierarchy of sequence-dependent features associated with prokaryotic translation. *Genome Res* **13**, 2665-2673.

Litsios, G. & Salamin, N. (2012). Effects of phylogenetic signal on ancestral state reconstruction. *Syst Biol* **61**, 533-538.

Lobry, J. R. & Gautier, C. (1994). Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. *Nucleic Acids Res* **22**, 3174-3180.

Lucks, J. B., Nelson, D. R., Kudla, G. R. & Plotkin, J. B. (2008). Genome landscapes and bacteriophage codon usage. *PLoS Comput Biol* **4**, e1000001.

Moran, N. A. & Wernegreen, J. J. (2000). Lifestyle evolution in symbiotic bacteria: insights from genomics. *Trends Ecol Evol* **15**, 321-326.

Moratorio, G., Iriarte, A., Moreno, P., Musto, H. & Cristina, J. (2013). A detailed comparative analysis on the overall codon usage patterns in West Nile virus. *Infect Genet Evol*.

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

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

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

Naya, H., Zavala, A., Romero, H., Rodriguez-Maseda, H. & Musto, H. (2004). Correspondence analysis of amino acid usage within the family Bacillaceae. *Biochem Biophys Res Commun* **325**, 1252-1257.

Novoa, E. M. & Ribas de Pouplana, L. (2012). Speeding with control: codon usage, tRNAs, and ribosomes. *Trends Genet* **28**, 574-581.

Novoa, E. M., Pavon-Eternod, M., Pan, T. & Ribas de Pouplana, L. (2012). A role for tRNA modifications in genome structure and codon usage. *Cell* **149**, 202-213.

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature* **401**, 877-884.

Pagel, M., Meade, A. & Barker, D. (2004). Bayesian estimation of ancestral character states on phylogenies. *Syst Biol* **53**, 673-684.

Pal, C., Papp, B. & Hurst, L. D. (2001). Highly expressed genes in yeast evolve slowly. *Genetics* **158**, 927-931.

Pal, C., Papp, B. & Hurst, L. D. (2003). Genomic function: Rate of evolution and gene dispensability. *Nature* **421**, 496-497; discussion 497-498.

Palacios, C. & Wernegreen, J. J. (2002). A strong effect of AT mutational bias on amino acid usage in Buchnera is mitigated at high-expression genes. *Mol Biol Evol* **19**, 1575-1584.

Park, D. & Choi, S. S. (2009). Why proteins evolve at different rates: the functional hypothesis versus the mistranslation-induced protein misfolding hypothesis. *FEBS Lett* **583**, 1053-1059.

Parmley, J. L. & Huynen, M. A. (2009). Clustering of codons with rare cognate tRNAs in human genes suggests an extra level of expression regulation. *PLoS Genet* **5**, e1000548.

Pavon-Eternod, M., David, A., Dittmar, K., Berglund, P., Pan, T., Bennink, J. R. & Yewdell, J. W. (2013). Vaccinia and influenza A viruses select rather than adjust tRNAs to optimize translation. *Nucleic Acids Res* **41**, 1914-1921.

Perez-Jimenez, R., Ingles-Prieto, A., Zhao, Z. M., Sanchez-Romero, I., Alegre-Cebollada, J., Kosuri, P., Garcia-Manyes, S., Kappock, T. J., Tanokura, M. & other authors (2011). Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat Struct Mol Biol* **18**, 592-596.

Philippe, H. & Roure, B. (2011). Difficult phylogenetic questions: more data, maybe; better methods, certainly. *BMC Biol* **9**, 91.

Precup, J. & Parker, J. (1987). Missense misreading of asparagine codons as a function of codon identity and context. *J Biol Chem* **262**, 11351-11355.

Qian, W., Yang, J. R., Pearson, N. M., Maclean, C. & Zhang, J. (2012). Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genet* **8**, e1002603.

R Development Core Team (2012). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Raiford, D. W., Heizer, E. M., Jr., Miller, R. V., Doom, T. E., Raymer, M. L. & Krane, D. E. (2012). Metabolic and translational efficiency in microbial organisms. *J Mol Evol* **74**, 206-216.

Ran, W. & Higgs, P. G. (2010). The influence of anticodon-codon interactions and modified bases on codon usage bias in bacteria. *Mol Biol Evol* **27**, 2129-2140.

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

Retchless, A. C. & Lawrence, J. G. (2011). Quantification of codon selection for comparative bacterial genomics. *BMC Genomics* **12**, 374.

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

Rima, B. K. & McFerran, N. V. (1997). Dinucleotide and stop codon frequencies in single-stranded RNA viruses. *J Gen Virol* **78 (Pt 11)**, 2859-2870.

Rispe, C., Delmotte, F., van Ham, R. C. & Moya, A. (2004). Mutational and selective pressures on codon and amino acid usage in Buchnera, endosymbiotic bacteria of aphids. *Genome Res* **14**, 44-53.

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

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

Rocha, E. P. & Danchin, A. (2004). An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol Biol Evol* **21**, 108-116.

Ronquist, F. (2004). Bayesian inference of character evolution. *Trends Ecol Evol* **19**, 475-481.

Roth, A. C. (2012). Decoding properties of tRNA leave a detectable signal in codon usage bias. *Bioinformatics* **28**, i340-i348.

Roychoudhury, S., Pan, A. & Mukherjee, D. (2011). Genus specific evolution of codon usage and nucleotide compositional traits of poxviruses. *Virus Genes* **42**, 189-199.

Satapathy, S. S., Dutta, M., Buragohain, A. K. & Ray, S. K. (2012). Transfer RNA gene numbers may not be completely responsible for the codon usage bias in asparagine, isoleucine, phenylalanine, and tyrosine in the high expression genes in bacteria. *J Mol Evol* **75**, 34-42.

Saunders, N. F., Thomas, T., Curmi, P. M., Mattick, J. S., Kuczek, E., Slade, R., Davis, J., Franzmann, P. D., Boone, D. & other authors (2003). Mechanisms of thermal adaptation revealed from the genomes of the Antarctic Archaea Methanogenium frigidum and Methanococcoides burtonii. *Genome Res* **13**, 1580-1588.

Schaber, J., Rispe, C., Wernegreen, J., Buness, A., Delmotte, F., Silva, F. J. & Moya, A. (2005). Gene expression levels influence amino acid usage and evolutionary rates in endosymbiotic bacteria. *Gene* **352**, 109-117.

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

Shabalina, S. A., Spiridonov, N. A. & Kashina, A. (2013). Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. *Nucleic Acids Res*.

Shackelton, L. A., Parrish, C. R. & Holmes, E. C. (2006). Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses. *J Mol Evol* **62**, 551-563.

Shanahan, T. (2011). Phylogenetic inertia and Darwin's higher law. *Stud Hist Philos Biol Biomed Sci* **42**, 60-68.

Sharp, P. M. & Li, W. H. (1986). Codon usage in regulatory genes in Escherichia coli does not reflect selection for 'rare' codons. *Nucleic Acids Res* **14**, 7737-7749.

Sharp, P. M. & Li, W. H. (1987). The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol Biol Evol* **4**, 222-230.

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

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 Res* **33**, 1141-1153.

Sharp, P. M., Cowe, E., Higgins, D. G., Shields, D. C., Wolfe, K. H. & Wright, F. (1988). Codon usage patterns in Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster and Homo sapiens; a review of the considerable within-species diversity. *Nucleic Acids Res* **16**, 8207-8211.

Shields, D. C. (1990). Switches in species-specific codon preferences: the influence of mutation biases. *J Mol Evol* **31**, 71-80.

Singer, G. A. & Hickey, D. A. (2000). Nucleotide bias causes a genomewide bias in the amino acid composition of proteins. *Mol Biol Evol* **17**, 1581-1588.

Singer, G. A. & Hickey, D. A. (2003). Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. *Gene* **317**, 39-47.

Smith, J. (2011). Superinfection drives virulence evolution in experimental populations of bacteria and plasmids. *Evolution* **65**, 831-841.

Sorensen, M. A., Kurland, C. G. & Pedersen, S. (1989). Codon usage determines translation rate in Escherichia coli. *J Mol Biol* **207**, 365-377.

Stedman, K. M., Kosmicki, N. R. & Diemer, G. S. (2013). Codon usage frequency of RNA virus genomes from high-temperature acidic-environment metagenomes. *J Virol* **87**, 1919.

Stoletzki, N. & Eyre-Walker, A. (2007). Synonymous codon usage in Escherichia coli: selection for translational accuracy. *Mol Biol Evol* **24**, 374-381.

Subramanian, S. & Kumar, S. (2004). Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics* **168**, 373-381.

Sueoka, N. (1961). Correlation between Base Composition of Deoxyribonucleic Acid and Amino Acid Composition of Protein. *Proc Natl Acad Sci U S A* **47**, 1141-1149.

Supek, F., Skunca, N., Repar, J., Vlahovicek, K. & Smuc, T. (2010). Translational selection is ubiquitous in prokaryotes. *PLoS Genet* **6**, e1001004.

Szathmary, E. (1993). Coding coenzyme handles: a hypothesis for the origin of the genetic code. *Proc Natl Acad Sci U S A* **90**, 9916-9920.

Tekaia, F. & Yeramian, E. (2006). Evolution of proteomes: fundamental signatures and global trends in amino acid compositions. *BMC Genomics* **7**, 307.

Tekaia, F., Yeramian, E. & Dujon, B. (2002). Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. *Gene* **297**, 51-60.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I. & other authors (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* **141**, 344-354.

Wagner, A. (2005). Energy constraints on the evolution of gene expression. *Mol Biol Evol* **22**, 1365-1374.

Wang, B., Shao, Z. Q., Xu, Y., Liu, J., Liu, Y., Hang, Y. Y. & Chen, J. Q. (2011). Optimal codon identities in bacteria: implications from the conflicting results of two different methods. *PLoS One* **6**, e22714.

Washenberger, C. L., Han, J. Q., Kechris, K. J., Jha, B. K., Silverman, R. H. & Barton, D. J. (2007). Hepatitis C virus RNA: dinucleotide frequencies and cleavage by RNase L. *Virus Res* **130**, 85-95.

Weissenbach, J. & Dirheimer, G. (1978). Pairing properties of the methylester of 5-carboxymethyl uridine in the wobble position of yeast tRNA3Arg. *Biochim Biophys Acta* **518**, 530-534.

Williams, P. D. (2012). New insights into virulence evolution in multigroup hosts. *Am Nat* **179**, 228-239.

Withers, M., Wernisch, L. & dos Reis, M. (2006). Archaeology and evolution of transfer RNA genes in the Escherichia coli genome. *RNA* **12**, 933-942.

Wong, E. H., Smith, D. K., Rabadan, R., Peiris, M. & Poon, L. L. (2010). Codon usage bias and the evolution of influenza A viruses. Codon Usage Biases of Influenza Virus. *BMC Evol Biol* **10**, 253.

Wong, J. T. (1975). A co-evolution theory of the genetic code. *Proc Natl Acad Sci U S A* **72**, 1909-1912.

Yamao, F., Andachi, Y., Muto, A., Ikemura, T. & Osawa, S. (1991). Levels of tRNAs in bacterial cells as affected by amino acid usage in proteins. *Nucleic Acids Res* **19**, 6119-6122.

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

Yang, Z., Nielsen, R. & Hasegawa, M. (1998). Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol Biol Evol* **15**, 1600-1611.

Young, M., Bolduc, B., Shaughnessy, D. P., Roberto, F. F., Wolf, Y. I. & Koonin, E. V. (2013). Reply to "codon usage frequency of RNA virus genomes from high-temperature acidic-environment metagenomes". *J Virol* **87**, 1920-1921.

Zavala, A., Naya, H., Romero, H. & Musto, H. (2002). Trends in codon and amino acid usage in Thermotoga maritima. *J Mol Evol* **54**, 563-568.

Agradecimientos